RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Drees
Statistique publique
de la santé et des solidarités

DREES MÉTHODES

N° 26 • avril 2026

Estimation des revenus des parents dans l'enquête Modes de garde et d'accueil des jeunes enfants 2021

Imputations post-appariements fiscaux et sociaux

Nathalie Missègue, en collaboration avec Hélène Guedj

Estimation des revenus des parents dans l'enquête Modes de garde et d'accueil des jeunes enfants 2021

Imputations post-appariements fiscaux et sociaux

Nathalie Missègue, en collaboration avec Hélène Guedj

Retrouvez toutes nos publications sur : Drees.solidarites-sante.gouv.fr

Retrouvez toutes nos données sur : data.Drees.solidarites-sante.gouv.fr

SYNTHÈSE

L'enquête Modes de garde et d'accueil des jeunes enfants (MDG) a été réalisée par la Drees en 2021. Le champ de l'enquête porte sur les ménages ayant au moins un enfant de moins de 6 ans résidant en France métropolitaine et comporte une extension à La Réunion auprès des ménages avec au moins un enfant de moins de 3 ans. Cette enquête dresse, tous les 6-7 ans environ, un panorama complet des solutions de garde et d'accueil adoptées par les parents pour leurs jeunes enfants, compte tenu de leurs contraintes professionnelles. Sur une semaine type, le calendrier d'accueil des enfants est décrit dans sa totalité (y compris les périodes passées avec leurs parents). Tous les intervenants assurant la prise en charge des enfants pendant une semaine de référence sont identifiés. La mesure du reste à charge des familles selon le mode d'accueil, compte tenu du coût qu'il représente, est également un des objectifs de l'enquête. Aussi, les revenus des familles doivent-ils être mesurés de manière la plus précise possible. C'est pourquoi les diverses composantes des ressources des familles (salaires, allocations chômage, etc.) sont évaluées grâce à un appariement avec les sources fiscales et sociales dont dispose l'Insee. Il s'agit d'un appariement effectué sur la base de données identifiantes (nom, prénom, etc.). Il est qualifié de « statistique » : lorsque l'appariement n'est pas certain, une personne ayant les mêmes caractéristiques au regard d'une partie seulement des données identifiantes est considérée comme la personne retrouvée.

L'enrichissement avec les données fiscales est de très bonne qualité : sur les 8 787 ménages de l'enquête, 8 348 ménages soit 95 % ont été complètement retrouvés dans les déclarations de revenus et d'imposition.

Par ailleurs, 8 538 individus sur 16 373 (soit 52 %) ont été retrouvés dans les fichiers sociaux comme percevant des prestations d'au moins une des caisses disponibles (CNAF et CCMSA).

Afin de disposer des revenus pour l'ensemble des parents de l'enquête, il est nécessaire d'imputer des revenus aux répondants et/ou à leurs conjoints qui sont actifs (actifs occupés ou au chômage) et qui n'ont pas été retrouvés à l'issue de l'enrichissement. En effet, lorsque le parent est actif occupé ou au chômage un revenu est attendu. La mise en œuvre des traitements distingue deux groupes d'imputation selon le statut isolé (familles monoparentales) ou en couple du parent répondant, tel que déclaré à l'enquête. Dans le second cas, on souhaite connaître les revenus des conjoints en plus de ceux du répondant. Au sein de chacun des deux groupes à traiter, la démarche consiste en une imputation par « *Random forest* » (forêts aléatoires). C'est une méthode d'imputation non paramétrique : elle ne fait pas de suppositions sur la distribution des variables à partir des données observables. C'est la méthode la plus performante parmi les trois méthodes qui ont été testées (cf. partie Imputation des revenus manquants). Pour chacun des deux groupes d'imputation, une fois les imputations réalisées, les distributions des revenus (l'ensemble des revenus, y compris ceux imputés) sont comparées à celles des seuls revenus issus de l'enrichissement. Il s'avère que les distributions après imputations diffèrent globalement très peu de celles avant imputations.

SOMMAIRE

INTRODUCTION

L'enquête Modes de garde et d'accueil des jeunes enfants menée auprès de ménages ayant au moins un enfant de moins de 6 ans a été réalisée par la Drees de 2002 à 2021, tous les 6-7 ans environ sur le territoire métropolitain et pour la première fois à La Réunion en 2021 auprès de ménages ayant au moins un enfant de moins de 3 ans. Cette enquête a été mise en place pour :

- Dresser un panorama complet des solutions de prise en charge et d'accueil adoptées par les parents pour leurs jeunes enfants, au regard de leurs contraintes professionnelles. Elle vise, notamment, à reconstituer de façon très précise, sur une semaine type, le calendrier d'accueil des enfants, incluant les périodes passées avec leurs parents, et à identifier l'ensemble des intervenants qui se succèdent pour s'occuper d'eux pendant la période considérée.
- Mesurer le reste à charge des familles selon le mode d'accueil.

Pour atteindre ce second objectif, il est particulièrement important de mesurer de manière fiable les ressources dont disposent les ménages. Ces derniers ont été interrogés, dans l'enquête, sur les ressources qu'ils perçoivent. Mais les montants déclarés par les enquêtés sont généralement sous-estimés (Burricand, 2013).

La collecte des revenus par voie d'appariement plutôt que par voie d'enquête est donc préférable, puisqu'elle permet de limiter les erreurs de mesure, de gagner en qualité sur la mesure des revenus individuels (notamment depuis que les déclarations de revenus sont préremplies) et d'alléger sensiblement les questionnaires (questionnement réduit sur les revenus). Elle permet également de disposer des revenus réellement perçus sur une année complète. L'appariement fiscal, réalisé par l'Insee, s'appuie sur deux sources administratives qui sont rapprochées des données d'enquête à l'aide de données identifiantes. Pour la mesure des revenus individuels perçus (salaires, allocations de chômage, revenus non salariaux, etc.), des revenus collectifs ou « non individualisables » (revenus fonciers, etc.) et des impôts, l'Insee mobilise les données de la Direction générale des finances publiques (DGFiP) issues des déclarations de revenus. Pour les prestations sociales, ce sont les données de la Caisse nationale des allocations familiales (CNAF) qui sont utilisées. Pour les allocations familiale[i]s et sociales des personnes relevant du régime agricole, l'Insee mobilise les données de la Caisse centrale de mutualité sociale agricole (CCMSA).

LES APPARIEMENTS RÉALISÉS

Appariements fiscaux

Les appariements fiscaux réalisés par l'Insee sont des appariements statistiques, effectués sur la base de données identifiantes : nom, prénom, etc. Ce qualificatif de « statistique » est lié au fait qu'il n'est pas toujours possible de retrouver exactement la personne concernée au sens où on dispose et on retrouve l'ensemble des données identifiantes. Dans ce cas, une personne ayant les mêmes caractéristiques au regard d'une partie des données identifiantes est considérée comme la personne retrouvée. L'opération d'enrichissement est réalisée en deux temps. Les personnes de l'enquête sont tout d'abord appariées avec les individus du fichier d'imposition des personnes (FIP) de la DGFiP. Le fichier est ensuite enrichi avec les données de déclarations fiscales des contribuables (fichier POTE de la DGFiP). Dans certains cas, il est possible d'apparier un individu issu du fichier d'enquête avec le fichier FIP sans toutefois parvenir à disposer de ses informations de revenus et d'imposition (fichier POTE). Cette différence peut s'expliquer de différentes façons : changement de situation familiale (décès du conjoint qui remplissait la déclaration de revenus, par exemple) ou de département des enquêtés (arrivée récente dans le logement actuel, par exemple), absence de déclaration fiscale, etc. L'appariement est réalisé en utilisant des combinaisons de variables suffisamment discriminantes (variables personnelles et de domicile) pour permettre une identification des individus. Le rapprochement est réalisé par appariement exact dans un premier temps, c'est-à-dire par comparaison de deux chaînes de caractères strictement identiques, puis par appariement par plus proche écho, en tenant compte de différences pouvant exister entre les chaînes de caractères. Au final, dans cette seconde phase, on retient l'individu fiscal le plus proche de l'individu enquêté.

Les données récupérées *via* l'enrichissement sont l'ensemble des informations déclarées à l'administration fiscale en 2022 sur les revenus 2021. Le document « Déclaration des revenus 2021 – Brochure pratique 2022 » édité par la DGFiP détaille de façon exhaustive l'ensemble des informations collectées et qui ont été transmises à la Drees par l'Insee, hors informations nominatives. Pour l'appariement avec les données fiscales, les données identifiantes sont : le nom, le prénom, le sexe, les dates et lieu de naissance, l'état matrimonial et l'adresse de résidence ; elles sont issues de l'enquête. Les données retournées à la Drees portent sur l'ensemble des membres du ménage.

L'enrichissement avec les données fiscales est de très bonne qualité : sur les 8 787 ménages de l'enquête, 8 348 ménages soit 95 % ont été retrouvés « complètement » dans les déclarations de revenus et d'imposition, et les informations les concernant ont donc pu être enrichies avec les données de ces déclarations de revenus. Ils sont enrichis complètement c'est à dire que l'ensemble des membres du ménage décrits dans l'enquête ont été enrichis. Pour certains ménages une partie seulement des revenus des individus de l'enquête MDG ont été retrouvés (par exemple, ceux du conjoint pour les couples n'ont pas été retrouvés), dans ce cas l'enrichissement est qualifié de « partiel ». Si l'on tient compte des ménages enrichis complètement et partiellement, le taux d'enrichissement est très élevé : 98 %.

Il a cependant été nécessaire d'imputer les revenus qui manquent pour les répondants actifs et leurs conjoints éventuels non retrouvés dans les fichiers de la DGFiP.

Appariements sociaux (CNAF, CCMSA)

Les appariements sociaux réalisés par l'Insee sont, comme les appariements fiscaux, des appariements statistiques. L'appariement avec les différentes caisses (CNAF, CCMSA) est réalisé de la même manière. Il diffère cependant quelque peu. En effet, les appariements sont réalisés au niveau de l'ensemble des individus du ménage interrogé, et les informations sont ensuite réagrégées au niveau du ménage. Contrairement aux appariements fiscaux, il n'est pas possible de distinguer parmi les ménages non appariés si cela résulte d'un problème d'appariement ou si le ménage n'est pas allocataire de prestations sociales.

8 538 individus sur 16 373 individus de l'enquête soumis à l'appariement avec les fichiers sociaux (soit 52 %) ont été retrouvés dans ces fichiers comme percevant des prestations d'au moins une des caisses disponibles (CNAF et CCMSA).

IMPUTATION DES REVENUS MANQUANTS

Méthode d'imputation

L'objectif premier de ces enrichissements est d'estimer le niveau de vie des ménages, lequel est égal au revenu disponible divisé par le nombre d'unités de consommation (cf. Annexe II, Données diffusées). Le second objectif est de disposer des revenus d'activité des parents. Ainsi, on a imputé ces deux types de revenus aux ménages de parents non enrichis, dès lors que le (ou les) parent(s) se déclare (nt) actif(s) (en emploi ou au chômage) puisque dans ce cas des revenus sont attendus.

Il y a 178 ménages de parents actifs à la date de l'enquête qui n'ont pas été retrouvés dans les déclarations fiscales 2021 et qui font donc l'objet d'imputations[1].

Il est essentiel, dans la mise en œuvre des traitements, de distinguer selon le statut isolé ou en couple des parents, tel que déclaré à l'enquête : en effet dans le second cas et contrairement au premier, des revenus de conjoints sont attendus si ces derniers sont actifs. Deux catégories de ménages sont donc traitées distinctement. D'un côté les répondants à l'enquête qui vivent seuls (familles monoparentales), de l'autre les répondants qui sont en couple et dont le conjoint vit dans le logement. Lorsque le conjoint ne vit pas dans le logement, ces ménages sont traités comme les familles monoparentales (on ne dispose d'aucune information sur le conjoint vivant ailleurs).

Nous avons testé trois méthodes d'imputation, l'objectif étant de choisir la plus performante. Ces méthodes sont non paramétriques c'est-à-dire qu'elles ne sont pas régies par des lois de probabilités paramétriques et ne font donc pas de supposition sur la distribution des données. Ces trois méthodes sont le hot-deck aléatoire par classes, les k plus proches voisins (KNN) et les forêts aléatoires. Pour choisir la méthode d'imputation la plus appropriée pour nos données, nous mettons en œuvre le procédé suivant d'évaluation des différentes méthodes d'imputation. Ce procédé consiste à simuler aléatoirement des valeurs de revenus manquantes au sein des ménages pour lesquels les revenus sont renseignés. On mesure les performances des différentes méthodes en comparant les valeurs imputées (pour ces valeurs manquantes simulées) aux vraies valeurs, c'est-à-dire les valeurs connues (voir l'annexe 1 pour la description des évaluations menées et des résultats obtenus). Au vu des résultats obtenus à l'issue de la mise en œuvre de ce procédé d'évaluation, la méthode la plus appropriée est celle des forêts aléatoires décrite ci-dessous. À l'issue de l'imputation par forêts aléatoires, on a une estimation du revenu disponible du ménage du parent, du revenu imposable (utile pour le calcul des coûts afférents aux différents modes d'accueil des jeunes enfants) ainsi que de son revenu d'activité propre et le cas échéant du revenu d'activité de son conjoint.

Nous utilisons donc une méthode de complétion basée sur les forêts aléatoires appelée missForest (package R), proposée par Stekhoven et Bühlmann (2011). Une « *Random Forest* » (ou forêt aléatoire) est une technique de *Machine Learning*. La technique d'imputation de missForest est basée sur l'algorithme *Random Forest* qui de par son caractère non paramétrique ne fait pas d'hypothèses explicites sur la forme de la fonction, mais tente plutôt d'estimer la fonction de la manière la plus proche des points de données. L'algorithme construit un modèle de forêts aléatoires, pour chaque variable, basé sur les données disponibles, puis utilise le modèle pour prédire les valeurs manquantes.

Dans *Random Forest* il y a d'abord le mot « *Forest* » (forêt). Cet algorithme va donc reposer sur des arbres que l'on appelle arbre de décision. Ces forêts associent deux idées : les arbres de décision et le bootstrap aggregating – bagging (l'agrégation de modèles).

Arbres de décision

Les arbres de décision - arbres de régression dans notre cas, de classification dans le cas de variables catégorielles à imputer - sont des modèles de prédiction (Loh, 2011). Ils sont bâtis à partir des données observables et leur structure est similaire à celle d'un organigramme. À chaque nœud de l'arbre, un test est réalisé sur une ou plusieurs variables (questions de l'enquête). Chaque branche représente le résultat du test correspondant au nœud qui précède la branche et chaque feuille (extrémité de l'arbre) est un résultat possible (une décision finale).

La figure ci-dessous, extraite de Jérémy Robert dans Random Forest : Forêt d'arbre de décision- Définition et fonctionnement, présente un schéma général d'arbre de décision appliqué à des variables catégorielles (notion de nœud, branche, feuille). Pour des variables catégorielles le résultat final est la réponse à une question.

[1] Les ménages n'ayant été retrouvés que partiellement dans les déclarations de revenus ne sont pas imputés (cf. supra pour la définition des ménages retrouvés partiellement).

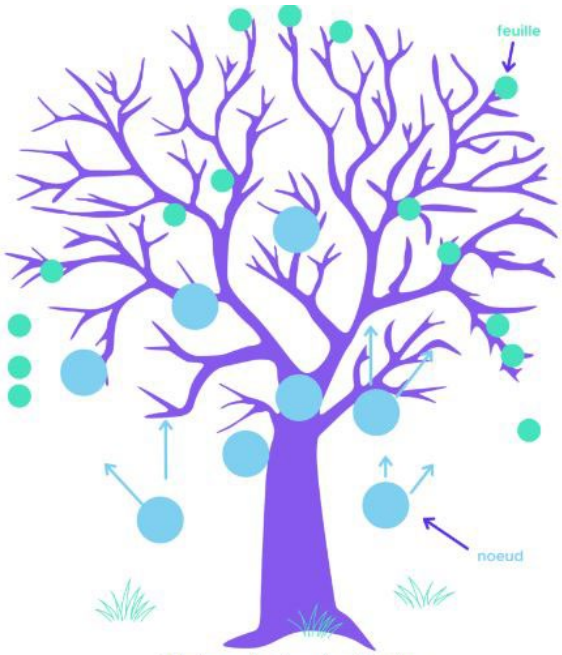
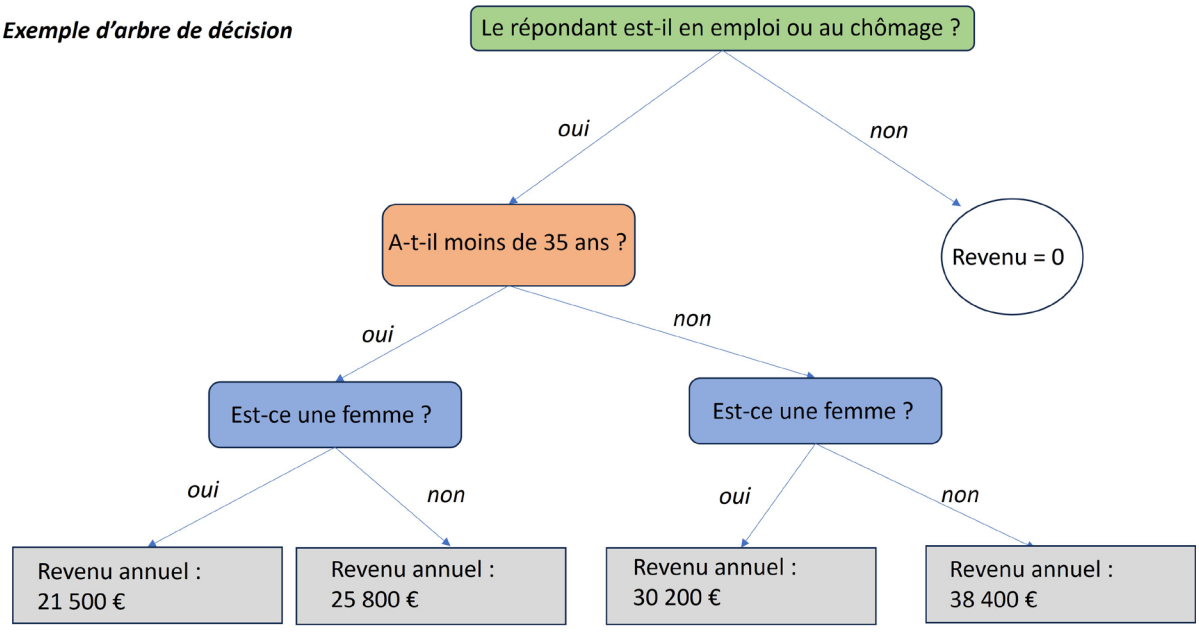


Schéma d'arbre de décision

*Sur l'arbre, **chaque question correspond à un noeud** c'est-à-dire à **un endroit où une branche se sépare en deux branches**. En fonction de la réponse à chaque question, nous allons nous orienter vers telle ou telle branche de l'arbre pour finalement **arriver sur une feuille** de l'arbre (ou extrémité) qui contiendra la réponse à notre question.*

Dans le cas d'un arbre de régression, le résultat final est une valeur, ici le revenu. La figure ci-dessous illustre un exemple d'arbre de régression avec le cas simple de trois variables auxiliaires de notre enquête (variables de test).

Exemple d'arbre de décision



Les tests effectués au niveau des nœuds ne sont pas arbitraires. Ils sont faits en sorte d'optimiser le gain d'information tiré du résultat du test. En effet, comment choisit-on l'ordre des questions à poser : pourquoi commencer par telle ou telle question ? Le principe est le suivant : à chaque nœud, l'algorithme se pose la question de savoir quelle question poser c'est-à-dire si on doit plutôt s'intéresser au statut face à l'emploi, à l'âge ou au sexe dans notre cas. Il va calculer pour chaque caractéristique le gain d'information que l'on obtiendrait si l'on choisissait cette

caractéristique. Puisqu'il faut maximiser le gain d'information, l'algorithme choisit la question et donc la caractéristique qui maximise ce gain.

Bagging

La spécificité de l'approche des forêts aléatoires de Breiman (1996, 2001) est de reposer sur le bagging. C'est-à-dire qu'à partir de l'ensemble des données observables, au lieu de générer un seul arbre de décision, t arbres de décision (indépendants) sont générés à partir de t sous-ensembles aléatoires de l'ensemble des données. Le résultat recherché est ensuite obtenu en effectuant une moyenne sur tous les arbres de décision. Le fait d'agréger les arbres de décision augmente la stabilité, la précision et la robustesse du modèle de prédiction. En effet, le bagging équilibre l'influence des points dans l'ensemble des données et réduit donc le biais qui pourrait être induit par l'un d'eux.

Algorithme de missForest

Les étapes de l'algorithme sont les suivantes.

- Étape 1 : une première imputation (temporaire) naïve est effectuée, par défaut une complétion par la moyenne, afin d'obtenir un échantillon d'apprentissage complet.
- Étape 2 : les variables sont classées par ordre croissant du pourcentage de données manquantes.
- Étape 3 : les données précédemment imputées sont réinitialisées à « manquantes » pour une seule des variables de la base de données.
- Étape 4 : les données manquantes de cette variable sont ensuite imputées à nouveau grâce à une forêt aléatoire (bâtie à partir des variables auxiliaires).
- Étape 5 : les étapes 3 et 4 sont répétées pour toutes les variables à imputer jusqu'à ce que chaque donnée manquante soit imputée. À cet instant, une itération est terminée.
- Étape 6 : les étapes 3 jusqu'à 5 sont répétées jusqu'à ce que la différence entre la matrice dernièrement imputée et la matrice imputée à l'itération précédente cesse de diminuer.

Plus précisément, on a les valeurs manquantes de quatre variables à imputer $Y_i,\ i = 1,2,3,4$[2]. Lors de la première itération, à l'étape 3, les données de la variable Y_1 sont réinitialisées à manquantes. Ces données manquantes sont ensuite imputées à l'étape 4 grâce à une forêt aléatoire (*via* des variables auxiliaires $X_j,\ j = 1,\dots,k$). Lors de la deuxième itération, les valeurs de la variable Y_2 sont mises à manquantes (étape 3) dans la matrice construite à la première itération. Les valeurs manquantes de Y_2 sont imputées (étape 4) à partir de cette matrice (dans laquelle les valeurs manquantes de la variable Y_1 ont été imputées précédemment).

L'algorithme s'exécute donc de façon itérative tout en mettant à jour la matrice des variables imputées et en vérifiant sa propre performance entre deux itérations. La vérification se fait en comparant le résultat de la valeur de remplacement actuelle avec la valeur précédente et en s'arrêtant aussitôt que la différence augmente.

Les forêts aléatoires ont été reconnues comme particulièrement efficaces en matière de prédictions dans de nombreux cas (par exemple Dagdoug, Goga et Haziza, 2021). De plus des études comparatives ont révélé que sur le plan des erreurs d'imputation, missForest a été la méthode la plus performante.

Les travaux menés ici montrent aussi que cette méthode est la plus performante (annexe 1), c'est pourquoi nous l'avons choisie.

Résultats des imputations de revenus

Sont présentés dans cette partie le bilan de la mise en œuvre des imputations[3] par groupe d'imputation (parents vivant seuls, parents en couple), ainsi que des éléments de distribution du revenu disponible des ménages et des revenus individuels d'activité (salaires, revenus non salariaux et allocations chômage) avant et après imputations (les résultats sont pondérés).

Revenus des parents vivant seuls

1 303 parents vivent seuls. On n'impute un revenu d'activité et un revenu disponible qu'aux ménages de parents vivant seuls se déclarant actifs (en emploi ou au chômage). 992 répondants vivant seuls sont en emploi ou au

[2] Imputations des salaires, allocations chômage, revenus non salariaux et revenu disponible.

[3] 178 ménages de parents font l'objet d'imputations.

chômage. Parmi eux, 60, soit 6 % se voient imputer des revenus (revenu disponible de leur ménage et revenus d'activité du parent actif).

Tableau 1 Distribution du revenu disponible des ménages de parents actifs vivant seuls, avant et après imputation (en euros annuels 2021)

Distribution	Ensemble des ménages de parents vivant seuls enrichis	Ensemble des ménages de parents vivant seuls (imputés et enrichis)	Écart de revenus : ménages de parents vivant seuls ensemble/enrichis (en %)
Moyenne	24 330	24 280	-0,2
Premier quartile (Q1)	19 330	19 460	0,7
Médiane	25 430	25 300	-0,5
Troisième quartile (Q3)	32 000	31 880	-0,4
Effectifs concernés	932	992	-

Lecture > Le revenu disponible moyen s'élève à 24 330 euros pour les répondants actifs vivant seuls ayant été enrichis, contre 24 280 euros pour ceux enrichis et imputés.
Champ > Ménages de parents vivant seuls, dont le parent répondant est actif.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

La distribution du revenu disponible des ménages de parents vivant seuls n'est pas affectée par les imputations réalisées.

Tableau 2 Distribution du revenu d'activité des parents actifs vivant seuls, avant et après imputation (en euros annuels 2021)

Distribution	Ensemble des ménages de parents vivant seuls enrichis	Ensemble des ménages de parents vivant seuls (imputés et enrichis)	Écart de revenus : ménages de parents vivant seuls ensemble/enrichis (en %)
Moyenne	17 660	17 500	-0,8
Premier quartile (Q1)	7 040	8 070	14,7
Médiane	15 910	15 890	-0,2
Troisième quartile (Q3)	23 720	23 450	-1,1
Effectifs concernés	932	992	-

Lecture > Le revenu d'activité moyen s'élève à 17 660 euros pour les répondants actifs vivant seuls ayant été enrichis, contre 17 500 euros pour ceux enrichis et imputés.
Champ > Ménages de parents vivant seuls, dont le parent répondant est actif.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

Les revenus d'activité des parents actifs vivant seuls situés dans le bas de la distribution sont plus élevés quand on tient compte des imputations. Le premier quartile est supérieur de 15 % à celui des ménages de parents vivant seuls enrichis. En effet, parmi les parents seuls enrichis, 19 % ont des revenus d'activité nuls (ceci est un peu contre-intuitif car ces parents sont actifs). Les imputations elles affectent un revenu d'activité strictement positif à ceux non enrichis puisqu'ils sont actifs. Ceci tire donc la distribution vers le haut. Toutefois, le reste de la distribution n'est pas significativement modifié.

Revenus des parents vivant en couple

On compte 7 484 ménages de parents vivant en couple. 7 393 ménages ont l'un des deux parents qui est actif.

Parmi eux, 118, soit 2 % seulement se voient imputer des revenus, soit le revenu disponible de leur ménage et les revenus d'activité des parents actifs.

Tableau 3 Distribution du revenu disponible des ménages de parents en couple actifs avant et après imputations (en euros annuels 2021)

Distribution	Ensemble des ménages de parents vivant en couple enrichis	Ensemble des ménages de parents vivant en couple (imputés et enrichis)	Écart de revenus : ménages de parents vivant en couple ensemble/enrichis (en %)
Moyenne	53 850	53 560	-0,5
Premier quartile (Q1)	37 320	37 160	-0,4
Médiane	48 700	48 490	-0,4
Troisième quartile (Q3)	62 870	62 560	-0,5
Effectifs concernés	7 063	7 181	-

Lecture > Le revenu disponible moyen s'élève à 53 850 euros pour les couples ayant été enrichis, contre 53 560 euros pour ceux enrichis et imputés.
Champ > Ménages de parents vivant en couple, dont le répondant et/ou le conjoint est actif.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

Les imputations réalisées ne modifient pas la distribution du revenu disponible des ménages de parents en couple.

Tableau 4 Distribution du revenu d'activité des répondants actifs en couple avant et après imputation

Distribution	Ensemble des répondants en couple enrichis	Ensemble des répondants en couple (imputés et enrichis)	Écart de revenus d'activité des répondants en couple ensemble/enrichis (en %)
Moyenne	22 990	22 940	-0,3
Premier quartile (Q1)	12 810	12 800	-0,1
Médiane	20 850	20 760	-0,4
Troisième quartile (Q3)	29 380	29 320	-0,2
Effectifs concernés	7 063	7 181	-

Lecture > Le revenu d'activité moyen s'élève à 22 994 euros pour les répondants actifs en couple ayant été enrichis, contre 22 936 euros pour ceux enrichis et imputés.
Champ > Ménages de parents vivant en couple, dont le répondant et/ou le conjoint est actif.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

La distribution des revenus d'activité des répondants vivant en couple n'est pas affectée par les imputations réalisées.

Tableau 5 **Distribution du revenu d'activité des conjoints actifs avant et après imputation (en euros annuels 2021)**

Distribution	Ensemble des conjoints enrichis	Ensemble des conjoints (imputés et enrichis)	Écart de revenus d'activité des conjoints ensemble/enrichis (en %)
Moyenne	26 460	26 390	-0,3
Premier quartile (Q1)	15 590	15 510	-0,5
Médiane	22 780	22 740	-0,2
Troisième quartile (Q3)	32 140	32 070	-0,2
Effectifs concernés	7 063	7 181	-

Lecture > Le revenu d'activité moyen s'élève à 26 460 euros pour les conjoints actifs ayant été enrichis, contre 26 390 euros pour ceux enrichis et imputés.
Champ > Ménages de parents vivant en couple, dont le répondant et/ou le conjoint est actif.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

Les imputations ne modifient pas la distribution des revenus d'activité des conjoints actifs.

POUR EN SAVOIR PLUS

Andridge, R. R., Little, R. J. (2009). The Use of Sample Weights in Hot Deck Imputation. *Journal of Official Statistics*, vol. 25, n°1, 21-36.

Arnold, C., Missègue, N. (2017). Appariement fiscal et social de l'enquête Bénéficiaires de minima sociaux (2012), Imputations Post Appariement. Drees, *Documents de travail,* série Sources et Méthodes, 64.

Boneschi, S., Missègue, N. (2021). L'estimation des revenus des seniors dans l'enquête CARE-Institutions, Imputations post-appariements fiscaux et sociaux. Drees, *Les Dossiers de la Drees,* série Méthodologie, 82.

Bousri I., Salah S. Arab, B., M. (2021). Validation d'une méthode d'imputation de données manquantes pour la reconstitution des séries de température. JAMA, vol. 5, 28-32.

Breiman (1996). Bagging Predictors. Machine Learning, vol. 24, 123-140.

Breiman (2001). Random Forest. Statistics Department, University of California, Berkeley, CA 94720.

Burricand, C. (2013). Transition from survey data to registers in France for Silc survey. In The use of registers in the context of EU-SILC: challenges and opportunities, *Eurostat Working Papers*, 2013 edition, 111-124.

Caron, N. Correction de la non-réponse par imputation ou par repondération. Insee*, Documents de travail, méthodologie statistique*, n°M0502.

Couvert, N., Missègue, N. (2019). L'estimation des revenus des seniors dans l'enquête CARE-Ménages - Imputations postappariements fiscaux et sociaux. Drees, *Documents de travail,* série Sources et Méthodes, 72.

Dagdoug M., Goga C. et Haziza D. (2021). Imputation par forêts aléatoires en théorie des sondages. Journées Théorie, Modélisation et Simulation.

Davezies, L., D'Haultfoeuille, X. (2009). Faut-il pondérer ? Ou l'éternelle question de l'économètre confronté à des données de sondage. Insee*, Documents de travail,* n°G2009/06.

Haziza, D. (2002). Inférence en présence d'imputation : un survol. Insee, Journées de méthodologie statistique.

Loh, W.-Y. (2011). Classification and regression trees. *WIREs Data Mining and Knowlegge Discovery*. John Wiley & Sons, Inc, vol. 1.

Stekhoven, D.J., Bühlmann P. (2012). MissForest - non-parametric missing imputation for mixed-type data. Bioinformatics. Vol. 28, n°1, 112-118.

Annexe 1. Performance des méthodes d'imputation

Nous évaluons les performances des méthodes que nous avons testées pour le revenu disponible qui est la principale grandeur cible (le revenu disponible permet de calculer le niveau de vie). Outre les forêts aléatoires, ces méthodes sont l'imputation par hot deck et par les K plus proches voisins (KNN). Le hot deck consiste à attribuer la valeur d'un répondant enrichi (donneur), sélectionné au hasard parmi l'ensemble des répondants, en remplacement de la valeur manquante du répondant non-enrichi (receveur). Pour les K plus proches voisins on affecte à la donnée manquante de l'individu i la médiane des valeurs des k plus proches voisins (k à définir et une métrique à choisir pour la notion de voisins).

Pour choisir la méthode d'imputation la plus appropriée pour nos données, nous mettons en œuvre le procédé d'évaluation des différentes méthodes d'imputation exposé ci-dessous.

Cette évaluation entre dans le cadre d'une simulation de Monte-Carlo. Il s'agit de simuler aléatoirement des valeurs manquantes (sur le sous-ensemble des ménages pour lesquels les revenus attendus sont renseignés) et de répéter l'expérience ci-dessous un grand nombre de fois. On va mesurer les performances des différentes méthodes en comparant les valeurs imputées aux valeurs connues.

On applique ce procédé pour chaque méthode envisagée :

- Pour le revenu disponible, des non-réponses sont tirées dans chacun des sous-échantillons constitués (9 sous-échantillons pour les répondants vivant seuls, chacun comportant 500 ménages, et 10 sous-échantillons pour les couples, de 700 ménages chacun[4]). Elles sont tirées aléatoirement.
- Les méthodes d'imputation à évaluer sont mises en œuvre sur chaque échantillon ainsi obtenu.
- Pour chacun des sous-échantillons avec valeurs manquantes imputées, les statistiques suivantes sont calculées :
  - La moyenne estimée du revenu ;
  - La médiane des erreurs (en valeur absolue) commises sur les valeurs imputées ;
  - Le maximum des erreurs (en valeur absolue) commises sur les valeurs imputées.
- L'ensemble de ces opérations est répété un grand nombre de fois. On a trois jeux de données simulées : dans le premier on a répété l'expérience 100 fois, 300 fois dans le deuxième et 500 fois dans le troisième.

À partir des trois statistiques calculées pour chaque sous-échantillon, on produit les trois indicateurs synthétiques suivants :

- L'écart-type Monte-Carlo de l'estimateur de la moyenne du revenu ;
- Le quantile à 95 % de la médiane des erreurs (en valeur absolue) ;
- Le quantile à 95 % du maximum des erreurs (en valeur absolue).

L'évaluation est donc menée sur plusieurs sous-échantillons. Concernant la comparaison des méthodes, les résultats qualitatifs de chaque sous-échantillon confirme la comparaison menée sur l'ensemble de l'échantillon (pour chaque sous-échantillon on réalise 100, 300 ou 500 simulations). Nous ne présentons ici que les résultats menés sur l'ensemble de l'échantillon.

Pour chaque méthose les mêmes variables auxiliaires sont bien sûr introduites. Ont été sélectionnées celles qui ressortent comme significatives dans une analyse de la variance du revenu disponible.

On utilise le calcul parallèle pour éviter des temps de calcul trop longs pour les forêts aléatoires. En effet, le calcul parallèle est l'exécution simultanée de différentes parties d'un calcul plus important sur plusieurs processeurs ou cœurs de calcul.

[4] Le nombre d'échantillons multiplié par le nombre d'observations dans chaque sous-échantillons (chaque sous-échantillon est de même taille) doit être inférieur au nombre d'observations dans la base totale.

Tableaux 1 **Performances des méthodes pour les répondants vivant seuls, selon le nombre de simulations**

100 simulations

Méthode	Revenu disponible	Indicateur
Hot deck	255	Écart-type de l'estimation de la moyenne
KNN	258	Écart-type de l'estimation de la moyenne
Forêts aléatoires	245	Écart-type de l'estimation de la moyenne
Hot deck	7 173	95 % de la médiane des erreurs
KNN	11 822	95 % de la médiane des erreurs
Forêts aléatoires	4 604	95 % de la médiane des erreurs
Hot deck	233 423	95 % du maximum des erreurs
KNN	324 039	95 % du maximum des erreurs
Forêts aléatoires	613 858	95 % du maximum des erreurs

Lecture > Pour les répondants vivant seuls l'écart-type de l'estimation de la moyenne est de 245 avec les forêts aléatoires, contre 255 pour le hot deck et 258 pour les K plus proches voisins (KNN) avec 100 simulations.
Champ > Ménages de parents vivant seuls.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

300 simulations

Méthode	Revenu disponible	Indicateur
Hot deck	232	Écart-type de l'estimation de la moyenne
KNN	288	Écart-type de l'estimation de la moyenne
Forêts aléatoires	236	Écart-type de l'estimation de la moyenne
Hot deck	7 246	95 % de la médiane des erreurs
KNN	12 536	95 % de la médiane des erreurs
Forêts aléatoires	4 407	95 % de la médiane des erreurs
Hot deck	314 779	95 % du maximum des erreurs
KNN	324 039	95 % du maximum des erreurs
Forêts aléatoires	599 137	95 % du maximum des erreurs

Lecture > Pour les répondants vivant seuls l'écart-type de l'estimation de la moyenne est de 236 avec les forêts aléatoires, contre 232 avec le hot deck et 288 pour les K plus proches voisins (KNN) avec 300 simulations.
Champ > Ménages de parents vivant seuls.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

500 simulations

Méthode	Revenu disponible	Indicateur
Hot deck	234	Écart-type de l'estimation de la moyenne
KNN	247	Écart-type de l'estimation de la moyenne
Forêts aléatoires	191	Écart-type de l'estimation de la moyenne
Hot deck	7 516	95 % de la médiane des erreurs
KNN	11 992	95 % de la médiane des erreurs
Forêts aléatoires	4 543	95 % de la médiane des erreurs
Hot deck	314 779	95 % du maximum des erreurs
KNN	324 039	95 % du maximum des erreurs
Forêts aléatoires	284 690	95 % du maximum des erreurs

Lecture > Pour les répondants vivant seuls l'écart-type de l'estimation de la moyenne est de 191 avec les forêts aléatoires, contre 234 pour le hot deck et 247 pour les K plus proches voisins (KNN) avec 500 simulations.
Champ > Ménages de parents vivant seuls.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

Tableaux 2 Performances des méthodes pour les couples, selon le nombre de simulations

100 simulations

Méthode	Revenu disponible	Indicateur
Hot deck	178	Écart-type de l'estimation de la moyenne
KNN	113	Écart-type de l'estimation de la moyenne
Forêts aléatoires	65	Écart-type de l'estimation de la moyenne
Hot deck	23 751	95 % de la médiane des erreurs
KNN	15 456	95 % de la médiane des erreurs
Forêts aléatoires	7 190	95 % de la médiane des erreurs
Hot deck	215 779	95 % du maximum des erreurs
KNN	134 680	95 % du maximum des erreurs
Forêts aléatoires	71 531	95 % du maximum des erreurs

Lecture > Pour les répondants en couple l'écart-type de l'estimation de la moyenne est de 65 avec les forêts aléatoires, contre 113 pour les K plus proches voisins (KNN) et 178 pour le hot deck avec 100 simulations.
Champ > Ménages de parents en couple.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

300 simulations

Méthode	Revenu disponible	Indicateur
Hot deck	191	Écart-type de l'estimation de la moyenne
KNN	118	Écart-type de l'estimation de la moyenne
Forêts aléatoires	68	Écart-type de l'estimation de la moyenne
Hot deck	23 246	95 % de la médiane des erreurs
KNN	14 301	95 % de la médiane des erreurs
Forêts aléatoires	7 484	95 % de la médiane des erreurs
Hot deck	214 622	95 % du maximum des erreurs
KNN	146 160	95 % du maximum des erreurs
Forêts aléatoires	89 975	95 % du maximum des erreurs

Lecture > Pour les répondants en couple l'écart-type de l'estimation de la moyenne est de 68 avec les forêts aléatoires, contre 118 pour les K plus proches voisins (KNN) et 191 pour le hot deck avec 300 simulations.
Champ > Ménages de parents en couple.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

500 simulations

Méthode	Revenu disponible	Indicateur
Hot deck	188	Écart-type de l'estimation de la moyenne
KNN	118	Écart-type de l'estimation de la moyenne
Forêts aléatoires	66	Écart-type de l'estimation de la moyenne
Hot deck	23 463	95 % de la médiane des erreurs
KNN	15 503	95 % de la médiane des erreurs
Forêts aléatoires	7 656	95 % de la médiane des erreurs
Hot deck	203 946	95 % du maximum des erreurs
KNN	175 359	95 % du maximum des erreurs
Forêts aléatoires	85 292	95 % du maximum des erreurs

Lecture > Pour les répondants en couple l'écart-type de l'estimation de la moyenne est de 66 avec les forêts aléatoires, contre 118 pour les K plus proches voisins (KNN) et 188 pour le hot deck avec 500 simulations.
Champ > Ménages de parents en couple.
Source > Drees, enquête Modes de garde et d'accueil des jeunes enfants 2021 ; Insee-DGFiP-Cnaf-Cnav-CCMSA.

La performance mesurée par l'écart-type de l'estimation de la moyenne est meilleure avec la méthode des forêts aléatoires pour les couples. Il en est de même pour les répondants vivant seuls, lorsque l'on réalise 100 ou 500 simulations. De plus toutes les simulations (100, 300, 500) menées sur chacun des sous-échantillons confirment cette meilleure performance de la méthode des forêts aléatoires aussi bien pour les répondants vivant seuls que pour les couples.

En termes de quantiles à 95 % de l'erreur médiane commise, les résultats confirment ce diagnostic quel que soit le type de ménage. Concernant le quantile à 95 % de l'erreur maximal commise ce résultat se confirme pour les couples. Il se confirme pour les répondants vivant seuls sur l'ensemble de l'échantillon avec un grand nombre de simulations (500). Sur les sous-échantillons (quel que soit le nombre de simulations), ce sont les imputations par forêts aléatoires qui présentent de meilleures performances dans la grande majorité des cas.

Annexe 2. Données diffusées

Tous les montants diffusés sont des montants annuels, relatifs à l'année 2021. Sont également mis à disposition des indicateurs précisant si les montants sont issus de l'appariement (répondants, conjoint retrouvés) ou bien sont imputés en partie ou en totalité. On met à disposition le montant des revenus individuels des répondants et de leurs éventuels conjoints. Ces revenus comprennent les revenus d'activité et de remplacement (revenus du travail, du chômage, les revenus non salariaux - revenus agricoles, revenus industriels et commerciaux et revenus non commerciaux) ainsi que les pensions alimentaires reçues et les rentes viagères à titre onéreux.

On met également à disposition les montants suivants. Sont également fournis des indicateurs précisant si les montants sont issus de l'appariement (répondants, conjoint retrouvés) ou bien sont imputés en partie ou en totalité.

- Le revenu d'activité du répondant (salaires + allocations chômage + revenus non salariaux[5]) et celui du conjoint.
- Le revenu d'activité de tous les membres du ménage.
- Le revenu individuel du répondant et celui du conjoint. Pour obtenir le revenu individuel, on ajoute au revenu d'activité les pensions alimentaires reçues et les rentes viagères à titre onéreux.
- Le revenu individuel de l'ensemble des membres du ménage.
- La pension alimentaire perçue par le répondant et celle perçue par le conjoint.
- Les pensions alimentaires versées (au niveau du ménage).
- Les prestations sociales perçues par les ménages. Elles comprennent les montants suivants :
  - Allocations familiales
  - Complément familial
  - Allocation de rentrée scolaire
  - Allocation d'éducation enfant handicapé
  - Allocation de soutien familial
  - Allocation journalière de présence parentale
  - Allocation de base de la PAJE
  - Allocation aux adultes handicapés
  - Compléments d'AAH
  - Prime d'activité
  - Revenu de solidarité active
  - Allocation logement familial
  - Allocation logement social
  - Aide personnalisée au logement
  - Prime exceptionnelle de fin d'année RSA
  - PREPARE et PREPARE majorée (+ CLCA)
  - Primes de naissance ou d'adoption
  - Revenu de solidarité Outre-mer
  - Allocation logement sociale étudiants
  - Minimum vieillesse MSA
- La PREPARE, la PREPARE majorée et le complément de libre choix d'activité (CLCA).
- Le revenu disponible des ménages de parents. Il comprend :
  - les revenus individuels des membres du ménage ;
  - les revenus non individualisables du ménage : revenus fonciers, revenus perçus à l'étranger, revenus non soumis au prélèvement libératoire (PL), revenus de valeurs mobilières soumis au PL auxquels on retire les pensions alimentaires versées ;
  - les prestations sociales (prestations familiales, minima sociaux, prestations logement) ;
  - les revenus des produits financiers exonérés d'impôt. Ces revenus ne figurent pas sur la déclaration de revenus. On a procédé à l'estimation[6] des montants des placements détenus par les ménages à partir des réponses à l'enquête, ainsi que des intérêts qu'ils génèrent.
    Sont retirés de la somme de ces revenus : l'impôt sur les revenus 2021, la taxe d'habitation, la CSG non déductible (imposable) et la CRDS (toujours imposable) sur les revenus d'activité et de remplacement, le PL sur valeurs mobilières, les prélèvements sociaux sur les revenus du patrimoine sans PL, les prélèvements sociaux sur valeurs mobilières avec PL, les prélèvements sociaux sur les revenus des placements exonérés d'impôt.

[5] Il peut arriver que les revenus non salariaux soient négatifs, lorsque l'activité indépendante conduit à un déficit (et non à un bénéfice).

[6] L'imputation est réalisée avec la méthode des résidus simulés. Elle consiste à imputer un montant d'épargne au ménage, ce montant doit se situer à l'intérieur de la tranche de montant d'épargne détenue déclarée par le répondant.

- Le niveau de vie du ménage. Le niveau de vie est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation (UC)[7]. Le niveau de vie est donc le même pour tous les individus d'un même ménage.
- Le nombre d'unités de consommation (UC) du ménage. Les unités de consommation sont calculées selon l'échelle d'équivalence dite de l'OCDE modifiée qui attribue 1 UC au premier adulte du ménage, 0,5 UC aux autres personnes de 14 ans ou plus et 0,3 UC aux enfants de moins de 14 ans.

[7] Pour comparer les niveaux de vie de ménages de taille ou de composition différente, on divise le revenu par le nombre d'unités de consommation (UC). Cette échelle d'équivalence (dite de l'OCDE) tient compte des économies d'échelle au sein du ménage. En effet, les besoins d'un ménage ne s'accroissent pas en stricte proportion de sa taille. Lorsque plusieurs personnes vivent ensemble, il n'est pas nécessaire de multiplier tous les biens de consommation (en particulier, les biens de consommation durables) par le nombre de personnes pour garder le même niveau de vie.

Drees Méthodes
N° 26 • avril 2026

Estimation des revenus des parents
dans l'enquête
Modes de garde et d'accueil des jeunes enfants 2021

Directeur de la publication
Thomas Wanecq

Responsable d'édition
Valérie Bauer-Eubriet

ISSN
2740-3564