Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Typologie des ménages dans la statistique STATPOP (HHTyp)

Description du processus de production et des tests de différentes méthodes d'imputation

EXPERIMENTAL STATISTICS

Neuchâtel, 2022

Éditeur:	Office fédéral de la statistique (OFS)	Concept de mise en page:	section PUB
Renseignements:	info.dem@bfs.admin.ch	Téléchargement:	www.experimental.bfs.admin.ch
Rédaction:	Sabrina Bravo, Estelle Crippa	Copyright:	OFS, Neuchâtel 2022
Domaine:	00 Bases statistiques des généralités		Reproduction autorisée seulement avec l'accord préalable des auteurs
Langue du texte original:	Français		

ESPACE DE L'EUROPE 10 CH-2010 NEUCHÂTEL WWW.STATISTIQUE.CH

Table des matières

1 But du document 4

2 Situation initiale 4

3 Objectifs 4

4 Résumé de la procédure 5

5 Sources de données pour les relations 6

6 Stratégie d'imputation - tests de différentes méthodes 7

6.1 Méthodes d'imputation 7
6.2 Performance des méthodes 9
6.2.1 Résultats pour les données de l'année 2018 10
6.2.2 Résumé des résultats 2018-2019 13
6.3 Estimation de la qualité 14
6.4 Choix de la méthode 16

7 Production de la variable HHTyp 17

7.1 Table Personne 17
7.2 Table Relation 17
7.3 Table Typologie 20
7.4 Imputations 20
7.4.1 Description technique de la méthode Arbre de décision 20
7.4.2 Variables du modèles 23
7.4.3 Echantillon d'apprentissage 25
7.4.4 Problématique du type : Couple de même sexe avec enfant 27

8 Estimation de la qualité 33

8.1 Indicateur de qualité 34
8.2 Taux de HHTyp manquants par commune 35
8.3 Analyse des erreurs 36

9 Perspectives 37

Annexes 38

Liste des abréviations

Abréviation	*Signification*
1P	Type de ménage : ménage d'une personne
A	Type de ménage : autre ménage de plusieurs personnes
AD	Méthode d'imputation statistique du type de ménage basé sur un Arbre de décision
ADRL	Méthode d'imputation statistique du type de ménage basé sur un Arbre de décision suivi d'une régression logistique pour améliorer l'imputation du type MSE
EBM	Enquête sur le budget des ménages
EKL	Section Revenus, consommation et conditions de vie de l'OFS
HHTyp	Nom de la variable Type de ménage (Household Type)
INFOSTAR	Registre informatisé de l'état civil géré par l'Office fédéral de la Justice (OFJ)
M	Type de ménage : couple marié sans enfant
ME	Type de ménage : couple marié avec au moins un enfant de moins de 25 ans (ou moins de 18 ans pour la typologie « 18 ans »)
ML	Méthode d'imputation statistique de la relation entre deux membres d'un ménage basé sur un algorithme de Machine Learning nommé Random Forest
MS	Type de ménage : couple de même sexe sans enfant
MSE	Type de ménage : couple de même sexe avec au moins un enfant de moins de 25 ans (ou moins de 18 ans pour la typologie « 18 ans »)
OFS	Office fédéral de la statistique
ORDIPRO	Système d'information destiné à la saisie des données relatives aux diplomates étrangers et fonctionnaires internationaux, exploité par le Département fédéral des affaires étrangères
PE	Type de ménage : parent seul avec au moins un enfant de moins de 25 ans (ou moins de 18 ans pour la typologie « 18 ans »)
RF	Méthode d'imputation statistique du type de ménage basé sur un Random Forest
RM	Méthode d'imputation statistique du type de ménage basé sur une Régression multinomiale
RS	Relevé structurel de l'OFS (enquête par échantillonnage qui fait partie du nouveau système de recensement fédéral de la population)
SILC	Enquête sur les revenus et les conditions de vie
STATPOP	Statistique de la population et des ménages
Stat-VD	Méthode d'imputation du type de ménage développé par l'Office statistique du canton de Vaud
UL	Type de ménage : couple en union libre sans enfant
ULE	Type de ménage : couple en union libre avec au moins un enfant de moins de 25 ans (ou moins de 18 ans pour la typologie « 18 ans »)

1 But du document

Ce document décrit les processus de production de la variable Type de ménage dans les données STATPOP, ainsi que les différentes méthodes d'imputation des données manquantes, testées dans le cadre de ce projet.

Dans le chapitre 4 se trouve un résumé de la procédure, sans les détails, dans le but de comprendre les tables principales et les processus de production de la variable Type de ménage. Dans le chapitre 6, les différentes techniques d'imputations testées sont montrées avec leurs résultats, pour arriver au choix final de la méthode. Finalement, c'est au chapitre 7 que la procédure de production est détaillée, de la préparation des différentes tables et la création de l'échantillon d'apprentissage jusqu'au modèle d'imputation. Au chapitre 8 est décrit le calcul de l'estimateur de qualité de la méthode.

2 Situation initiale

Jusqu'en 2000, la variable « Type de ménage », qui permet de décrire la configuration des relations entre les personnes qui vivent dans le même ménage (p. ex. couple marié avec enfants, ménages de personnes non-apparentées, etc.), était relevée dans les recensements exhaustifs de la population qui avaient lieu tous les dix ans. Avec la mise en place du nouveau système de recensement à partir de 2010, cette variable est disponible dans le Relevé Structurel (RS), enquête annuelle réalisée sur un échantillon d'environ 300 000 personnes âgées de 15 ans ou plus, faisant partie de la population résidante permanente et vivant dans des ménages privés. Les résultats du Relevé Structurel permettent d'obtenir des estimations directes et fiables pour des niveaux géographiques d'au moins 15 000 personnes. Pour avoir des estimations à un niveau géographique plus fin, les données du Relevé Structurel regroupées sur cinq ans (pooling) permettent d'obtenir des estimations jusqu'à des ensembles d'environ 3 000 personnes. Les résultats de la Typologie des ménages ne sont, par conséquent, pas disponibles de façon exhaustive, c'est-à-dire qu'ils ne sont disponibles ni pour tous les ménages, ni pour l'ensemble des communes de Suisse.

Le type de ménage est une information importante non seulement dans le domaine des ménages (au niveau suisse le Relevé Structurel est tout à fait adapté aux besoins), mais aussi en lien avec d'autres thématiques. Pour des besoins de pondérations des enquêtes sur les revenus et les conditions de vie (SILC) et sur le budget des ménages (EBM), la section Revenus, consommation et conditions de vie (EKL) a défini, pour l'ensemble de la population, des types de ménages principaux (personnes seules, ménages de deux adultes avec ou sans enfants, adultes seuls avec enfants) basés uniquement sur les différences d'âge entre membres du ménage. Une variable supplémentaire précise la présence d'au moins un couple marié dans le ménage. L'Office statistique du Canton de Vaud a aussi développé, il y a quelques années, un programme pour produire une typologie de ménage basée sur les données des registres. Les résultats ne sont pas diffusés directement, mais uniquement en lien avec d'autres données, notamment pour des analyses sur la pauvreté. Ce sont leurs travaux, dont la typologie est proche de celle du Relevé Structurel, qui ont donné l'impulsion pour le lancement de ce projet de développement de la typologie des ménages exhaustive à l'OFS.

3 Objectifs

Le but de ce projet est d'avoir une variable Type de ménage (HHTyp) pour l'ensemble des ménages privés de la population résidante permanente au domicile principal. Cela permettra d'avoir :

- Des résultats sur les types de ménage à un niveau géographique fin (p. ex. pour avoir la répartition des types de ménage par commune) ;

- Des types de ménage au niveau individuel comme variable d'analyse dans d'autres domaines (p. ex. appariement avec d'autres données pour comparer les revenus de travail par type de ménage à un niveau géographique fin).

4 Résumé de la procédure

Dans le Relevé Structurel, le type de ménage n'est pas demandé directement dans le questionnaire. Il est calculé sur la base de l'ensemble des relations entre les différents membres d'un ménage, communiquées par les personnes de l'échantillon. Sachant que certaines sources de données utilisées pour la production de STATPOP nous renseignent sur les relations entre les personnes, il a été décidé, pour ce projet, de suivre la méthode du Relevé Structurel, à savoir de calculer le type de ménage sur la base des relations.

Les modalités de la typologie à produire, reprises directement de la typologie du Relevé Structurel, sont les suivantes :

- Ménage d'une personne (1P)
- Couple marié sans enfant (M)
- Couple en union libre[1] sans enfant (UL)
- Couple de même sexe[2] sans enfant (MS)
- Parent seul avec au moins un enfant[3] de moins de 25 ans (PE)
- Couple marié avec au moins un enfant de moins de 25 ans (ME)
- Couple en union libre avec au moins un enfant de moins de 25 ans (ULE)
- Couple de même sexe avec au moins un enfant de moins de 25 ans (MSE)
- Autre ménage de plusieurs personnes (A)

La typologie se crée en plusieurs étapes. Sur la base de la formation des ménages[4] définie dans STATPOP, une **table Personne** est créée avec un identifiant ménage. Les personnes sont numérotées par ordre décroissant en fonction de leur âge (voir tableau 1).

Tableau 1 : Exemple de « Table Personne » avec 3 ménages M1, M2, M3

Ménage	Personne
M1	P1
M1	P2
M1	P3
M2	P1
M3	P1
M3	P2

Le ménage M1 a trois personnes P1, P2, P3, le ménage M2 a une personne P1, le ménage M3 à deux personnes P1, P2.

A partir de la table Personne, une **table Relation** est créée (voir tableau 2). Elle contient chaque paire de membres du ménage auquel elle appartient, ainsi que la relation entre les deux personnes.

---

[1] Dans les couples en union libre sont considérés uniquement les couples hétérosexuels.

[2] Dans les couples de même sexe sont considérés les couples en union libre et les couples en partenariat enregistré.

[3] Un enfant est la fille ou le fils d'un autre membre du ménage indépendamment de l'âge de l'enfant.

[4] Un ménage est formé de toutes les personnes partageant le même logement.

Tableau 2 : Exemple de « Table Relation »

Ménage	Personne1	Personne2	Relation
M1	P1	P2	Epoux
M1	P1	P3	Père
M1	P2	P1	Epouse
M1	P2	P3	Mère
M1	P3	P2	Fille
M1	P3	P1	Fille
M3	P1	P2	Concubin
M3	P2	P1	Concubine

Le ménage M2 n'y est pas, car il n'est formé que d'une seule personne. Chaque relation possède sa symétrie (P1-P2 et P2-P1). Pour un ménage de n personnes, la table Relation aura n(n-1) lignes (si n>1).*

Le programme du Relevé Structurel qui définit le type de ménage en fonction des relations du ménage est ensuite utilisé pour créer une **table Typologie** (voir tableau 3). Le type de ménage (HHTyp) ne peut être calculé que si toutes les relations du ménage sont connues.

Tableau 3 : Exemple de « Table Typologie » avec 3 ménages M1, M2, M3

Ménage	HHTyp
M1	Couple marié avec enfant
M2	Ménage d'une personne
M3	Couple en union libre sans enfant

5 Sources de données pour les relations

Les données utilisées dans ce projet sont celles de la statistique STATPOP et du Relevé Structurel. Les informations de STATPOP qui donnent directement des relations proviennent des deux sources suivantes :

- *Registre informatisé de l'état civil* (**INFOSTAR**) : dans les données INFOSTAR reprisent dans STATPOP se trouvent les liens « père-enfant », « mère-enfant », « épouse-époux » et « partenaires enregistrés ». Toutes les personnes qui ont eu un événement d'état civil (p. ex. être né, avoir un enfant, se marier, etc.) en Suisse depuis la fin des années 1990 se trouvent dans le registre INFOSTAR. Environ 85% de la population résidante se trouvent actuellement dans INFOSTAR. Pour avoir une information de « liens », il est nécessaire que les deux personnes liées soient inscrites dans INFOSTAR.

- *Registre des diplomates et fonctionnaires internationaux* (**ORDIPRO**) : toutes les personnes ayant une autorisation de travail en Suisse délivrée par le DFAE ainsi que les membres de leur famille et leur personnel de maison sont inscrits dans le registre ORDIPRO. La variable donnant les relations entre la personne de référence (diplomate ou fonctionnaire international) et certains membres majeurs de sa famille, comme par exemple les époux, les concubins, les enfants, est reprise dans une table spécifique de STATPOP.

En plus des relations données directement par STATPOP, les relations sont complétées des trois manières suivantes :

- *Relevé Structurel* (**RS**) à partir de 2010 : reprise d'informations relevées annuellement directement auprès d'un échantillon de la population. La relation annoncée dans le Relevé Structurel est reprise pour toutes les paires de personnes vivant toujours ensemble dans un ménage entre le moment

du relevé et la production de la typologie, avec des conditions sur l'état civil pour les concubins, partenaires, époux, apparentés et non-apparentés. Tous les types de liens qui existent dans le Relevé Structurel sont ainsi repris.

- ***Algorithmes déterministes*** : sur la base de variables démographiques présentes dans les données STATPOP, il est possible de définir des relations entre deux personnes avec 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 époux-épouse.

- ***Relations éloignées*** : sur la base des relations obtenues avec STATPOP et le Relevé Structurel, l'information peut être déduite pour certaines paires n'ayant pas encore de relation et qui ont des liens de parenté éloignés. Il est par exemple possible de définir un lien petit-enfant/grand-parent, si le lien entre l'enfant et un de ses parents est déjà défini et que le lien entre le grand-parent et le parent en question aussi.

Le nombre d'attributions de relations basées sur les données à disposition dans STATPOP et dans le Relevé Structurel (appelées « relations sûres » dans la suite de ce document), atteint environ 84% des paires (voir le détail dans le chapitre 7). Sur la base des relations connues de ces 84% de paires, un type de ménage peut être calculé pour 86% des ménages. Seuls les ménages dont toutes les paires ont une relation connue peuvent se voir attribuer un type de ménage. Les ménages composés d'une seule personne reçoivent automatiquement le type « ménage d'une personne ». Les ménages, dont au moins une relation n'est pas définie, n'ont pas de type de ménage.

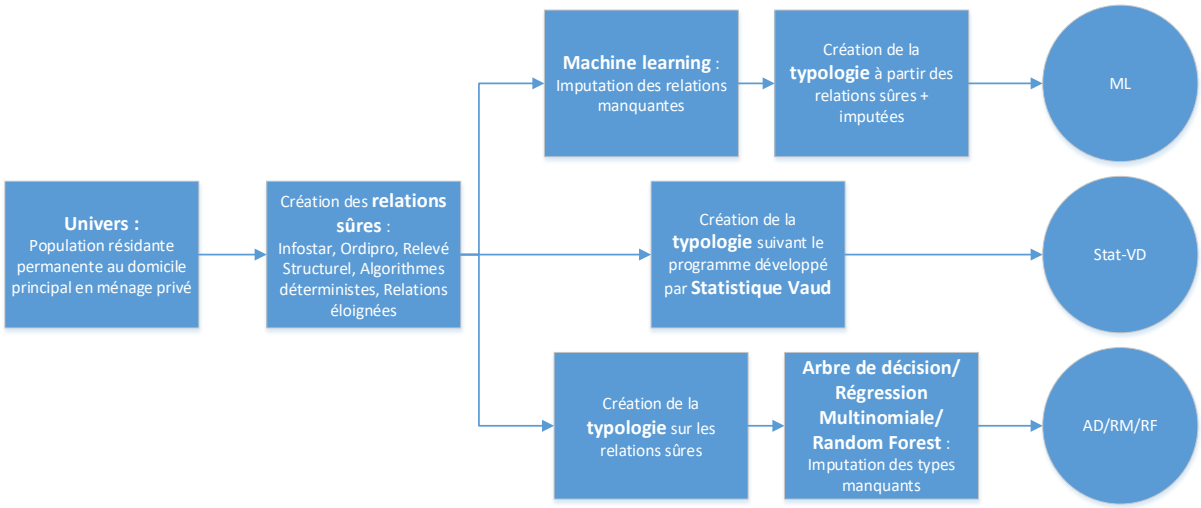
6 Stratégie d'imputation - tests de différentes méthodes

Pour les 13,8% de ménages (environ 500 000 ménages) pour lesquels un type n'a pas pu être défini sur la base des relations sûres (détaillées aux chapitres 7.1 à 7.3), des imputations doivent être réalisées. Dans ce chapitre, nous parlons des différentes méthodes appliquées pour imputer les données manquantes, nous comparons ensuite les résultats de chaque méthode pour les années 2018 et 2019 (chapitre 6.2) pour conclure par le choix du modèle d'imputation (chapitre 6.4). Le modèle choisi sera détaillé au chapitre 7.4.

6.1 Méthodes d'imputation

Nous avons testé différentes méthodes d'imputation afin d'obtenir un type de ménage pour l'ensemble des ménages. Nous partons de l'effectif et des ménages de STATPOP pour créer la table **Personne**. Ensuite, en se basant sur les données de STATPOP et du Relevé Structurel, nous créons la table **Relation**, dans laquelle est indiquée, pour chaque paire de personnes du même ménage, leur relation (père, mère, fille, fils, épouse, etc.). Dans cette table se trouvent des **relations manquantes** : soit parce qu'on ne connait pas la relation (aucune présence dans INFOSTAR, ORDIPRO, RS et ne répond pas aux critères des algorithmes déterministes), soit parce qu'il n'y a pas de lien familial ou de concubinage entre les deux personnes (p.ex. colocataire ou sous-locataire) et que leur relation « non-apparenté » ne peut pas être reprise du RS (pas de présence ou changement de ménage).

Figure 1 : Schéma des différentes méthodes d'imputation testées



A partir de la table **Relation,** nous avons testé trois stratégies d'imputation (voir figure 1) :

1. **Imputer les relations manquantes** pour avoir une table de relation complète qui permet ensuite de calculer le type de ménage. Pour imputer les relations, nous avons utilisé un algorithme de machine learning appelé *Random Forest*. L'algorithme crée un certain nombre d'arbres sur la base d'un échantillon de données ne contenant que des relations connues, l'échantillon d'apprentissage. Chaque arbre a pour but de séparer au mieux les relations en fonction de variables explicatives choisies. Selon les tests effectués, le nombre d'arbres a été fixé à 300 et le nombre de variables explicatives testées à chaque embranchement (nœud) à 8. En appliquant ces arbres aux données à imputer, la relation choisie par le plus grand nombre d'arbres est imputée.
   ⇨ Typologie « ML »

2. Utiliser le *programme développé par l'Office statistique du canton de Vaud*, qui crée des relations et **donne directement un type de ménage pour l'ensemble des ménages**. Le programme est basé d'une part sur les relations sûres et d'autres part sur des algorithmes déterministes qui tiennent compte notamment de l'état civil et de la différence d'âge entre les membres du ménage ainsi que de certaines variables comme le lieux d'origine par exemple.
   ⇨ Typologie « Stat-VD »

3. Calculer le type de ménage pour tous les ménages dont toutes les relations sont connues et ensuite **imputer les types de ménage manquants**. Cette stratégie diffère du premier point, car ici, c'est directement le type de ménage qui est imputé et non les relations. Pour cette troisième stratégie, nous essayons trois algorithmes. Pour chacun de ces algorithmes, nous avons choisi des variables au niveau du ménage dans le but d'expliquer la variable Type de ménage (p. ex. nombre de personnes dans le ménage, nombre de personnes avec l'état civil « Marié » dans le ménage, commune de domicile urbaine-rurale, taille de la commune, etc.) et un échantillon d'apprentissage avec le type de ménage connu. Les trois algorithmes sont les suivants :

   a. *Arbre de décision* : en utilisant les données d'apprentissage, un arbre est créé dans le but de séparer au mieux les Types de ménage en fonction des variables explicatives. Puis dans un deuxième temps, l'arbre de décision est appliqué aux ménages dont le type n'est pas connu, et impute la valeur qui correspond.
      ⇨ Typologie « AD »

b. *Régression multinomiale* : cette méthode permet d'estimer la probabilité de chaque type de ménage en fonction des variables explicatives. Après avoir estimé les paramètres du modèle avec l'échantillon d'apprentissage, nous l'appliquons aux données à imputer. Le type de ménage ayant la probabilité estimée la plus élevée est alors imputé.
   ⇨ Typologie « RM »

c. *Random forest* : cette fois l'algorithme du Random forest est appliqué sur la variable Type de ménage. L'algorithme crée un certain nombre d'arbres qui ont pour but de séparer au mieux les types de ménage, sur la base de l'échantillon d'apprentissage et des variables explicatives. Selon les tests effectués, le nombre d'arbres a été fixé à 1 000 et le nombre de variables testées à chaque nœud à 6. En appliquant ces arbres sur les données à imputer, le type de ménage choisi par le plus grand nombre d'arbres est imputé.
   ⇨ Typologie « RF »

L'algorithme développé par Statistique Vaud (Stat-VD) est un algorithme déterministe qui ne contient donc pas d'aléatoire et imputera chaque année de la même manière. Au contraire, dans notre cas, la Régression multinomiale (RM) se base chaque année sur un échantillon d'apprentissage pour ajuster son modèle. Ce dernier est ensuite appliqué aux données à imputer. Autrement dit, il n'est pas aléatoire, mais peut ainsi capturer les changements de comportement de la population au fil du temps. L'Arbre de décision (AD) contient quant à lui une faible part d'aléatoire. En effet, chaque année, l'arbre complet est créé sans aléatoire à partir de l'échantillon d'apprentissage. Cependant, pour éviter un « overfitting »[5], un algorithme de pruning (expliqué au chapitre 7.4.1) est appliqué et réduit la taille de l'arbre. Cet algorithme de pruning contient de l'aléatoire. Les deux algorithmes de Random Forest (ML et RF), par contre, utilisent fortement l'aléatoire. Pour chaque arbre, un tirage aléatoire avec remise est effectué dans l'échantillon d'apprentissage de base pour constituer l'échantillon d'apprentissage de cet arbre, et pour chaque nœud dans l'arbre, un sous-ensemble de variables est sélectionné aléatoirement parmi les variables explicatives du modèle[6].

6.2 Performance des méthodes

Pour évaluer les résultats des différentes méthodes, nous avons comparé nos résultats aux estimations du Relevé Structurel. Afin de disposer de la même population que le Relevé Structurel, nous avons enlevé de l'effectif STATPOP, les personnes provenant de la source ORDIPRO, car ces personnes ne figurent pas dans le Relevé Structurel[7].

Trois tests sont effectués sur les résultats des cinq méthodes pour évaluer la qualité de l'imputation :

1. Comparer la distribution des types de ménages de l'ensemble de la population à la distribution estimée du Relevé Structurel.

2. Comparer la distribution des imputations à une distribution estimée des types de ménage manquants basée sur l'estimation du Relevé Structurel (distribution cible des imputations pour s'approcher de la distribution du RS).

---

[5] Nous parlons d'overfitting ou de « surapprentissage » lorsqu'un modèle correspond trop précisément à un ensemble de données. Ce qui pourrait le rendre moins fiable sur de nouvelles données et entraîner des instabilités.

[6] Le nombre de variables explicatives sélectionné aléatoirement est défini dans le paramètre « mtry ».

[7] Les personnes issues d'ORDIPRO affectent environ 10 700 ménages (dont env. 8 200 dans le canton de Genève) sur un total d'environ 3 800 000 ménages.

3. Comparer individuellement les imputations avec le Relevé Structurel de l'année considérée (qui n'est pas utilisé pour les relations sûres[8]). A savoir, comparer le type de ménage des ménages imputés qui ont été interrogés dans le Relevé Structurel et dont nous connaissons le type six mois après la production des Types de ménage.

Pour comparer la distribution des imputations à une distribution estimée du Relevé Structurel (point 2 ci-dessus), nous regardons la distribution des Types de ménages sans imputation, uniquement basés sur les relations sûres. Nous calculons ensuite la répartition que devraient avoir les ménages dont le type n'est pas connu (appelé « manquants » dans la suite du document) pour que la distribution sur la population totale soit identique à l'estimation du Relevé Structurel. Cela nous donne une estimation de la distribution des manquants (voir tableau 4).

Tableau 4 : Exemple d'estimation de distribution des types de ménage manquants, en %

	Estimation Relevé structurel	Relation Sure	Différence	Estimation distribution des manquants
Type de ménage manquant	0,00	13,69	-	-
Ménage d'une personne	35,69	35,69	0,00	0,00
Parent seul avec au moins un enfant de moins de 25 ans	4,64	4,49	0,15	1,10
Autre ménage de plusieurs personnes	7,59	4,62	2,97	21,69
Couple marié sans enfant	20,11	18,88	1,23	8,99
Couple en union libre sans enfant	6,53	1,76	4,77	34,82
Couple de même sexe sans enfant	0,60	0,23	0,37	2,71
Couple marié avec au moins un enfant de moins de 25 ans	22,12	18,60	3,52	25,67
Couple en union libre avec au moins un enfant de moins de 25 ans	2,70	2,02	0,67	4,90
Couple de même sexe avec au moins un enfant de moins de 25 ans	0,03	0,01	0,01	0,10
TOTAL	**100**	**100**	**13,69**	**100**

La colonne « Estimation du Relevé structurel » donnera l'estimation du Relevé structurel pour le tableau 5 (résultats pour le point 1 ci-dessus) et la colonne « Estimation distribution des manquants » donnera l'estimation du Relevé structurel pour le tableau 6 (résultats pour le point 2 ci-dessus).

6.2.1 Résultats pour les données de l'année 2018

Dans le tableau 5, nous examinons la différence de distribution entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation sur l'ensemble de la population (point 1 du chapitre 6.2). Nous y ajoutons une mesure de distance définie comme suit :

$$dist = \sqrt{\sum_{i \in A} (x_i - y_i)^2}$$

A : ensemble des modalités des Types de ménage (le type « manquant » est inclus)
x_i : valeur en pourcent de l'estimation du RS pour l'année 2018 pour le type de ménage i
y_i : valeur en pourcent de la méthode d'imputation (p. ex. « ML ») pour le type de ménage i

[8] En raison des calendriers de production, les données du Relevé Structurel de l'année considérée ne sont pas encore disponibles lors de la production de la typologie STATPOP. Pour les relations sûres, seules les années 2010 à l'année t-1 du Relevé Structurel sont utilisées. C'est pour cela qu'il est possible de faire cette comparaison individuelle avec les résultats du Relevé Structurel pour les années passées en gardant la même méthode que pour la production.

Tableau 5 : Différence de distribution en % des Types de ménage sur la population entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation

	Estimation RS2018	ML	Diff^2 ML	Stat-VD	Diff^2 Stat-VD	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	35,686	35,686	0,000	35,686	0,000	35,686	0,000	35,686	0,000	35,686	0,000
PE	4,639	5,186	0,300	5,152	0,264	5,270	0,398	5,232	0,352	5,227	0,346
A	7,590	8,288	0,487	8,144	0,307	7,756	0,028	8,007	0,174	7,850	0,068
M	20,109	19,505	0,365	19,373	0,542	19,397	0,507	19,447	0,438	19,396	0,509
UL	6,529	7,508	0,957	7,915	1,919	7,735	1,453	7,512	0,965	7,695	1,357
MS	0,605	0,303	0,091	0,246	0,129	0,319	0,082	0,252	0,125	0,355	0,063
ME	22,118	20,632	2,211	20,424	2,872	20,877	1,542	20,873	1,550	20,784	1,780
ULE	2,697	2,871	0,031	3,045	0,121	2,934	0,056	2,955	0,067	2,982	0,082
MSE	0,027	0,021	0,000	0,015	0,000	0,026	0,000	0,036	0,000	0,025	0,000
dist			**2,107**		**2,481**		**2,017**		**1,916**		**2,051**

On remarque que les résultats de toutes les méthodes sont assez proches de la distribution estimée par le Relevé Structurel. Ceux de l'algorithme développé par Statistique Vaud sont légèrement plus éloignés.

Nous nous concentrons maintenant uniquement sur les imputations. Le tableau 6 contient les distributions des imputations, ainsi que la distance à la distribution visée, selon l'estimation sur la base du Relevé Structurel et les différentes méthodes d'imputation (point 2 du chapitre 6.2).

Tableau 6 : Différence de distribution en % des Types de ménage sur les imputations entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation

	Estimation RS2018	ML	Diff^2 ML	Stat-VD	Diff^2 Stat-VD	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF
.	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
PE	1,103	5,103	15,994	4,853	14,061	5,711	21,235	5,436	18,769	5,400	18,458
A	21,694	26,790	25,960	25,743	16,387	22,912	1,482	24,743	9,294	23,596	3,615
M	8,991	4,581	19,451	3,617	28,885	3,791	27,039	4,157	23,375	3,782	27,142
UL	34,823	41,964	51,001	44,938	102,318	43,625	77,473	41,998	51,477	43,330	72,375
MS	2,715	0,510	4,860	0,090	6,890	0,628	4,355	0,138	6,639	0,889	3,334
ME	25,668	14,811	117,874	13,293	153,137	16,600	82,230	16,576	82,664	15,925	94,932
ULE	4,904	6,182	1,632	7,450	6,478	6,637	3,000	6,787	3,546	6,992	4,360
MSE	0,102	0,058	0,002	0,016	0,007	0,096	0,000	0,165	0,004	0,086	0,000
dist			**15,387**		**18,115**		**14,725**		**13,992**		**14,974**

Il est normal que la distance soit ici plus élevée, car nous avons « zoomé » sur les imputations. Nous remarquons notamment que la méthode d'imputation Machine learning (ML) laisse des types de ménage manquants (0,001%). En effet, les relations créées par cette méthode peuvent ne pas être cohérentes pour l'algorithme qui crée le type de ménage en fonction des relations.

Le tableau 7 nous montre le résultat des comparaisons individuelles entre les imputations 2018 et les résultats du Relevé Structurel 2018 (point 3 du chapitre 6.2). Ainsi, en multipliant par les poids du Relevé Structurel, nous pouvons estimer le nombre d'erreurs effectuées par l'algorithme. Un petit ajustement est fait pour caler sur le nombre de manquants.

$$\frac{1}{n}\sum_{i\in I} c\, w_i \mathbf{1}_{\{x_i \neq y_i\}}$$

n : nombre de ménages du domaine d'analyse sous revue (p. ex. population totale, imputée)
I : ensemble des ménages dans l'intersection RS ∩ Imputations
x_i : Type de ménage selon le RS pour le ménage i
y_i : Type de ménage imputé pour le ménage i
w_i : poids du RS pour le ménage i
c : nombre de ménages imputés divisé par la somme des poids ménage du RS de l'intersection RS ∩ Imputations ($\sum_{i\in I} w_i$), ainsi $\sum_{i\in I} c\, w_i$ = nombre de ménages imputés
$\mathbf{1}_{\{x_i \neq y_i\}}$: Indicatrice des erreurs d'imputation, elle vaut 1 si l'imputation est incorrecte ($x_i \neq y_i$) et vaut 0 sinon.

Tableau 7 : Estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS

	ML	Stat-VD	RM	AD	RF
% imputations incorrectes estimé sur les manquants	10,61	14,61	12,15	10,55	9,99
% imputations incorrectes estimé sur la population totale	1,45	2,00	1,66	1,45	1,37

En 2018, il y a 28 725 observations non pondérées qui correspondent à 455 796,65 observations pondérées.

En comparaison individuelle, les méthodes Stat-VD et Régression multinomiale (RM) sont un peu moins bonnes que les autres, bien que la méthode RM était très bonne en distribution. Nous notons que l'Arbre de décision (AD) et le Random Forest sur les Types de ménage (RF) sont les deux meilleures méthodes selon les trois types de tests réalisés. Cette analyse a également été effectuée par canton et les résultats se trouvent à l'annexe 1.

6.2.2 Résumé des résultats 2018-2019

Dans ce sous-chapitre se trouve un résumé des résultats pour les différents tests effectués, leurs détails sont disponibles à l'annexe 1. Les trois types de tests ont aussi été effectués avec les données de l'année 2019, ainsi que pour une typologie supplémentaire qui considère les ménages avec enfants de moins de 18 ans (à la place de 25 ans). Cette deuxième typologie (18 ans) répond mieux aux besoins de certains utilisateurs. Elle n'a cependant pas été testée avec les méthodes Machine learning et Statistique Vaud.

Dans le tableau 8 se trouve la distance (définie au chapitre 6.2.1) des distributions totales et imputées entre les différentes méthodes et le Relevé Structurel (points 1 et 2 du chapitre 6.2).

Tableau 8 : Distance entre la distribution des Types de ménage des différentes méthodes et celle estimée par le Relevé Structurel

2018	25 ans	ML	Stat-VD	RM	AD	RF	18 ans	RM	AD	RF
% imputations incorrectes estimé sur les manquants		15,39	18,12	14,72	13,99	14,97		11,75	11,54	11,82
% imputations incorrectes estimé sur le total		2,11	2,48	2,02	1,92	2,05		1,61	1,58	1,62

2019	25 ans	ML	Stat-VD	RM	AD	RF	18 ans	RM	AD	RF
% imputations incorrectes estimé sur les manquants		13,96	16,71	13,50	13,27	13,48		10,56	10,20	10,39
% imputations incorrectes estimé sur le total		1,85	2,22	1,79	1,76	1,79		1,40	1,35	1,38

Le tableau 9 montre l'estimation du pourcentage d'erreur estimé en comparant individuellement les imputations aux résultats du Relevé Structurel (point 3 du chapitre 6.2).

Tableau 9 : Estimation du pourcentage d'erreur en comparant individuellement les imputations au RS

2018	25 ans	ML	Stat-VD	RM	AD	RF	18 ans	RM	AD	RF
% imputations incorrectes estimé sur les manquants		10,61	14,61	12,15	10,55	9,99		11,25	9,80	9,29
% imputations incorrectes estimé total		1,45	2,00	1,66	1,45	1,37		1,54	1,34	1,27

2019	25 ans	ML	Stat-VD	RM	AD	RF	18 ans	RM	AD	RF
% imputations incorrectes estimé sur les manquants		11,00	15,31	12,56	11,16	10,85		11,56	10,59	10,07
% imputations incorrectes estimé total		1,46	2,03	1,66	1,48	1,44		1,53	1,40	1,33

En regardant les deux tableaux 8 et 9, nous constatons que les deux meilleures méthodes d'imputation sont le Random Forest (RF) et l'Arbre de décision (AD). Effectivement, la méthode du Random Forest sur les relations (ML) laisse des Types de ménages manquants; et en comparaison individuelle, les méthodes Statistique Vaud (Stat-VD) et la Régression multinomiale (RM) ont des résultats un peu moins bons que l'Arbre de décision (AD) et le Random forest (RF), bien que la Régression multinomiale soit proche en distribution.

6.3 Estimation de la qualité

Pour effectuer les tests de performance décrits dans le chapitre 6.2, nous avons utilisé les données du Relevé Structurel. Cependant, si la typologie est publiée au niveau communal, les données du Relevé Structurel ne sont pas assez précises pour la majorité des communes afin de procéder à une estimation de la qualité, étant donné qu'il s'agit d'une enquête par échantillonnage dont le plan de sondage ne peut pas couvrir ces exigences. De plus, pendant la production de la variable Type de ménage, nous n'avons pas accès à ces données et il est important de pouvoir évaluer la qualité des imputations afin que l'utilisateur puisse juger de la pertinence des résultats pour ses besoins. Dans ce chapitre, nous utilisons les estimateurs de qualité fournis par les algorithmes Random Forest et Arbre de décision et vérifions leur fiabilité avec les comparaisons individuelles à l'aide du Relevé Structurel.

Les estimateurs de qualité fournis par les algorithmes, fonctionnent de la manière suivante :

- Random Forest : chacun des 1 000 arbres choisit un type de ménage. Le type de ménage majoritaire est imputé et son estimation d'erreur d'attribution du type de ménage est la proportion d'arbres n'ayant pas voté pour ce type. Par exemple, pour un ménage dont le type est manquant, il pourrait y avoir 200 arbres qui imputeraient « Parent seul avec enfant », 350 qui imputeraient « Couple en union libre » et 450 qui imputeraient « Autre type de ménage ». Dans ce cas, « Autre type de ménage » est imputé et l'erreur est estimée à 550/1 000 = 0,55.

- Arbre de décision : quand l'arbre se construit sur la base de l'échantillon d'apprentissage, ce dernier est partitionné à l'aide de branches qui sont définies par des règles discriminantes jusqu'à arriver au bout de chaque ramification, appelé feuille. A la fin du processus, les ménages de l'échantillon d'apprentissage sont répartis dans l'ensemble des feuilles de l'arbre (chaque ménage n'appartient qu'à une seule feuille). Lorsque l'arbre est appliqué à un ménage à imputer, le type de ménage majoritaire dans la feuille lui est attribué. L'erreur d'attribution du type de ménage est estimée par la proportion de ménages de l'échantillon d'apprentissage qui n'est

pas de ce type. Supposons par exemple que, dans la feuille, il y avait 75 ménages de l'échantillon d'apprentissage avec les types suivants : 15 « Parent avec enfant », 20 « Couples en union libre » et 40 « Autre type de ménage ». Dans ce cas, le type imputé est « Autre type de ménage » et l'erreur est estimée à 35/75 = 0,47.

Le tableau 10 montre les différences entre les estimations d'erreur calculées en comparant individuellement les imputations avec le Relevé Structurel (point 3 du chapitre 6.2) et celles des estimateurs de qualité des méthodes Arbre de décision ou Random Forest, sur la population totale et sur les imputations.

Tableau 10 : Comparaison des estimations d'erreur du Relevé Structurel avec les résultats des estimateurs de qualités des algorithmes pour la typologie 25 ans de 2018 et 2019

	Méthode	Population totale	Erreur selon l'estimateur de qualité	Erreur selon l'estimation du RS	Diff	Imputations	Erreur selon l'estimateur de qualité	Erreur selon l'estimation du RS	Diff
2018	RF		1,5226	1,3676	0,1550		11,1181	9,9862	1,1319
	AD		1,5389	1,4451	0,0938		11,2369	10,5525	0,6844
2019	RF		1,4143	1,4381	0,0238		10,6710	10,8512	0,1802
	AD		1,4784	1,4788	0,0004		11,1545	11,1580	0,0035

On remarque que l'estimateur de qualité de l'Arbre de décision est plus proche de celui fournit par le Relevé Structurel que l'erreur calculée pour le Random Forest.

Dans le tableau 11, nous vérifions la précision des estimateurs de qualité par type de ménage. La colonne « N_imput » contient le nombre d'imputations du Type par la méthode d'imputation. La colonne « %Erreur selon AD, RS ou RF » correspond à l'estimation de l'erreur, soit donnée par la méthode (AD ou RF), soit donnée par le Relevé Structurel (RS). La colonne « N_RS » correspond au nombre de ménages dans le Relevé Structurel qui ont permis de faire l'estimation de l'erreur. Si ce chiffre est petit, le « %Erreur selon RS » n'est pas fiable, car son estimation est faite sur trop peu de données. La colonne « Diff des % » nous donne la différence entre l'estimation de l'erreur selon la méthode et l'estimation de l'erreur selon le Relevé Structurel.

Tableau 11 : Comparaison des estimateurs de qualité avec celui du Relevé Structurel par catégorie de Type de ménage (HHTyp)

		Arbre de décision					Random Forest				
25 ans	HHTyp	N_imput	%Erreur selon AD	%Erreur selon RS	N_RS	Diff des %	N_imput	%Erreur selon RF	%Erreur selon RS	N_RS	Diff des %
2018	PE	27 957	10,19%	8,51%	131	1,68%	27 772	17,22%	7,54%	128	9,68%
	A	127 263	16,98%	18,67%	701	-1,69%	121 363	15,96%	12,11%	497	3,85%
	M	21 379	11,71%	12,07%	101	-0,36%	19 450	18,17%	6,37%	51	11,80%
	UL	216 008	10,32%	10,96%	1466	-0,64%	222 862	6,71%	11,34%	1540	4,63%
	MS	711	3,82%	25,00%	7	-21,18%	4 572	39,81%	47,27%	101	7,46%
	ME	85 258	6,22%	4,29%	262	1,93%	81 908	8,21%	3,10%	195	5,11%
	ULE	34 910	8,32%	10,08%	214	-1,76%	35 964	16,11%	10,14%	217	5,97%
	MSE	849	37,64%	50,00%	13	-12,36%	444	47,63%	12,50%	2	35,13%
2019	PE	27 240	13,10%	12,04%	172	1,06%	26 570	16,72%	10,68%	146	6,04%
	A	120 532	14,72%	16,24%	685	1,52%	119 120	15,66%	15,11%	644	0,55%
	M	19 824	12,94%	12,26%	97	0,68%	17 710	18,83%	6,72%	46	12,11%
	UL	223 716	11,48%	11,47%	1661	0,01%	224 669	6,19%	11,46%	1656	5,27%
	MS	391	22,02%	40,00%	6	17,98%	3 750	43,20%	41,77%	64	1,43%
	ME	78 557	3,79%	4,66%	285	0,87%	77 238	7,83%	3,97%	240	3,86%
	ULE	33 938	10,68%	9,02%	194	1,66%	34 550	16,01%	9,13%	203	6,88%
	MSE	64	0,00%	0,00%	0	0,00%	655	41,22%	60,00%	9	18,78%

Les différences entre l'estimation du Relevé Structurel et l'estimation de la méthode (AD ou RF) qui dépassent 2% et dont la colonne « N_RS » est supérieure à 30 unités sont surlignées en jaune. Nous constatons que l'estimation d'erreur de l'Arbre de décision (AD) est nettement plus proche de celle du Relevé Structurel.

Le tableau 11 nous montre que l'estimation de l'erreur par le Random Forest est majoritairement éloignée de l'estimation faite par le Relevé Structurel. Par exemple, l'estimation de l'erreur du type Couple en union libre (UL) est sous-estimée par le Random Forest de 4 à 5%. Etant donné que ce type est le plus imputé (220 000 ménages), environ 10 000 de ces ménages sont mal imputés sans être repérés par l'estimateur de qualité du Random Forest. Il y a aussi beaucoup de différences dans les autres catégories comme par exemple en 2018, l'erreur sur le type Parent seul avec enfant (PE) est estimée à 17,51%, alors que le Relevé Structurel l'estime à 7,74% (surestimation de l'erreur d'environ 10%).

En conclusion, le Random Forest est légèrement meilleur pour imputer les types de ménage manquants, par contre son estimateur de qualité n'est pas fiable. L'Arbre de décision quant à lui, est très légèrement moins bon au niveau performance, mais l'erreur d'attribution du type de ménage fournit une appréciation acceptable de la qualité de ses imputations.

6.4 Choix de la méthode

Les résultats de toutes les méthodes sont très proches. En effet, nous pouvons estimer que sur l'entier de la table ménage, elles ont toutes une erreur inférieure à 2%. Cependant, les méthodes de Statistique

Vaud et Régression Multinomiale donnent des résultats légèrement moins bons au niveau de la comparaison individuelle. La méthode Machine Learning (qui impute des relations plutôt que directement des types de ménage), bien qu'ayant de bons résultats, laisse des ménages sans valeur de Type de ménage.

Le choix final consiste à départager les méthodes Arbre de décision et Random Forest. Selon les résultats du chapitre précédent, la méthode de **l'Arbre de décision** a été choisie pour imputer les types de ménage. Cette méthode est choisie pour la production de la Typologie des ménages, car, bien que son résultat en comparaison individuelle avec le Relevé Structurel soit légèrement moins bon que le Random forest, elle nous permet d'estimer la qualité des imputations de manière acceptable pendant la production.

Dans la suite du document, nous décrivons la méthode Arbre de décision en détail.

7 Production de la variable HHTyp

Dans ce chapitre, toute la procédure de création de la variable Type de ménage (HHTyp) est décrite. Nous commençons par la production des tables : table **Personne**, table **Relation** et table **Typologie**. Ensuite, nous détaillons la méthode d'imputation, l'échantillon d'apprentissage et les différentes variables du modèle.

7.1 Table Personne

Nous partons de la table de l'effectif STATPOP appariée avec la table des ménages STATPOP. Cette table est filtrée pour ne garder que la population résidante permanente au domicile principal en ménage privé.

```
populationtype='1'
typeOfHousehold='1'
```

7.2 Table Relation

Sur la base de la table **Personne**, nous préparons la table **Relation**. Pour chaque ménage, nous créons toutes les combinaisons de deux personnes (Personne A et Personne B). Nous ajoutons leurs caractéristiques STATPOP nécessaires. Pour la personne B, toutes ces caractéristiques sont préfixées par « B_ ».

Nous y ajoutons la variable Relation qui représente la relation entre la personne A et la personne B.

VN	Age	Sexe	…	B_VN	B_Age	B_Sexe	…	Relation

Les valeurs de la variable Relation sont reprises des différentes « sources » dans l'ordre ci-dessous :

1. INFOSTAR : informations contenues dans la table de l'effectif STATPOP
2. ORDIPRO : informations contenues dans une table spécifique de STATPOP
3. RS : table des relations des enquêtes 2010 à année-1
4. Algorithme déterministe : basé sur les informations contenues dans l'effectif STATPOP
5. Relations éloignées : calculées sur la base des relations définies par les points 1 à 4, nous appliquons un algorithme selon des déductions logiques et déterministes qui ajoute des relations éloignées telles qu'oncle, tante, cousin, grand-parent, arrière-grand-parent, beau-frère, etc.

Les relations obtenues selon les différentes sources se trouvent dans le tableau 12.

Tableau 12 : Relations possibles données à une paire selon la source

	INFOSTAR	ORDIPRO	RS	Algorithme déterministe	Relations éloignées
Epoux/se	X	X	X	X	
Partenaire enregistré-e	X	X	X	X	
Concubin-e		X	X	X[9]	
Père/mère	X	X	X		
Fils/Fille	X	X	X		
Frère ou demi-frère/sœur ou demi-sœur	X	X	X		
Beau-père/belle-mère (époux/se du parent)	X		X		X
Beau-fils/belle-fille (enfant de l'époux/se)	X		X		X
Grand-père/grand-mère	X	X	X		X
Petit-fils/petite-fille	X	X	X		X
Apparenté[10]		X	X		X
Non-Apparenté[11]	X		X		X

Nous ajoutons la relation seulement si celle-ci est vide. Autrement dit, si INFOSTAR et le Relevé Structurel possèdent tous deux l'information d'une même « paire », INFOSTAR va d'abord remplir la variable Relation, et ensuite, comme l'information est déjà remplie, le Relevé Structurel ne l'écrase pas et celle d'INFOSTAR est conservée.

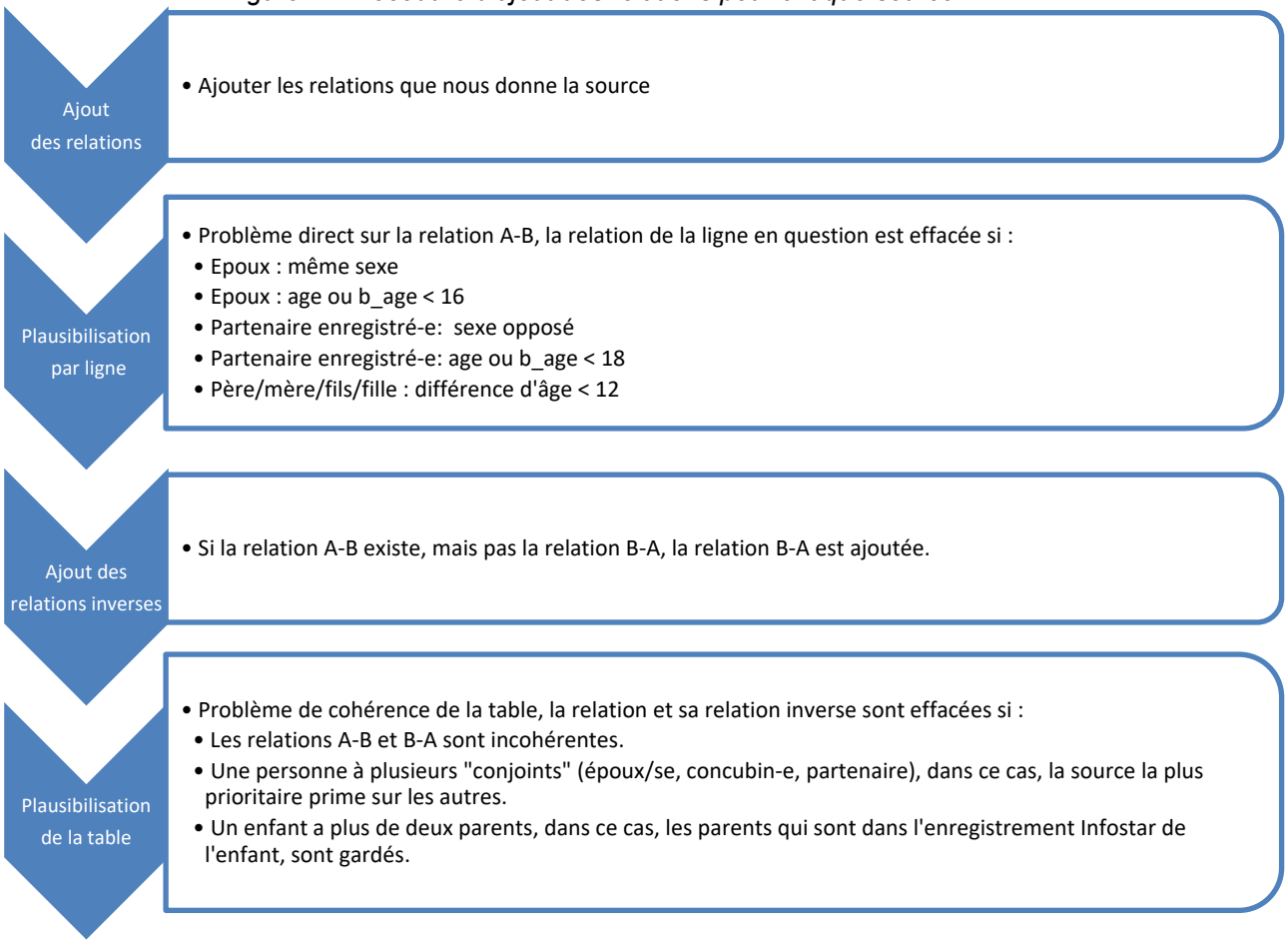
Pour chaque source, après avoir ajouté les relations, nous effectuons des traitements de consolidations des données avant de passer à la source suivante. Ce processus est décrit dans la figure 2.

9 Pour les couples non-mariés ensemble ayant un enfant commun dans le même ménage.

10 Dans la catégorie « apparenté » sont inclus notamment les oncles, tantes, cousins.

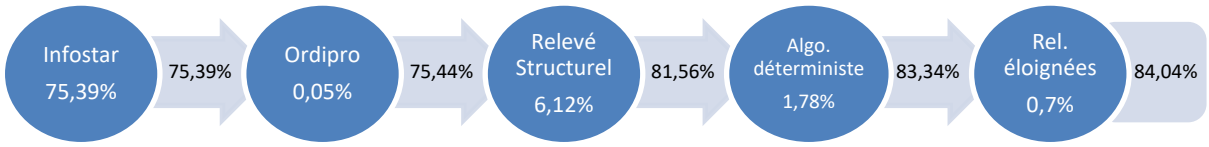
11 Dans la catégorie « non-apparenté » sont inclus notamment les beaux-parents (parents de l'époux/se), les beaux-enfants (époux/se des enfants), les enfants du/de la concubin-e, le/la concubin-e du parent.

Figure 2 : Procédure d'ajout des relations pour chaque source[12]



Pour l'année 2018, les sources ont contribué à renseigner les relations, selon les pourcentages décrits dans la figure 3.

Figure 3 : Pourcentage des relations trouvées selon la source



Il reste 15,96% de relations manquantes à la fin de la reprise des relations sûres des différentes sources à disposition.

[12] Plausibilisation de la table : les incohérences AB-BA sont rares (50 cas sur 16 000 000) et proviennent des données sources. Avec l'introduction du « Mariage pour tous », les plausibilisations seront adaptées.

7.3 Table Typologie

Pour attribuer un type de ménage, nous utilisons le programme développé pour le Relevé Structurel. Si le ménage contient toutes les relations, alors un type de ménage (HHTyp) y est associé. Le tableau 13 donne la répartition des Types de ménages basés sur les relations sûres pour l'année 2018.

Tableau 13 : Répartition des types de ménage basés sur les relations sûres pour 2018

Type de ménage (HHTyp)	Répartition en %	Répartition absolue
Manquant	13,78	518 512
Ménage d'une personne	35,63	1 340 644
Parent seul avec au moins un enfant de moins de 25 ans	4,46	167 930
Autre ménage de plusieurs personnes	4,61	173 400
Couple marié sans enfant	18,88	710 155
Couple en union libre sans enfant	1,76	66 201
Couple de même sexe sans enfant	0,23	8 815
Couple marié avec au moins un enfant de moins de 25 ans	18,61	700 055
Couple en union libre avec au moins un enfant de moins de 25 ans	2,02	76 147
Couple de même sexe avec au moins un enfant de moins de 25 ans	0,01	493
TOTAL	**100**	**3 762 352**

Pour 2018, 13,78% des ménages ont au moins une relation du ménage manquante, nous le voyons car le type (HHTyp) ne peut pas être renseigné.

7.4 Imputations

Dans ce chapitre, nous décrivons la méthode d'imputation Arbre de décision, qui impute les Types de ménage manquants de la table **Typologie**. Nous définissions ensuite les variables du modèle et décrivons l'échantillon d'apprentissage. Pour finir, nous décrivons la procédure spécifique développée pour les couples de même sexe avec enfants.

7.4.1 Description technique de la méthode Arbre de décision

Un arbre de décision est un modèle statistique de classification. Une liste de variables, catégorielles ou continues, ainsi qu'un ensemble d'observations dont le résultat est connu, appelé échantillon d'apprentissage, lui sont fournis.

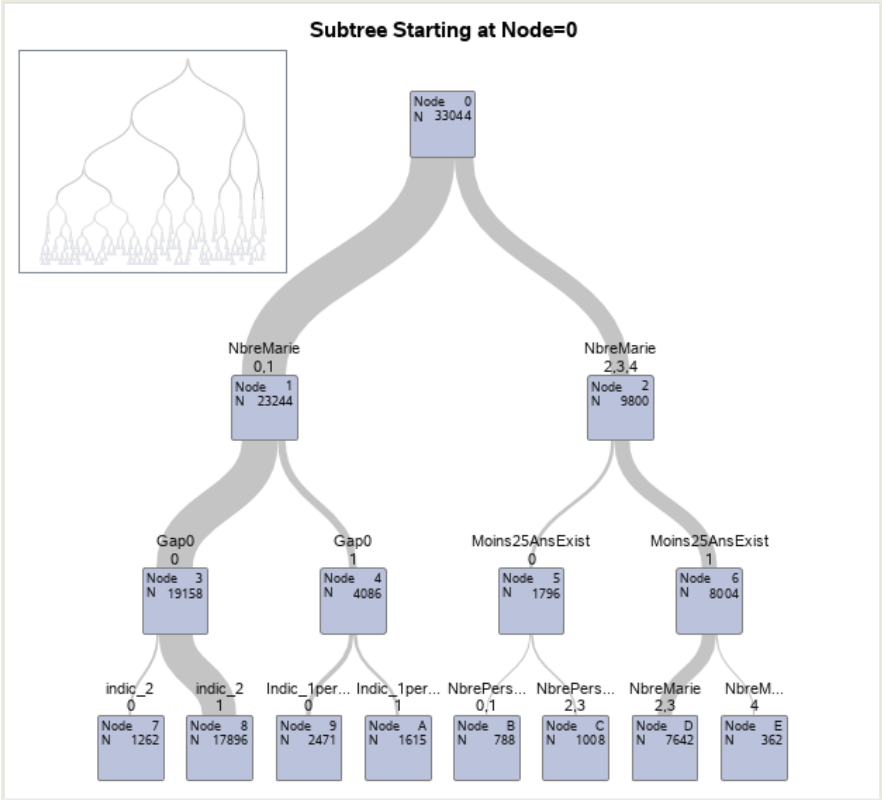
L'arbre se construit comme suit[13] :
Nous partons de la racine avec toutes les données de l'échantillon d'apprentissage, puis les variables sont testées l'une après l'autre jusqu'à trouver celle dont une modalité ou une valeur discriminante (pour les variables continues) sépare au mieux les différents types de ménage entre les deux nœuds qui en résultent. Ainsi, les données d'apprentissage sont partitionnées selon la modalité/valeur discriminante. Ensuite, pour chaque nouveau nœud et la partition des données d'apprentissage qui lui correspond, la même procédure est opérée jusqu'à l'obtention de l'arbre complet.

Pour tester si une branche peut être séparée, nous utilisons l'option « Grow CHAID ». Cette option effectue des séparations basées sur des tests statistiques (Chi-carré). Pour chaque variable disponible, nous testons H0 : « Indépendance entre la réponse (Type de ménage) et la variable testée pour la

[13] Dans SAS, l'Arbre de décision est créé via la procédure « PROC HPSPLIT », sa documentation SAS : The HPSPLIT Procedure (sas.com).

branche ». Nous considérons que les tests sont indépendants avec un niveau[14] α=0,02. Si H0 est rejetée par au moins un test, une branche peut être créée et sinon la procédure de partition s'arrête. Une illustration des premiers embranchements de l'arbre créé avec l'échantillon d'apprentissage pour l'année 2018 se trouve à la figure 4.

Figure 4 : Début de l'arbre de décision pour 2018



La racine de l'arbre (Node 0) contient les 33 044 ménages de l'échantillon d'apprentissage. Le premier nœud sépare les ménages selon le nombre de mariés dans le ménage. Les 23 244 ménages avec 0 ou 1 personne mariée se retrouvent dans le nœud 1 (Node 1) alors que les 9 800 autres ménages se retrouvent dans le nœud 2 (Node 2), ils contiennent 2 personnes mariées ou plus (NbreMarie=2, 3 ou 4). De chaque côté de l'arbre, les branches continuent de séparer en fonction des variables explicatives.

Ensuite, pour éviter un « overfitting » sur les données d'apprentissage, un algorithme de pruning est appliqué. Il s'agit de l'algorithme « Cost-Complexity »[15] dont le but est de diminuer le nombre de feuilles jusqu'à trouver un équilibre entre éviter l'« overfitting » et garder le but d'imputation. Nous partons de la notion de coût,

$$CC(T) = R(T) + \alpha\,|T|$$

où

T : un sous-arbre
$CC(T)$: le coût du sous-arbre
$R(T)$: le pourcentage d'erreur sur l'échantillon d'apprentissage avec ce sous-arbre
α : le coût par feuille
$|T|$: le nombre de feuilles du sous-arbre.

---

[14] De manière générale, les paramètres (mtry, nombre d'arbres, le niveau, les options des méthodes, …) ont été optimisés sur la base des résultats de différents tests de comparaison avec le Relevé Structurel.

[15] Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

Nous commençons avec $\alpha = 0$, dans ce cas, le sous-arbre qui minimise le coût est l'arbre complet, disons T_0. Puis, la valeur de α est augmentée, jusqu'à ce qu'un autre sous-arbre ait le coût plus bas que T_0, disons T_1 et appelons le α limite, α_1. Ainsi nous pouvons construire une suite finie de $(T_0, 0)$,…, (T_m, α_m) puisque le nombre de sous-arbres est fini. Quand α tend vers l'infini, le T_m qui minimise le coût est la racine sans aucune branche (le coût de chaque feuille est devenu trop grand). La méthode exacte pour trouver les T_i est décrite dans la publication de Breiman et al.[16] et s'appelle « Weakest-link pruning ». Il reste à trouver le α optimal.

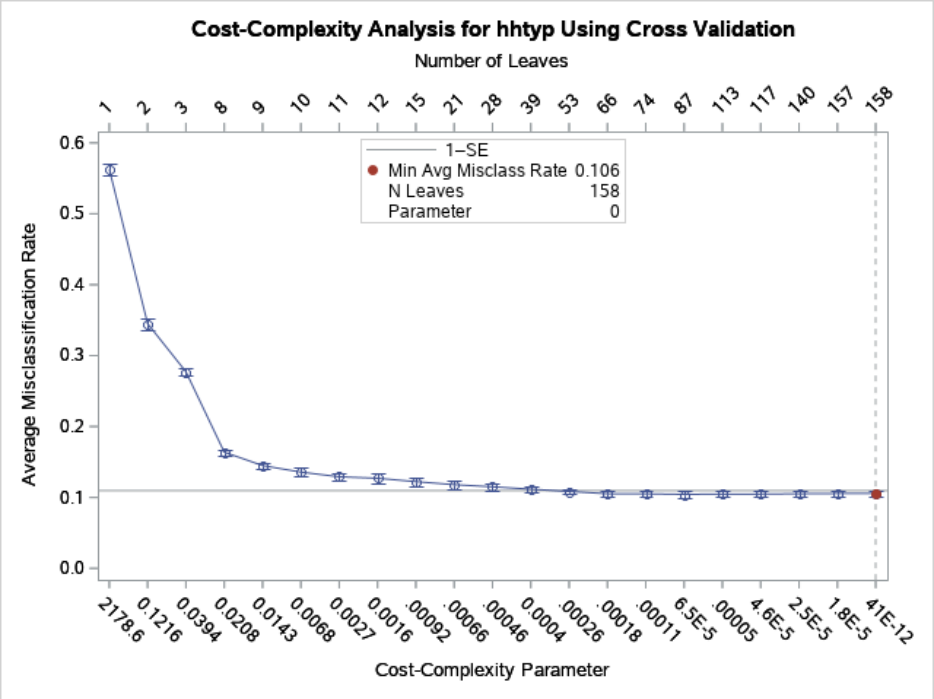
Nous commençons par définir $\beta_i = \sqrt{\alpha_i \times \alpha_{i+1}}$ pour $\mathrm{i} = 0, \ldots, \mathrm{m} - 1$. A partir de ce point, l'aléatoire entre en jeu. L'échantillon d'apprentissage est divisé aléatoirement en k sous-ensembles[17] de taille environ égale. Pour $\mathrm{j} = 1, \ldots, \mathrm{k}$, le sous-ensemble j est mis de côté. Les autres sous-ensembles permettent la construction d'un nouvel arbre complet. Ensuite, pour chaque coût par feuille, $\alpha = \beta_i$ (i=0,…,m-1), un sous-arbre T_{ij}, qui minimise le coût $CC(T_{ij})$, est sélectionné. Puis, cet arbre est appliqué au $j^{\text{ème}}$ sous-ensemble et l'erreur quadratique moyenne, ASE_{ij}, est calculé. Une fois fait sur les k sous-ensembles, nous calculons l'erreur moyenne pour chaque i.

$$ASE_i = \frac{1}{k}\sum_{j=1}^{k} ASE_{ij}$$

Finalement, le i qui minimise ASE_i, détermine l'arbre finale T_i (sous-arbre de l'arbre complet correspondant au coût par feuille α_i).

Dans la figure 5, fournie par la procédure « hpsplit », « Cost-Complexity Parameter » correspond au β_i, « Average Missclassification Rate » correspond au ASE_i et « Number of Leaves » nous indique le nombre de feuilles correspondant au β_i. La barre verticale nous indique où ce minima est atteint.

Figure 5 : Pruning pour l'année 2018, une prévision d'erreur moyenne à 10,6 % dans les imputations

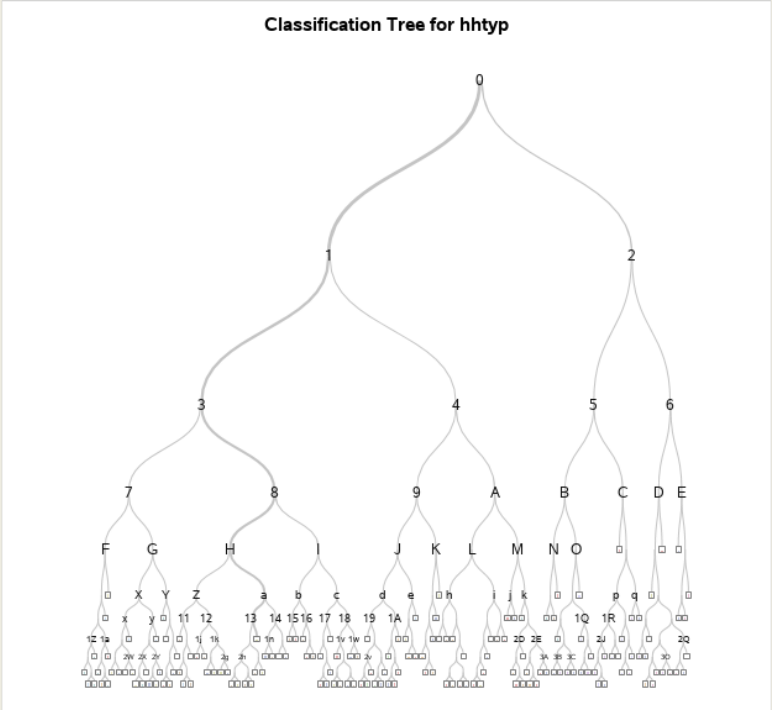


16 Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

17 Par défaut k=10.

Pour 2018, l'arbre entier avait 199 feuilles et après l'algorithme de pruning, c'est le sous-arbre avec 159 feuilles qui a été retenu. L'illustration de l'arbre final pour l'année 2018 se trouve à la figure 6.

Figure 6 : Arbre 2018, après pruning



7.4.2 Variables du modèles

Nous définissons préalablement une notion de « Gap ». Le but des Gap est de séparer les différentes générations au sein d'un ménage afin de transmettre au modèle une description de chaque génération du ménage.

On commence par trier les personnes du ménage de la plus jeune à la plus âgée et chaque personne du ménage sera traitée l'une après l'autre en partant de la personne la plus jeune. Si la personne la plus jeune a moins de 25 ans, elle va dans le Gap0, si elle a 25 ans ou plus, elle va dans Gap1. Ensuite, nous regardons la différence d'âge entre la personne précédente et la personne actuelle, si cette différence est égale ou plus petite que 15 ans, alors elle va dans le même Gap que la personne précédente, sinon elle va dans le Gap suivant. Pour chaque ménage, il y a autant de Gap que de générations. Une fois que l'ensemble des personnes du ménage est traité, si le Gap0 est non-vide et que le Gap1 est vide, alors les personnes du Gap0 sont transférées au Gap1, ainsi les « adultes » sont dans le Gap1 (voir tableau 14).

Tableau 14 : Exemple d'attribution de la variable Gap pour le ménage M1

Ménage	Personne	Age	Gap
M1	P1	80	2
M1	P2	50	1
M1	P3	48	1
M1	P4	26	0
M1	P5	23	0

Exemple d'une famille avec deux enfants, deux parents et un grand-parent. Premier enfant < 25 ans (Gap0), deuxième enfant, 26-23=3 <= 15 (Gap0), premier parent, 48-26=22 > 15 (Gap1), deuxième parent 50-48=2 <= 15 (Gap1), grand-parent 80-50=30 > 15 (Gap2). Si dans ce ménage, il n'y avait pas l'enfant de 23 ans, P4 serait dans le Gap1, P3 et P2 dans le Gap2 et P1 dans le Gap3.

L'idée est que les « enfants » soient dans le Gap0, que les « parents » dans le Gap1 et les « grands-parents » dans le Gap2. Dans le cas d'un jeune couple avec enfant, l'enfant sera en Gap0 et le couple en Gap1 (même si les parents ont moins de 25 ans).

Le choix final des variables explicatives utilisées dans le modèle d'imputation sont listées dans le tableau 15. Pour construire ces variables, nous apparions les données STATPOP avec :

1. La Statistique des bâtiments et des logements : nombre de pièces dans le logement
2. Les Niveaux géographiques de la Suisse : urbain-rural

Tableau 15 : Variables utilisées dans le modèle d'imputation

Variable	**Catégo-riel**	**Explication**
householdByNa-tioPerm	X	1 : Ménage suisse: membres tous suisses 2 : Ménage suisse-étranger: au moins un suisse et un étranger 3 : Ménage étranger: tous étrangers
foreignHouse-holdPerm	X	1 : Ménage suisse-étranger (au moins un Suisse et un étranger) avec tous les étrangers du même pays 2 : Ménage suisse-étranger (au moins un Suisse et un étranger) avec des étrangers de différents pays 3 : Ménage suisse-étranger (au moins un Suisse et un étranger) avec tous les étrangers sans indication de nationalité 4 : Ménage étranger (tous étrangers) avec tous les étrangers du même pays 5 : Ménage étranger (tous étrangers) avec des étrangers de différents pays 6 : Ménage étranger (tous étrangers) avec tous les étrangers sans indication de nationalité -8 : Ménage suisse (tous suisses) -9 : Ménage avec au moins un étranger sans indication de nationalité et au moins un étranger avec indication de nationalité. Regroupés en (1, 3), (4, 6), (2, -9), (5) et (-8)
NbreFemmePlus25	X	Nombre de femmes de plus de 25 ans (0, 1, 2+)
NbreHommePlus25	X	Nombre d'hommes de plus de 25 ans (0, 1, 2+)
Moins25AnsExist	X	Il existe une personne de moins de 25 ans dans le ménage (0, 1)
nb_famille_4[18]	X	Nombre de noms de famille différents dans le ménage. (1, 2, 3, 4+)
NbrePart	X	Nombre de personnes avec l'état civil = partenariat enregistré (0, 1, 2+)
NbreMarie	X	Nombre de personnes avec état civil = marié (0,1, 2, 3, 4+)
indic_2	X	Le ménage contient exactement 2 personnes (0, 1)
MaxDiffAge		La plus grande différence d'âge du ménage
Gap0	X	Il existe une personne dans le Gap0 (0, 1)

[18] Si nom1= « Smith », nom2= « Granny » et nom3= « Granny-Smith », alors nb_famille = 2, parce que nom1 et nom2 sont contenus dans nom3.

Variable	Catégo-riel	Explication
Gap2	X	Il existe une personne dans le Gap2 (0, 1)
Indic_1pers_gap1	X	Il existe exactement une personne dans le Gap1 (0, 1)
CatNbreHommeGap1	X	Nombre d'hommes dans le Gap1 (0, 1, 2, 3+)
CatNbreFemmeGap1	X	Nombre de femmes dans le Gap1 (0, 1, 2, 3+)
NbrePersGap0	X	Nombre de personnes dans le Gap0 (0, 1, 2+)
NbrePersGap2	X	Nombre de personnes dans le Gap2 (0, 1, 2, 3+)
NbreMarieGap0	X	Nombre de personnes qui ont l'état civil = marié dans le Gap0 (0, 1, 2+)
NbreMarieGap3	X	Nombre de personnes qui ont l'état civil = marié dans le Gap3 (0, 1, 2+)
MoyAgeGap0		Moyenne d'âge des personnes qui sont dans le Gap0
MoyAgeGap1		Moyenne d'âge des personnes qui sont dans le Gap1
MoyAgeGap2		Moyenne d'âge des personnes qui sont dans le Gap2
NbreGap	X	Nombre de Gap non vide (1, 2, 3+)
SigmaAge		Ecart-type de la variable âge dans le ménage.
SymAge		Coefficient d'asymétrie (troisième moment) de la variable âge. Si elle n'est pas calculable (SigmaAge=0), alors 2,82[19]
PieceParPers		Nombre de pièces dans le logement divisé par le nombre de personnes du ménage
PieceParPers_2	X	S'il y a exactement 2 personnes dans le ménage, indique le nombre de pièces du logement. Sinon : 0 (0, 1, 2, 3, 4, 5, …)
Taille		Nombre d'habitants dans la commune
Urbain	X	1 : Urbain 2 : Rural 3 : Intermédiaire

7.4.3 Echantillon d'apprentissage

Un échantillon d'apprentissage est un ensemble de ménages dont le type est connu. Il permet de construire l'arbre qui sera utilisé pour imputer les données manquantes. Pour choisir notre échantillon d'apprentissage, nous essayons de nous rapprocher au mieux des données que l'on veut imputer.

Nous refaisons le processus de préparation des données de l'année précédente en y ajoutant le Relevé Structurel de l'année précédente. Et nous obtenons la table **Relation**, puis la table **Typologie** avec une

[19] La valeur 2,82 est ici pour assurer une certaine continuité de la variable. Si SigmaAge=0, le calcule de SymAge contient un 0/0 qui n'est pas défini. Cette valeur 2,82 a été choisie car c'est la valeur de SymAge pour une différence de 1 an entre deux personnes.

colonne supplémentaire « Contribution RS », qui contient l'information dans les cas où le Relevé Structurel de l'année précédente a contribué à obtenir le type de ménage. Ces ménages qui étaient manquants l'année précédente, mais ne le sont plus grâce au Relevé Structurel de l'année précédente, constitueront l'échantillon d'apprentissage pour cette année (environ 30 000 ménages), voir figure 7.

Figure 7 : Définition de l'échantillon d'apprentissage

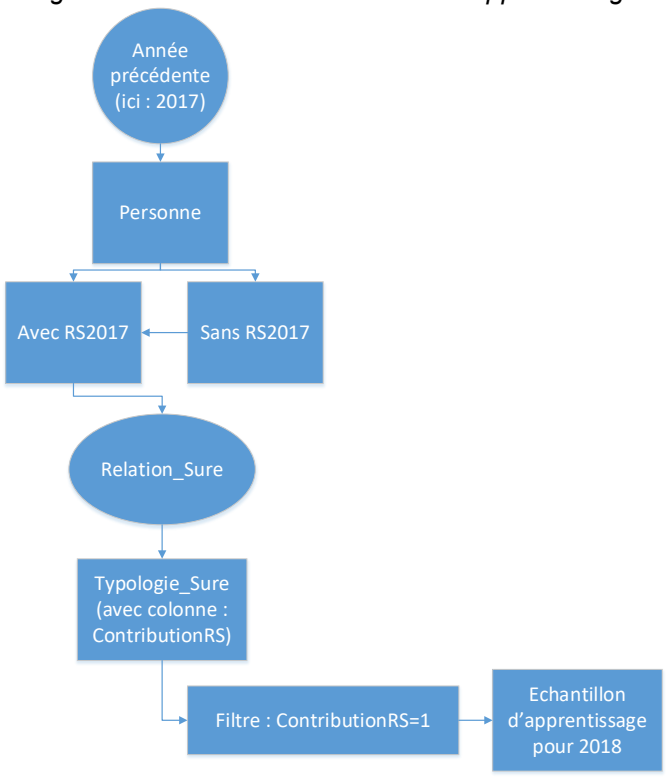
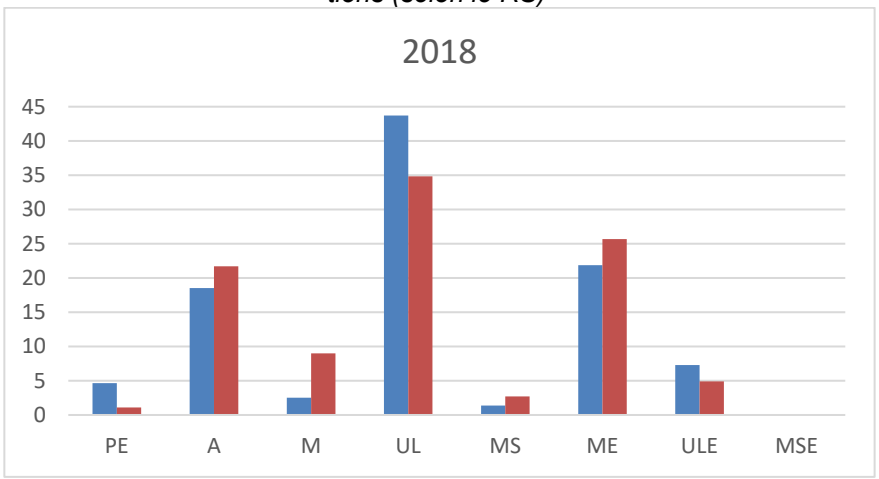


Figure 8 : Répartition de l'échantillon d'apprentissage comparée à la distribution attendue des imputations (selon le RS)



Le Relevé Structurel est basé sur un tirage aléatoire afin de récolter le type de ménage d'un échantillon (de ménages). L'intersection entre l'échantillon des ménages du Relevé Structurel et la sous-population des ménages à imputer peut être vu comme un domaine d'analyse à l'intérieur de la population du Relevé Structurel. Les distributions de cette intersection et des imputations attendues sont montrées à

la figure 8. De plus, cette source est primordiale, car, dans les relations sûres, elle donne presque intégralement les catégories Couple en union libre (UL) et Autre type de ménage (A). En effet, les colocations et les couples en union libre ne sont que rarement renseignés dans les sources plus prioritaires.

7.4.4 Problématique du type : Couple de même sexe avec enfant

Après différents tests de qualité effectués sur les données finales, il a été découvert que le type de ménage « Couple de même sexe avec enfant de moins de 18/25 ans » était imputé de manière instable et fluctuait au fil des années (tableau 16).

Tableau 16 : Nombre de ménages avec le type « Couple de même sexe avec enfant de moins de 25/18 ans » après imputation

Année	Imputation (25 ans)	Imputation (18 ans)	Total (25 ans)	Total (18 ans)
2017	81	58	510	419
2018	876	620	1 369	1 039
2019	93	136	643	617
2020	0	414	619	968

La catégorie « Couple de même sexe avec enfant de moins de 18/25 ans » étant peu représentée, nous avons constaté que l'algorithme avait tendance à la regrouper avec la catégorie « Parent seul avec enfant de moins de 18/25 ans ». Ce groupement est particulièrement visible dans la matrice de confusion de l'échantillon d'apprentissage pour l'année 2020 (voir tableau 17).

Tableau 17 : Matrice de confusion de l'arbre de décision sur l'échantillon d'apprentissage pour l'année 2020

AD	HHTyp prédit							
HHTyp connu	**PE**	**A**	**M**	**UL**	**MS**	**ME**	**ULE**	**MSE**
PE	1 086	7	0	0	0	8	126	0
A	75	3 827	34	1 621	0	162	33	0
M	0	39	631	0	0	49	0	0
UL	23	255	31	13 085	0	2	20	0
MS	1	419	0	0	15	0	1	0
ME	2	21	3	0	0	5 260	0	0
ULE	15	6	0	3	0	44	2 084	0
MSE	15	0	0	0	0	0	1	0

Les catégories « Parent seul avec enfant de moins de 18/25 ans » (PE) et « Couple de même sexe avec enfant de moins de 18/25 ans » (MSE) se confondent, car il arrive que dans la catégorie « PE », il y ait une personne non-apparentée qui habite avec le parent seul et l'enfant, ce qui pourrait se confondre avec un couple de même sexe avec enfant. De plus, le type « MSE » est très peu représenté et il se retrouve absorbé avec les ménages du type « Parent seul avec enfant de moins de 18/25 ans » (PE).

Pour résoudre ce problème, nous avons décidé de créer la catégorie provisoire « Parent seul ou Couple de même sexe avec enfant de moins de 18/25 ans » dans l'arbre principal. Cette catégorie est ensuite séparée en deux types (PE et MSE) via une régression logistique.

7.4.4.1 Echantillon d'apprentissage pour la régression logistique

Pour simplifier la suite du texte, nous allons abréger :

- PE : « Parent seul avec enfant de moins de 18/25 ans »,
- MSE : « Couple de même sexe avec enfant de moins de 18/25 ans »,
- PMSE : « Parent seul ou Couple de même sexe, avec enfant de moins de 18/25 ans ».

Après analyse, nous avons constaté que dans la catégorie « PE », il y a deux types distincts : ceux qui ressemblent à la catégorie MSE et ceux qui ne lui ressemblent pas. Ceci se traduit dans l'arbre de décision par des feuilles PMSE dans lesquels il n'y a quasiment que des PE (appelé PMSE_PE) et des feuilles PMSE dans lesquels les catégories PE et MSE sont bien représentées (appelé PMSE_PMSE). Il est aussi possible d'avoir une feuille PMSE_PMSE dont les MSE sont majoritaires.

Pour avoir plus de données de la catégorie MSE (environ 20 ménages par année dans l'échantillon d'apprentissage), nous appliquons l'arbre de décision aux échantillons d'apprentissage des deux années précédentes. Par exemple, pour imputer l'année 2019, l'arbre est créé avec l'échantillon d'apprentissage pour 2019 (qui contient des ménages du Relevé Structurel de 2018). Cet arbre est ensuite appliqué aux échantillons d'apprentissage pour les années 2017, 2018 et 2019. Nous déterminons ensuite si les feuilles PMSE sont du type PMSE_PE ou PMSE_PMSE. Une feuille est définie PMSE_PMSE selon les deux règles suivantes (exemple au tableau 18) :

1. En choisissant les feuilles contenant le plus de ménages de type MSE, jusqu'à avoir au moins 80% des ménages MSE ;
2. S'il y a plus 20% de MSE dans la feuille.

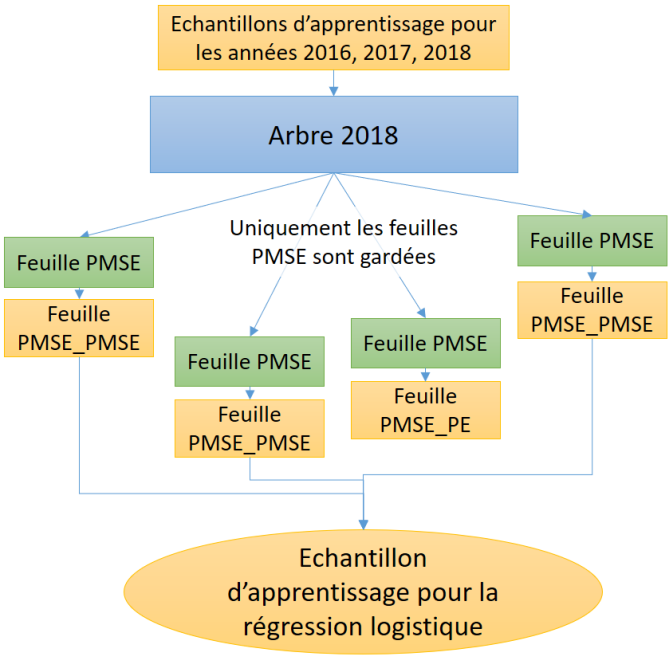
Tableau 18 : Feuilles PMSE pour l'année 2019 (typologie 25 ans)

Numéro de feuille PMSE	Ménages pas du type MSE	Ménage du type MSE	Cumul des ménages MSE	Jusqu'à avoir 80% des ménages MSE	> 20% de MSE	Feuille PMSE_PMSE
72	101	40	40	1	1	1
86	3	8	48	1	1	1
84	8	5	53	0	1	1
38	107	2	55	0	0	0
21	2 064	1	56	0	0	0
10	13	0	56	0	0	0
74	33	0	56	0	0	0
Nombre total des ménages	**2 329**	**56**	-	-	-	-

Les 80% des 56 ménages MSE donnent 44,8 ménages, autrement dit, après avoir ordonné par la colonne « Ménage du type MSE », nous prenons les feuilles jusqu'à avoir 44,8 ménages de type MSE (les deux premières lignes). Ensuite pour la feuille 84, il y a 5 ménages MSE contre 8 ménages pas MSE, autrement dit, il y a plus de 20% de ménage MSE, elle est donc aussi sélectionnée.

Nous formons notre échantillon d'apprentissage pour la régression logistique en prenant l'ensemble des ménages contenus dans les feuilles PMSE_PMSE (voir figure 9 et tableau 19). Les feuilles PMSE_PE seront imputés de la même manière que les autres types de ménage, les détails se trouvent au chapitre 7.4.4.3.

Figure 9 : Schéma de la procédure de création de l'échantillon d'apprentissage pour la régression logistique



Seules les feuilles de l'arbre 2018 (t), qui avaient été définies comme PMSE sont gardées (vert). Cet arbre de l'année t est appliqué aux échantillons d'apprentissage des années t-2, t-1 et t. En regardant le résultat de « l'imputation » des échantillons d'apprentissage, les feuilles PMSE_PMSE et PMSE_PE peuvent être définies. Les ménages des années t-2, t-1 et t contenus dans les feuilles PMSE_PMSE forment l'échantillon d'apprentissage de la régression logistique.

Tableau 19 : Echantillon d'apprentissage de la régression pour l'année 2019 (typologie 25 ans)

HHTyp	Ménages pas du type MSE	Ménage du type MSE	Total
PE	114	0	**114**
A	25	0	**25**
UL	1	0	**1**
MS	1	0	**1**
ME	2	0	**2**
MSE	0	57	**57**
Total	**143**	**57**	**200**

La technique ci-dessus, permet de garder les proportions entre les différents types de ménages tombant dans les feuilles PMSE_PMSE. Nous pouvons supposer qu'il n'y aura pas de ménage à double. En effet, si une même composition de ménage était déjà existante une année, l'année suivante, le type de ce ménage sera défini comme sûr et ne pourra plus être dans un échantillon d'apprentissage. De plus, comme les tirages aléatoires du Relevé Structurel sont coordonnés négativement, il est aussi peu probable d'avoir deux fois le même ménage tiré successivement.

7.4.4.2 Régression logistique

Il y a chaque année environ 2 800 ménages imputés par la régression logistique. Le but de cette régression logistique est de séparer les ménages MSE des autres types de ménages (en particulier les PE) dans les feuilles PMSE_PMSE. Pour créer le modèle logistique, les poids du Relevé Structurel sont utilisés pour spécifier au modèle la pondération des ménages. Dans le tableau 20, se trouvent les variables utilisées dans la régression logistique.

Tableau 20 : Variables du modèle de régression logistique

Variable	Catégo-riel	Explication
Taille*		Nombre d'habitants dans la commune
Urbain*	X	1 : Urbain 2 : Rural 3 : Intermédiaire Regroupés en (1) et (2, 3)
ForeignHouse-holdPerm*	X	1 : Ménage suisse-étranger (au moins un Suisse et un étranger) avec tous les étrangers du même pays 2 : Ménage suisse-étranger (au moins un Suisse et un étranger) avec des étran-gers de différents pays 3 : Ménage suisse-étranger (au moins un Suisse et un étranger) avec tous les étrangers sans indication de nationalité 4 : Ménage étranger (tous étrangers) avec tous les étrangers du même pays 5 : Ménage étranger (tous étrangers) avec des étrangers de différents pays 6 : Ménage étranger (tous étrangers) avec tous les étrangers sans indication de nationalité -8 : Ménage suisse (tous suisses) -9 : Ménage avec un au moins un étranger sans indication de nationalité et au moins un étranger avec indication de nationalité. Regroupés en (1, 3), (2, 4, 6, -9), (5) et (-8)
PartDivorce	X	Il existe une personne ayant comme état civil « En partenariat » ou « Partenariat dissous » (0, 1)
NbrePartGap1	X	Nombre de personnes ayant l'état civil « En partenariat » dans le Gap1 (0 ou 1, 2, …)
Cat_2msex_gap1	X	Il existe exactement deux personnes de même sexe dans le Gap1 (0, 1)
NbreMarieGap1	X	Nombre de personnes ayant l'état civil « Marié » dans le Gap1 (0, 1, …)
MoyenneAge		Moyenne de la variable âge dans le ménage
KurtAge		Kurtosis (quatrième moment) de la variable âge. Si elle n'est pas calculable[20] (SigmaAge=0), alors 4
nb_famille_gap1	X	Nombre de noms de famille différents dans le Gap1 (1, 2+)
nb_famille_gap0	X	Nombre de noms de famille différents dans le Gap0 (0 ou 1, 2+)
Age2Msexe		Âge de la personne la plus jeune après avoir choisi les deux personnes plus âgées de même sexe[21]. S'il n'y a pas deux personnes de même sexe, alors -9.
DiffAgeMsexe		Différence d'âge des deux personnes plus âgées de même sexe[21]. S'il n'y a pas deux personnes de même sexe, alors -9.

** Déjà présente dans le tableau 15 (chapitre 7.4.2)*

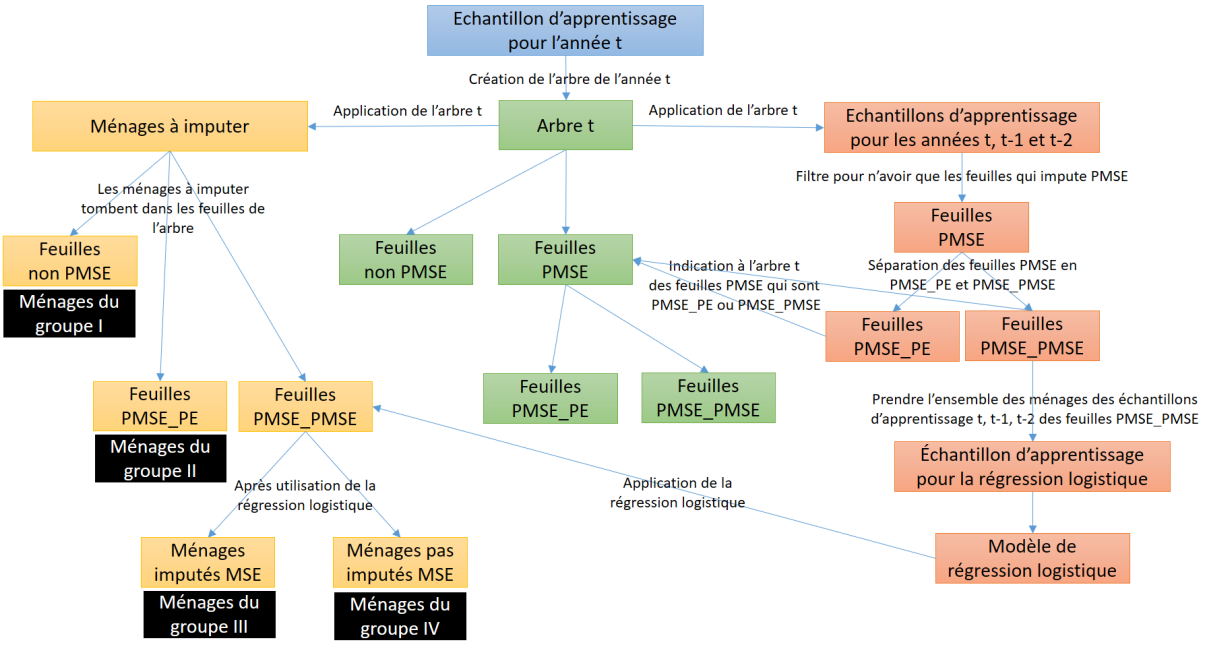
[20] La valeur 4 est choisie pour assurer une certaine continuité de la variable. Si SigmaAge=0, le calcul de KurtAge contient un 0/0 qui n'est pas défini. Cette valeur 4 correspond au KurtAge pour une différence de 1 an entre deux personnes.

[21] Soient P1, P2, P3 les trois personnes les plus âgées du ménage, d'âge A1, A2, A3 respectivement (avec A1 > A2 > A3). Si P1 et P2 sont de même sexe, alors DiffAgeMsexe = A2-A1 et Age2Msexe=A2. Si P1 et P3 sont de même sexe (mais pas P2), alors DiffAgeMsexe = A3-A1 et Age2Msexe=A3. Si P2 et P3 sont de même sexe (mais pas P1), alors DiffAgeMsexe =A3-A2 et Age2Msexe=2.

7.4.4.3 Imputation et calcul de l'estimateur de qualité

Il existe quatre cas de figure pour lesquels l'imputation et l'estimateur de qualité des ménages imputés doivent être définis. Ces cas de figure sont résumés dans la figure 10.

Figure 10 : Résumé de la procédure et mise en évidence des quatre possibilités d'imputation



- *Ménages du groupe I : ménages à imputer qui sont dans une feuille de l'arbre qui n'est pas du type PMSE*
- *Ménages du groupe II : ménages à imputer qui sont dans une feuille de l'arbre PMSE_PE*
- *Ménages du groupe III : ménages à imputer qui sont dans une feuille de l'arbre PMSE_PMSE et imputé MSE après utilisation de la régression logistique*
- *Ménage du groupe IV : ménages à imputer qui sont dans une feuille de l'arbre qui est du type PMSE_PMSE et pas imputé MSE après utilisation de la régression logistique*

Les imputations ainsi que les estimateurs de qualité des quatre groupes de ménages décrits dans la figure 10 sont définis de la manière suivante :

- Ménages du groupe I : ils sont imputés comme précédemment, autrement dit, quand l'échantillon d'apprentissage crée l'arbre de décision, à chaque feuille est rattaché un type majoritaire et la proportion de ce type dans la feuille. La valeur imputée est le type majoritaire et l'estimateur de qualité, la proportion de ce type de la feuille dans laquelle le ménage se trouve.

- Ménages du groupe II : ils suivent la même règle que celle du groupe I, la différence est que le type majoritaire et sa proportion dans la feuille doivent être recalculés en séparant PMSE en PE et MSE. Les ménages peuvent ensuite être imputés par le type majoritaire de leur feuille et avoir comme estimateur de qualité la proportion de ce type.

- Ménages des groupes III et IV : la régression logistique indique la probabilité pour chaque ménage d'être de type MSE ou non. Pour définir la probabilité d'un type non MSE, il manque la probabilité conditionnelle de ce type, sachant qu'il n'est pas MSE. Cette probabilité conditionnelle peut être obtenue en utilisant les proportions pondérées de l'échantillon d'apprentissage de la régression logistique avec les poids du Relevé Structurel, mais sans les ménages de type MSE (voir le tableau 21 et la formule qui le suit). Le type de ménage pour les groupes III et IV est imputé avec le type dont la probabilité p est la plus élevée. L'estimateur de qualité est défini par $1 - p$.

Tableau 21 : Echantillon d'apprentissage pondéré de la régression logistique sans le type MSE pour l'année 2019 (typologie 25 ans)

HHTyp	Ménages pas du type MSE	% Ménage pas du type MSE
PE	1977,30	75,05
A	517,98	19,66
UL	10,22	0,39
MS	55,67	2,11
ME	73,61	2,79

La probabilité qu'un ménage soit de type non MSE est défini par

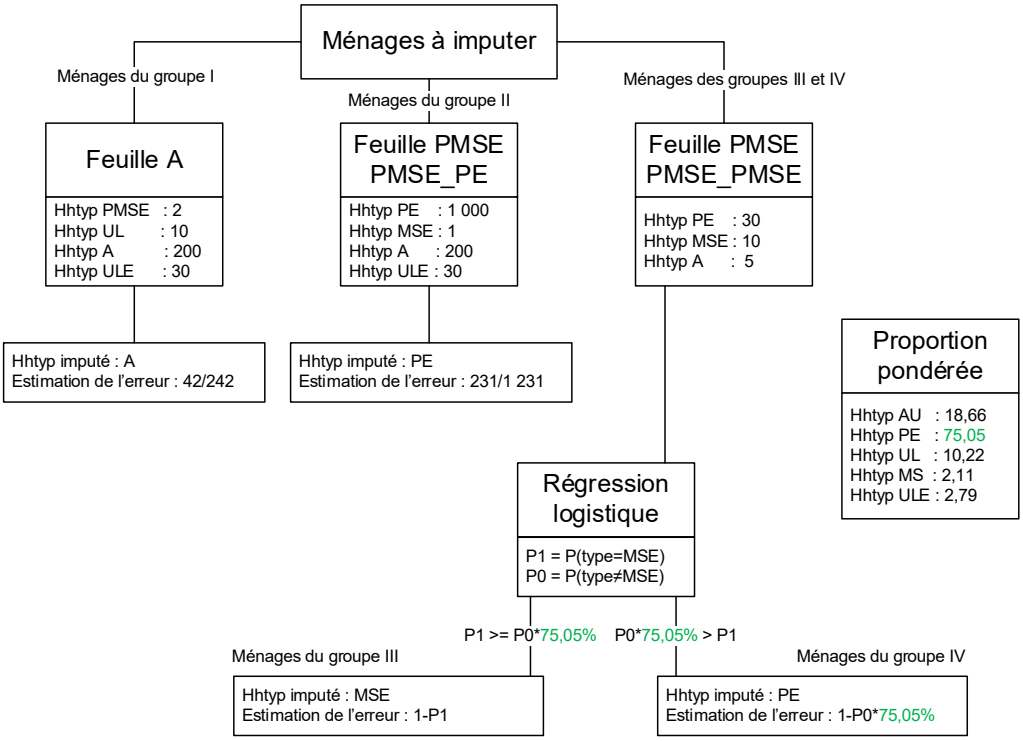
$$P(type = X) = P(type \neq MSE)\, P(type = X \,|type \neq MSE)$$

avec

P	estimation de la probabilité
X	type de ménage (pas MSE)
$P(type \neq MSE)$	estimation de la probabilité donnée par le résultat de la régression logistique. En effet, la régression logistique donne pour chaque ménage la probabilité estimée que ce ménage soit du type MSE ou non.
$P(type = X \,\|type \neq MSE)$	estimation de la probabilité donnée par la proportion pondérée du type X dans l'échantillon d'apprentissage auquel le type MSE a été retiré. Effectivement, en supposant que $type \neq MSE$, la probabilité du type X, revient à regarder sa proportion pondérée dans l'échantillon d'apprentissage (sans MSE).

La figure 11 montre un exemple d'imputation pour les groupes de ménages I, II, III et IV.

Figure 11 : Résumé avec un exemple des quatre types d'imputation avec leur estimateur de qualité



Dans le tableau 22 se trouve la comparaison du nombre d'imputations de type MSE, avant et après le traitement de la régression logistique.

Tableau 22 : Nombre total de ménages du type MSE après imputation, sans et avec la régression logistique et estimation de ce nombre par le Relevé Structurel

	Avec ORDIPRO				**Sans ORDIPRO**					
	Sans logistique		Avec logistique		Sans logistique		Avec logistique		Estimation du RS	
Année	25 ans	18 ans	25 ans	18 ans	25 ans	18 ans	25 ans	18 ans	25 ans	18 ans
2017	510	419	904	870	506	820	896	863	899	738
2018	1 369	1 039	1 019	894	1 340	1 031	925	891	1 015	966
2019	643	617	1 063	1 009	614	1 425	1 060	1 006	1 054	884
2020	619	968	1 108	1 001	616	672	1 102	1 060	1 262	1 151

Les résultats sans ORDIPRO ont été produits pour pouvoir être comparés au Relevé Structurel. En plus de l'instabilité de l'imputation sans régression logistique au fil des années, nous constatons également une instabilité, avec ou sans ORDIPRO, dans l'imputation sans régression logistique et au contraire une stabilité avec la régression logistique.

Les différents résultats de performance présentés au chapitre 6.2 sont mis en annexe avec une colonne supplémentaire « ADRL » (Arbre de Décision avec Régression Logistique).

8 Estimation de la qualité

Les données du Relevé Structurel ne sont disponibles que plusieurs mois après la production de la variable Type de ménage. Pendant la production, la qualité des imputations est évaluée par l'estimateur de qualité défini pour l'Arbre de décision avec régression logistique (ADRL).

8.1 Indicateur de qualité

Comme décrit au chapitre 7.4.4.3, pour chaque ménage imputé, l'algorithme de l'Arbre de décision avec régression logistique fournit une valeur qui nous permet d'estimer la qualité de l'imputation.

L'estimation de la qualité des imputations pour un domaine d'analyse donné est définie par

$$P_{erreur} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

avec

P_{erreur}	estimation de la probabilité d'erreur pour le domaine d'analyse
n	nombre de ménages dans le domaine d'analyse
p_i	estimation de la probabilité d'erreur pour le ménage i dans le domaine d'analyse. Si le ménage n'a pas été imputé, cette probabilité vaut 0.

Le domaine d'analyse peut par exemple correspondre à l'ensemble des ménages de Suisse, d'un canton ou d'une commune ou plus simplement d'un sous-ensemble quelconque de ménages.

Ce P_{erreur} sera notamment utilisé dans les analyses pour construire un indicateur de qualité qui s'articule en quatre modalités :

1. « Qualité suffisante » : pour les domaines d'analyse sans imputation et les domaines d'analyse de 100 ménages ou plus dont $P_{erreur} \leq 2\%$
2. « Qualité peut ne pas être suffisante » : pour les domaines d'analyse de 100 ménages ou plus dont $P_{erreur} > 2\%$
3. « Petit domaine : données volatiles, car < 100 ménages » : pour les domaines d'analyse ayant moins de 100 ménages et qui contiennent des imputations.
4. « Taux d'imputation élevé » : pour les types de ménage dans un domaine d'analyse ayant un taux d'imputation élevé[22]. Cette modalité ne dépend pas du P_{erreur}, mais dépend du taux d'imputation et de ce qui va être publié. Si le domaine d'analyse correspond aux communes et que dans une commune, tous les types « Couple en union libre sans enfant » sont imputés, tous les ménages avec ce type de la commune auront la valeur 4 dans l'indicateur (pour autant que ces ménages n'aient pas déjà une valeur 2 ou 3).

Même si P_{erreur} est en-dessous de 2%, il est possible que l'erreur effective soit plus grande que 2% due à la variance de l'estimateur. Cependant, en fixant le seuil à 2%, nous nous protégeons d'une erreur « trop grande ». Prenons l'exemple des résultats des Types de ménage (25 ans) de l'année 2018 avec les communes comme domaine d'analyse. Les communes de 100 ménages ou plus, dont $P_{erreur} \leq 2\%$, ont une moyenne[23] de P_{erreur} de 1,23% et un écart-type[24] moyen de 0,41%. Dans le tableau 23 se trouve la répartition des communes selon les quatre modalités de l'indicateur de qualité.

---

[22] Pour les types de ménage « A », « UL », « ULE » et « MSE », le seuil est fixé à 95% d'imputation. Pour les autres types, qui sont en général moins imputés, le seuil est fixé à 50%.

[23] Distribution des P_{erreur} pour les communes de 100 ménages ou plus dont $P_{erreur} \leq 2\%$: moyenne : 1,23, écart-type : 0,30, médiane : 1,22, MAD : 0,20.

[24] L'écart-type est calculé en supposant les imputations comme des variables aléatoires indépendantes suivant une loi de Bernoulli de paramètre p_i.

Tableau 23 : Répartition des communes selon les modalités de l'indicateur pour l'année 2018

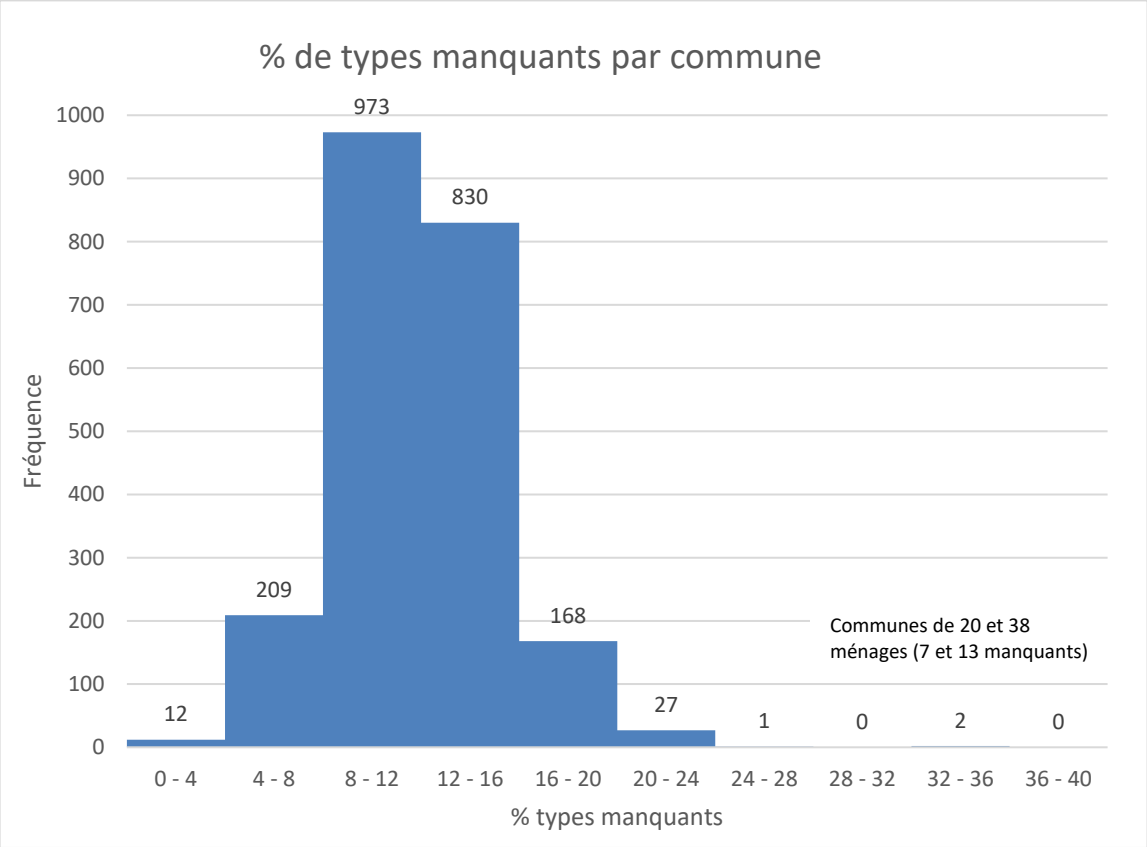
Indicateur de qualité	Nombre de communes 25 ans	Nombre de communes 18 ans
Qualité suffisante pour tous les types de ménage	1 724	1 701
Qualité peut ne pas être suffisante	72	30
Petit domaine : données volatiles, car < 100 ménages	143	143
Qualité suffisante pour une partie des types et taux d'imputation élevé pour au moins un type	283*	348*
Total	**2 222**	**2 222**

** pour 240 communes sur 283 (resp. 256 sur 348), seul le type MSE a un taux d'imputation élevé, car il y a souvent un seul ménage de ce type dans la commune.*

8.2 Taux de HHTyp manquants par commune

La figure 12 représente le nombre de communes par pourcentage de Types de ménage manquants.

Figure 12 : Répartition des communes par taux de Types de ménage manquants pour 2018



La première barre nous indique qu'il y a 12 communes qui ont entre 0% et 4% de Types de ménage manquants.

Dans les analyses, nous avons constaté que les communes de 100 ménages ou plus estimées à plus de 2% d'erreur ont en moyenne plus de valeurs manquantes (19,2%) que celles estimées à moins de 2% d'erreur (11,6%). Les résultats de 2019 se trouvent à l'annexe 2.

8.3 Analyse des erreurs

Pour identifier les erreurs d'imputation, nous comparons les types de ménage imputés avec les informations du Relevé Structurel. Dans le tableau 24 sont représentés les ménages imputés qui ont été interrogés la même année dans le Relevé Structurel multipliés par leur poids et mis en pourcentage. La somme de toutes les cases vaut 100%, la somme de la diagonale donne l'estimation du ratio d'imputations correctes et la somme hors diagonale donne l'estimation du ratio d'imputations incorrectes selon le Relevé Structurel.

Tableau 24 : Matrice de comparaison entre les types imputés et les types du RS pour 2018

ADRL	HHTyp RS							
HHTyp imputé	**PE**	**A**	**M**	**UL**	**MS**	**ME**	**ULE**	**MSE**
PE	4,72	0,27	.	0,07	.	.	0,05	0,07
A	0,09	16,71	0,12	0,79	2,03	0,04	0,03	.
M	.	0,20	2,33	0,10	.	0,02	.	.
UL	.	5,15	.	42,02	.	.	0,01	.
MS	.	0,02	.	.	0,06	.	.	.
ME	0,03	0,44	0,22	.	.	17,85	0,11	.
ULE	0,46	0,15	.	0,02	.	.	5,71	0,01
MSE	0,02	.	.	.	0,01	.	.	0,03

En couleur, se trouvent les différences de plus de 1%. La confusion entre les types Couple en union libre sans enfant (UL) et Autre type de ménage (A) (5,15%+0,79%=5,94%) et la confusion entre les ménages de Couple de même sexe sans enfant (MS) et les Autre type de ménage (A) (0,02+2,03=2,05%) représentent à elles seules 7,99% d'erreur sur 10,57% (somme hors diagonale). Autrement dit, elles représentent environ 76% des erreurs d'imputation.

Une analyse plus approfondie des données a relevé que les erreurs sont principalement dues aux ménages de 2 personnes. Il y a une confusion entre les Couples en union libre sans enfant et les colocataires ou sous-locataires et entre les Couples de même sexe sans enfant et les colocataires ou sous-locataires. Effectivement, lorsqu'il y a un ménage de deux personnes de sexe opposé, l'algorithme a tendance à imputer des Couples en union libre, alors que lorsque les deux personnes sont de même sexe, l'algorithme a tendance à imputer des colocataires.

Quand l'algorithme impute le type Couple en union libre sans enfant, dans environ 89% des cas, le résultat est le même que celui du Relevé Structurel (42,02/47,18) et dans environ 11% des cas, il se « trompe » avec colocataire (5,15/47,18). Quand l'algorithme impute le type Autre type de ménage, dans environ 84% des cas, le résultat est le même que celui du Relevé Structurel (16,71/19,81) et dans environ 14% des cas, il se « trompe » avec les types Couple en union libre sans enfant ou Couple de même sexe sans enfant ((0,79+2,03)/19,81). Pour 2018, nous estimons avoir 1,45% d'erreurs d'imputation, dont 1,10% est dû à ces confusions UL-A et H-A et 0,35% aux autres erreurs.

9 Perspectives

Le fait de publier la Typologie des ménages STATPOP en tant que statistique expérimentale nous permettra d'approfondir les tests pour, si possible, améliorer la qualité des imputations et de l'estimation de l'erreur. Il s'agirait, par exemple, d'améliorer l'échantillon d'apprentissage et de mieux distinguer les types de ménage UL-A et H-A qui représentent 1,10% des erreurs (sur 1,45% d'erreur). Il sera aussi nécessaire de plausibiliser le type de ménage imputé avec les caractéristiques des personnes qui composent le ménage[25] (en 2018, pour la typologie 25 ans, il y a 169 ménages incohérents sur 518 512 ménages imputés et pour la typologie 18 ans, il y a 29 ménages incohérents). Cette plausibilisation sera ajoutée dans le choix du type lors de l'imputation.

Bien que seuls les résultats issus des imputations par la méthode Arbre de décision avec régression logistique seront mis à disposition en raison de la meilleure estimation de la qualité, des pistes d'améliorations continueront aussi d'être recherchées pour la méthode Random Forest et potentiellement pour d'autres méthodes. Les deux méthodes d'imputation Arbre de décision et Random Forest vont continuer à être utilisées en parallèle lors de la production des types de ménage. Ceci permettra de contrôler que les résultats issus des deux méthodes ne divergent pas et de garantir une bonne qualité des résultats. A moyen terme, la variable Type de ménage devrait être intégrée dans la production courante de STATPOP.

En plus des méthodes d'imputation, deux autres points vont encore être approfondis. Il s'agira, d'une part, d'évaluer les possibilités de calculer les types de ménages pour les années passées. En d'autres termes, il faudra définir à partir de quelle année le nombre de relations reprises du Relevé Structurel est suffisant pour assurer une bonne qualité de la typologie. D'autre part, des tests sur une typologie avec des modalités agrégées seront réalisés. L'idée étant d'intégrer les Couples en union libre sans enfant (de sexe différent ou de même sexe) dans la catégorie Autre ménage de plusieurs personnes, afin d'améliorer la qualité de l'imputation. Cette typologie supplémentaire offrirait certes moins de possibilités d'analyses du point de vue thématique, mais nous pouvons espérer que les résultats des imputations seront de bien meilleure qualité. En effet, comme vu au chapitre 8.3 pour l'année 2018, parmi les 1,45% d'erreurs estimées suite aux imputations, 1,10% proviendraient de la confusion entre ces trois types spécifiques.

[25] Exemple d'incohérence : imputation du type « Couple marié avec enfants de moins de 25 ans » pour un ménage avec uniquement des personnes de plus de 25 ans.

Annexes

ANNEXE 1 : TABLEAU DE RÉSULTATS DES MÉTHODES – ANNÉES 2018-2019 (SUITE DU CHAPITRE 6.2)

Tableaux 2018 pour la typologie « 25 ans »

Tableau 25 : différence de distribution des Types de ménage sur la population entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 1 du chapitre 6.2)

	Estimation RS2018	ML	Diff^2 ML	Stat-VD	Diff^2 Stat-VD	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	35,686	35,686	0,000	35,686	0,000	35,686	0,000	35,686	0,000	35,686	0,000	35,686	0,000
PE	4,639	5,186	0,300	5,152	0,264	5,270	0,398	5,232	0,352	5,227	0,346	5,240	0,361
A	7,590	8,287	0,487	8,144	0,307	7,756	0,028	8,007	0,174	7,850	0,068	8,009	0,175
M	20,109	19,505	0,365	19,373	0,542	19,397	0,507	19,447	0,438	19,396	0,509	19,448	0,437
UL	6,530	7,508	0,957	7,915	1,919	7,735	1,453	7,512	0,965	7,695	1,357	7,514	0,968
MS	0,605	0,303	0,091	0,245	0,129	0,319	0,082	0,252	0,125	0,355	0,063	0,252	0,125
ME	22,118	20,632	2,211	20,424	2,872	20,877	1,542	20,873	1,550	20,784	1,780	20,871	1,555
ULE	2,697	2,871	0,031	3,045	0,121	2,934	0,056	2,954	0,067	2,982	0,082	2,956	0,067
MSE	0,027	0,021	0,000	0,015	0,000	0,026	0,000	0,036	0,000	0,025	0,000	0,025	0,000
dist			**2,107**		**2,481**		**2,017**		**1,916**		**2,051**		**1,921**

Tableau 26 : différence de distribution des types de ménage sur les imputations entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 2 du chapitre 6.2)

	Estimation RS2018	ML	Diff^2 ML	Stat-VD	Diff^2 Stat-VD	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
PE	1,103	5,103	15,994	4,853	14,061	5,711	21,235	5,436	18,769	5,400	18,458	5,490	19,245
A	21,695	26,790	25,960	25,743	16,387	22,912	1,482	24,743	9,294	23,596	3,615	24,753	9,355
M	8,991	4,581	19,451	3,617	28,885	3,791	27,039	4,157	23,375	3,782	27,142	4,164	23,302
UL	34,823	41,964	51,001	44,938	102,318	43,625	77,473	41,998	51,477	43,330	72,375	42,008	51,631
MS	2,715	0,510	4,860	0,090	6,890	0,628	4,355	0,138	6,639	0,889	3,334	0,138	6,639
ME	25,668	14,811	117,874	13,293	153,137	16,600	82,230	16,576	82,664	15,925	94,932	16,563	82,904
ULE	4,904	6,182	1,632	7,450	6,478	6,636	3,000	6,787	3,546	6,992	4,360	6,798	3,587
MSE	0,102	0,058	0,002	0,016	0,007	0,096	0,000	0,165	0,004	0,086	0,000	0,084	0,000
dist			**15,387**		**18,115**		**14,725**		**13,992**		**14,974**		**14,024**

Tableau 27 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS (point 3 du chapitre 6.2)

	ML	Stat-VD	RM	AD	RF	ADRL
% imputations incorrectes estimé sur les manquants	10,61	14,61	12,15	10,55	9,99	10,57
% imputations incorrectes estimé sur la population totale	1,45	2,00	1,66	1,45	1,37	1,45

Tableau 28 : Estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS par canton

%Erreur Canton	n 2018 ∩ Manquant	ML		Stat-Vd		RM		AD		RF		ADRL	
		Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total
AG	1862	11,37	1,52	12,92	1,72	12,17	1,62	10,16	1,35	9,94	1,33	10,18	1,36
AI	26	8,76	0,81	9,33	0,86	11,99	1,11	8,76	0,81	11,99	1,11	8,76	0,81
AR	133	12,48	1,48	18,54	2,20	19,38	2,30	17,05	2,02	12,74	1,51	17,05	2,02
BE	2986	11,16	1,42	14,77	1,88	12,74	1,62	11,37	1,45	11,12	1,42	11,33	1,44
BL	904	9,06	1,31	13,87	2,01	9,79	1,42	9,65	1,40	8,15	1,18	9,65	1,40
BS	579	11,68	1,84	14,50	2,28	13,43	2,12	11,90	1,87	11,81	1,86	11,90	1,87
FR	1003	7,69	1,20	16,25	2,53	10,02	1,56	8,03	1,25	7,54	1,17	8,07	1,26
GE	1583	10,92	1,42	13,37	1,73	16,16	2,09	12,71	1,65	11,01	1,43	12,87	1,67
GL	117	5,87	0,88	10,76	1,62	9,82	1,47	4,82	0,72	5,88	0,88	4,82	0,72
GR	491	11,08	1,20	18,82	2,04	10,38	1,13	9,15	0,99	8,92	0,97	9,34	1,01
JU	169	9,32	1,01	12,19	1,33	10,77	1,17	8,96	0,97	6,53	0,71	8,69	0,95
LU	1803	11,50	1,35	14,85	1,74	13,41	1,57	10,74	1,26	10,88	1,27	10,80	1,26
NE	686	8,43	0,74	10,75	0,94	9,86	0,86	8,30	0,73	7,58	0,66	8,30	0,73
NW	141	11,90	1,79	15,50	2,33	9,61	1,44	9,94	1,49	10,42	1,57	9,94	1,49
OW	100	13,35	1,62	16,50	2,00	15,24	1,85	13,39	1,62	12,46	1,51	13,39	1,62
SG	1471	9,41	1,31	13,67	1,90	9,36	1,30	8,06	1,12	8,14	1,13	8,20	1,14
SH	232	11,30	1,56	10,70	1,48	9,19	1,27	8,38	1,16	7,00	0,97	8,38	1,16
SO	832	9,48	1,38	12,25	1,78	10,65	1,55	8,79	1,28	9,25	1,35	8,79	1,28
SZ	509	12,10	1,91	17,76	2,80	13,66	2,15	12,80	2,02	12,89	2,03	12,80	2,02
TG	736	9,52	1,26	14,45	1,91	9,00	1,19	8,89	1,18	8,27	1,09	8,89	1,18
TI	1414	6,79	0,60	9,07	0,81	8,32	0,74	6,50	0,58	5,89	0,52	6,50	0,58
UR	71	9,89	0,94	19,07	1,81	12,64	1,20	11,38	1,08	7,55	0,72	11,38	1,08
VD	4038	10,80	1,37	13,46	1,71	12,27	1,56	11,45	1,45	10,88	1,38	11,51	1,46
VS	885	7,54	0,94	11,85	1,48	10,05	1,26	7,66	0,96	7,09	0,89	7,66	0,96
ZG	804	9,67	1,40	10,89	1,57	10,29	1,49	8,48	1,23	8,02	1,16	8,48	1,23
ZH	5150	12,02	2,07	17,35	2,98	13,81	2,37	12,22	2,10	11,32	1,95	12,20	2,10

Tableaux 2018 pour la typologie « 18 ans »

Ici seules les méthodes Régression Multinomiale (RM), Arbre de décision (AD), Random Forest (RF) et Arbre de décision avec régression logistique (ADRL) ont été testées.

Tableau 29 : différence de distribution des Types de ménage sur la population entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 1 du chapitre 6.2)

	Estimation RS2018	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	35,686	35,686	0,000	35,686	0,000	35,686	0,000	35,686	0,000
PE	3,064	3,398	0,111	3,386	0,103	3,397	0,111	3,396	0,110
A	14,214	14,170	0,002	14,442	0,052	14,225	0,000	14,463	0,062
M	20,324	19,409	0,837	19,385	0,882	19,396	0,861	19,381	0,889
UL	6,600	7,740	1,299	7,593	0,987	7,713	1,238	7,582	0,963
MS	0,610	0,322	0,083	0,242	0,135	0,346	0,070	0,242	0,135
ME	17,130	16,677	0,205	16,617	0,264	16,607	0,274	16,621	0,260
ULE	2,345	2,572	0,052	2,621	0,076	2,604	0,067	2,605	0,068
MSE	0,026	0,026	0,000	0,027	0,000	0,026	0,000	0,024	0,000
dist			**1,609**		**1,581**		**1,619**		**1,577**

Tableau 30 : différence de distribution des types de ménage sur les imputations entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 2 du chapitre 6.2)

	Estimation RS2018	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
PE	0,650	3,087	5,939	2,997	5,507	3,080	5,906	3,074	5,873
A	30,482	30,160	0,104	32,149	2,776	30,562	0,006	32,302	3,310
M	10,561	3,881	44,625	3,703	47,031	3,786	45,902	3,674	47,424
UL	35,338	43,659	69,242	42,591	52,604	43,462	66,000	42,505	51,367
MS	2,751	0,645	4,435	0,067	7,203	0,822	3,724	0,067	7,203
ME	16,652	13,344	10,943	12,902	14,062	12,832	14,591	12,930	13,860
ULE	3,458	5,118	2,756	5,472	4,056	5,349	3,577	5,356	3,605
MSE	0,109	0,106	0,000	0,119	0,000	0,107	0,000	0,092	0,000
dist			**11,749**		**11,543**		**11,820**		**11,517**

Tableau 31 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS (point 3 du chapitre 6.2)

	RM	AD	RF	ADRL
% imputations incorrectes estimé sur les manquants	11,25	9,80	9,29	9,75
% imputations incorrectes estimé sur la population totale	1,54	1,34	1,27	1,33

Tableau 32 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS par canton

%Erreur	n	RM		AD		RF		ADRL	
Canton	2018 ∩ Manquant	Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total
AG	1862	11,33	1,51	9,72	1,30	9,34	1,24	9,69	1,29
AI	26	8,76	0,81	8,76	0,81	11,99	1,11	8,76	0,81
AR	133	16,29	1,93	15,48	1,84	12,18	1,45	15,48	1,84
BE	2986	12,19	1,55	10,84	1,38	10,46	1,33	10,74	1,37
BL	904	8,42	1,22	8,58	1,25	7,18	1,04	8,54	1,24
BS	579	12,56	1,98	11,32	1,78	10,90	1,72	11,32	1,78
FR	1003	8,82	1,38	7,26	1,13	6,95	1,08	7,16	1,12
GE	1583	14,71	1,91	10,42	1,35	9,57	1,24	10,26	1,33
GL	117	6,40	0,96	3,19	0,48	3,19	0,48	3,19	0,48
GR	491	9,60	1,04	8,14	0,88	8,67	0,94	8,14	0,88
JU	169	10,66	1,16	8,48	0,92	6,97	0,76	8,48	0,92
LU	1803	12,36	1,45	10,26	1,20	10,06	1,18	10,26	1,20
NE	686	8,65	0,76	6,81	0,60	6,51	0,57	6,81	0,60
NW	141	9,74	1,46	9,47	1,42	9,47	1,42	9,47	1,42
OW	100	13,01	1,58	12,38	1,50	10,23	1,24	12,38	1,50
SG	1471	8,45	1,18	7,47	1,04	7,51	1,05	7,46	1,04
SH	232	6,73	0,93	6,82	0,94	5,17	0,72	7,31	1,01
SO	832	9,00	1,31	8,24	1,20	8,00	1,16	8,24	1,20
SZ	509	12,80	2,02	11,36	1,79	11,46	1,81	11,36	1,79
TG	736	8,95	1,18	8,51	1,13	7,92	1,05	8,51	1,13
TI	1414	7,45	0,66	6,00	0,53	5,43	0,48	6,07	0,54
UR	71	12,64	1,20	9,75	0,93	5,92	0,56	9,75	0,93
VD	4038	11,33	1,44	10,74	1,36	10,12	1,29	10,72	1,36
VS	885	9,03	1,13	7,29	0,91	6,74	0,84	7,09	0,89
ZG	804	9,96	1,44	8,52	1,23	7,91	1,14	8,41	1,22
ZH	5150	13,13	2,26	11,48	1,97	10,88	1,87	11,38	1,96

Tableaux 2019 pour la typologie « 25 ans »

Tableau 33 : différence de distribution des Types de ménage sur la population entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 1 du chapitre 6.2)

	Estimation RS2019	ML	Diff^2 ML	Stat-VD	Diff^2 Stat-VD	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	36,041	36,041	0,000	36,041	0,000	36,041	0,000	36,041	0,000	36,041	0,000	36,041	0,000
PE	4,741	5,200	0,211	5,165	0,180	5,278	0,288	5,248	0,257	5,230	0,239	5,238	0,247
A	7,515	8,125	0,371	7,960	0,198	7,718	0,041	7,815	0,090	7,778	0,069	7,812	0,088
M	19,799	19,345	0,206	19,238	0,314	19,264	0,286	19,320	0,229	19,264	0,285	19,320	0,229
UL	6,754	7,598	0,712	8,013	1,584	7,780	1,052	7,719	0,930	7,744	0,979	7,719	0,930
MS	0,638	0,318	0,102	0,257	0,145	0,323	0,099	0,255	0,146	0,344	0,087	0,255	0,146
ME	21,722	20,388	1,779	20,187	2,357	20,551	1,373	20,552	1,369	20,518	1,451	20,552	1,369
ULE	2,762	2,963	0,040	3,123	0,130	3,009	0,061	3,034	0,074	3,050	0,083	3,035	0,074
MSE	0,028	0,023	0,000	0,017	0,000	0,037	0,000	0,016	0,000	0,032	0,000	0,028	0,000
dist			**1,850**		**2,215**		**1,789**		**1,759**		**1,787**		**1,756**

Tableau 34 : différence de distribution des types de ménage sur les imputations entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 2 du chapitre 6.2)

	Estimation RS2019	ML	Diff^2 ML	Stat-VD	Diff^2 Stat-VD	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
PE	1,579	5,042	11,996	4,775	10,221	5,627	16,394	5,402	14,619	5,269	13,620	5,332	14,085
A	21,642	26,239	21,139	24,997	11,260	23,169	2,333	23,903	5,112	23,623	3,924	23,878	5,003
M	7,543	4,119	11,725	3,315	17,875	3,507	16,293	3,931	13,044	3,512	16,248	3,931	13,044
UL	37,087	43,453	40,525	46,583	90,171	44,827	59,898	44,365	52,963	44,554	55,749	44,365	52,963
MS	2,963	0,549	5,831	0,089	8,260	0,591	5,627	0,078	8,328	0,744	4,927	0,078	8,326
ME	24,406	14,341	101,290	12,822	134,179	15,566	78,148	15,579	77,917	15,317	82,603	15,579	77,917
ULE	4,679	6,192	2,289	7,400	7,405	6,544	3,481	6,730	4,208	6,852	4,721	6,736	4,232
MSE	0,100	0,063	0,001	0,017	0,007	0,169	0,005	0,013	0,008	0,130	0,001	0,101	0,000
dist			**13,957**		**16,715**		**13,497**		**13,274**		**13,483**		**13,250**

Tableau 35 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS (point 3 du chapitre 6.2)

	ML	Stat-VD	RM	AD	RF	ADRL
% imputations incorrectes estimé sur les manquants	11,00	15,31	12,56	11,16	10,85	11,16
% imputations incorrectes estimé sur la population totale	1,46	2,03	1,66	1,48	1,44	1,48

Tableau 36 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS par canton

%Erreur Canton	n 2018 ∩ Manquant	ML		Stat-Vd		RM		AD		RF		ADRL	
		Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total
AG	1936	10,85	1,43	14,03	1,84	12,04	1,58	10,94	1,44	10,45	1,37	10,94	1,44
AI	38	4,41	0,51	0,68	0,08	2,34	0,27	0,00	0,00	0,00	0,00	0,00	0,00
AR	137	12,39	1,50	13,33	1,61	12,30	1,49	9,31	1,13	13,42	1,62	9,31	1,13
BE	3263	10,71	1,37	14,15	1,82	12,81	1,64	11,61	1,49	10,94	1,40	11,55	1,48
BL	835	9,79	1,24	13,71	1,74	11,50	1,46	9,87	1,25	10,27	1,30	9,73	1,23
BS	596	11,76	1,76	18,12	2,71	15,23	2,28	13,28	1,98	13,03	1,95	13,28	1,98
FR	936	9,90	1,40	17,40	2,46	11,58	1,64	9,53	1,35	9,51	1,35	9,53	1,35
GE	1686	12,66	1,66	17,16	2,25	15,82	2,08	12,88	1,69	12,97	1,70	13,03	1,71
GL	96	7,41	0,86	14,30	1,67	5,77	0,67	6,52	0,76	5,46	0,64	6,52	0,76
GR	512	12,84	1,53	23,90	2,84	13,37	1,59	12,15	1,44	12,74	1,51	12,77	1,52
JU	136	6,61	0,54	8,79	0,72	7,65	0,62	7,85	0,64	6,04	0,49	8,54	0,70
LU	2130	12,01	1,50	14,74	1,84	13,15	1,64	11,91	1,48	11,75	1,46	11,93	1,49
NE	765	7,87	0,77	8,37	0,82	8,81	0,86	7,82	0,76	7,03	0,69	7,82	0,76
NW	143	9,11	1,29	14,33	2,03	8,83	1,25	11,16	1,58	11,26	1,60	11,16	1,58
OW	99	11,32	1,45	12,72	1,63	13,29	1,71	12,69	1,63	11,40	1,46	12,69	1,63
SG	1493	10,15	1,36	15,06	2,02	10,55	1,41	9,11	1,22	8,95	1,20	9,07	1,22
SH	225	9,70	1,24	12,31	1,57	9,33	1,19	7,73	0,99	8,08	1,03	7,73	0,99
SO	870	9,83	1,43	14,14	2,05	10,42	1,51	9,45	1,37	8,51	1,23	9,37	1,36
SZ	525	8,98	1,37	16,01	2,44	9,89	1,51	10,02	1,53	9,98	1,52	10,02	1,53
TG	754	10,42	1,34	15,72	2,02	13,27	1,71	11,24	1,45	11,15	1,43	11,24	1,45
TI	1216	8,25	0,63	10,61	0,82	9,30	0,71	7,67	0,59	7,64	0,59	7,55	0,58
UR	89	5,91	0,71	16,48	1,99	8,16	0,98	8,16	0,98	5,99	0,72	8,16	0,98
VD	3864	10,83	1,31	12,67	1,53	13,19	1,60	11,51	1,39	10,40	1,26	11,54	1,40
VS	824	8,25	0,96	12,34	1,43	10,69	1,24	8,73	1,01	8,38	0,97	8,73	1,01
ZG	755	9,62	1,30	11,39	1,54	11,22	1,51	9,84	1,33	9,24	1,25	9,84	1,33
ZH	5089	12,85	2,10	18,10	2,96	14,03	2,29	12,85	2,10	12,72	2,08	12,82	2,10

Tableaux 2019 pour la typologie « 18 ans »

Ici seules les méthodes Régression Multinomiale (RM), Arbre de décision (AD), Random Forest (RF) et Arbre de décision avec régression logistique (ADRL) ont été testées.

Tableau 37 : différence de distribution des Types de ménage sur la population entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 1 du chapitre 6.2)

	Estimation RS2019	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	36,041	36,041	0,000	36,041	0,000	36,041	0,000	36,041	0,000
PE	3,098	3,426	0,107	3,416	0,101	3,417	0,102	3,413	0,099
A	14,065	14,052	0,000	14,166	0,010	14,065	0,000	14,233	0,028
M	19,922	19,267	0,429	19,268	0,428	19,270	0,425	19,265	0,432
UL	6,795	7,789	0,988	7,684	0,791	7,749	0,911	7,671	0,767
MS	0,641	0,311	0,108	0,258	0,146	0,330	0,097	0,258	0,146
ME	16,973	16,438	0,286	16,438	0,286	16,421	0,305	16,403	0,325
ULE	2,441	2,640	0,039	2,700	0,067	2,676	0,055	2,689	0,062
MSE	0,023	0,036	0,000	0,027	0,000	0,031	0,000	0,026	0,000
dist			**1,400**		**1,352**		**1,376**		**1,363**

Tableau 38 : différence de distribution des types de ménage sur les imputations entre l'estimation du Relevé Structurel et les différentes méthodes d'imputation (point 2 du chapitre 6.2)

	Estimation RS2019	RM	Diff^2 RM	AD	Diff^2 AD	RF	Diff^2 RF	ADRL	Diff^2 ADRL
.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
1P	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
PE	0,587	3,060	6,115	2,986	5,759	2,994	5,794	2,959	5,627
A	30,546	30,443	0,011	31,308	0,581	30,544	0,000	31,812	1,602
M	8,474	3,530	24,442	3,539	24,350	3,553	24,219	3,515	24,587
UL	37,397	44,898	56,259	44,106	45,015	44,598	51,854	44,007	43,689
MS	2,984	0,501	6,165	0,097	8,335	0,639	5,500	0,101	8,310
ME	16,484	12,451	16,265	12,450	16,276	12,319	17,348	12,183	18,499
ULE	3,446	4,945	2,249	5,401	3,822	5,218	3,141	5,319	3,509
MSE	0,081	0,173	0,008	0,112	0,001	0,136	0,003	0,104	0,001
dist			**10,560**		**10,205**		**10,386**		**10,287**

Tableau 39 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS (point 3 du chapitre 6.2)

	RM	AD	RF	ADRL
% imputations incorrectes estimé sur les manquants	11,56	10,59	10,07	10,57
% imputations incorrectes estimé sur la population totale	1,53	1,40	1,33	1,40

Tableau 40 : estimation du pourcentage d'erreurs en comparant individuellement les imputations au RS par canton

%Erreur Canton	n 2018 ∩ Manquant	RM		AD		RF		ADRL	
		Manquant	Total	Manquant	Total	Manquant	Total	Manquant	Total
AG	1936	10,98	1,44	10,54	1,39	9,85	1,29	10,48	1,38
AI	38	3,03	0,35	0,00	0,00	0,00	0,00	0,00	0,00
AR	137	11,99	1,45	11,03	1,34	12,49	1,51	10,26	1,24
BE	3263	12,32	1,58	10,83	1,39	10,22	1,31	10,93	1,40
BL	835	9,50	1,20	9,49	1,20	8,84	1,12	9,34	1,18
BS	596	13,50	2,02	12,36	1,85	12,17	1,82	12,59	1,88
FR	936	10,56	1,50	9,14	1,29	8,74	1,24	9,01	1,28
GE	1686	13,72	1,80	12,17	1,60	11,55	1,52	11,74	1,54
GL	96	6,26	0,73	6,23	0,73	5,46	0,64	4,85	0,57
GR	512	12,28	1,46	12,38	1,47	12,52	1,49	12,33	1,47
JU	136	7,24	0,59	7,44	0,61	4,67	0,38	7,72	0,63
LU	2130	12,62	1,57	11,49	1,43	11,39	1,42	11,50	1,43
NE	765	7,90	0,77	7,37	0,72	6,57	0,64	7,12	0,70
NW	143	9,11	1,29	7,95	1,13	9,28	1,32	7,51	1,07
OW	99	12,24	1,57	11,40	1,46	11,40	1,46	10,95	1,41
SG	1493	9,21	1,23	8,60	1,15	7,73	1,04	8,58	1,15
SH	225	8,83	1,13	7,05	0,90	8,08	1,03	7,40	0,94
SO	870	10,28	1,49	9,22	1,34	8,49	1,23	9,50	1,38
SZ	525	9,87	1,50	9,85	1,50	9,36	1,43	9,77	1,49
TG	754	12,04	1,55	11,25	1,45	10,61	1,36	10,85	1,40
TI	1216	9,10	0,70	7,44	0,57	6,68	0,51	7,58	0,58
UR	89	8,16	0,98	7,10	0,86	5,99	0,72	7,10	0,86
VD	3864	11,78	1,43	10,71	1,30	9,61	1,16	10,47	1,27
VS	824	9,67	1,12	8,03	0,93	7,42	0,86	8,62	1,00
ZG	755	9,97	1,34	9,10	1,23	8,67	1,17	9,01	1,21
ZH	5089	13,02	2,13	12,15	1,99	11,88	1,94	12,20	1,99

ANNEXE 2 : TABLEAU DU TAUX DE MANQUANTS PAR COMMUNE EN 2019 (SUITE DU CHAPITRE 8.2)

Figure 13 : Répartition des communes par taux de Types de ménage manquants pour 2019

