Journées de méthodologie statistique de l'Insee 2022

Diffuser une base anonymisée : utopie ou réalitée ?

Julie Djiriguian (*), Nathalie Missègue (**), Layla Ricroch (***), Alii

(*) Unité innovation et sécurité des données
(**) Cheffe de projet statistique
(***) Bureau Handicap dépendance
Drees

julie.djiriguian@sante.gouv.fr
nathalie.missegue@sante.gouv.fr
layla.ricroch@sante.gouv.fr

Mots-clés : Anonymisation, confidentialité, méthodes perturbatrices, open data.

Domaines. 11. Institutionnel, open science (Anonymisation).

Résumé

Cet article décrit le processus mis en œuvre à la Drees pour anonymiser une base de données individuelles de nature administrative et quasi-exhaustive. Il s'agit de remontées individuelles (RI) de données sur les bénéficiaires de l'allocation personnalisée d'autonomie (APA) vivant à domicile en 2017. Cette base de données provenant des systèmes de gestion des conseils départementaux est nommée ici RI-APA à domicile ou plus simplement RI-APA. L'objectif est d'en diffuser une version en *open data*, pour suivre les préconisations du rapport Bothorel « Pour une politique publique de la donnée » [1] et permettre ainsi un premier accès sans délai à des chercheurs en attente des autorisations d'accès et de traitement aux données confidentielles exhaustives - démarches plus longues du fait des contraintes spécifiques aux données de santé. L'intérêt est également de fournir une base ouverte sur laquelle faire tourner le modèle de microsimulation Autonomix qui a été diffusé en *open code* (Ouverture du code du modèle de microsimulation AUTONOMIX sous GitLab [2]). L'approche conduite ici pour l'anonymisation s'appuie sur des travaux réalisés à l'Insee et présentés aux JMS 2015 ([3], [4]).

Il s'agit, en s'appuyant sur la théorie et en ayant échangé avec l'Insee sur ces questions [1], de mettre en œuvre des méthodes de floutage/retraitements des données qui permettent de protéger le fichier des trois principaux risques. Nous nous sommes focalisés sur la minimisation du risque de ré-identification (*identity disclosure*). Pour autant, le risque de divulgation d'attribut et de révélation inférentielle (*attribute disclosure and probabilistic attack*) ont été considérés et traités.

1. Département des méthodes statistiques.

L'approche générale retenue pour anonymiser la base de données des RI-APA à domicile fait partie de la famille des méthodes de génération de données partiellement synthétiques. Á notre connaissance, la méthode perturbatrice que nous utilisons pour flouter certaines variables quantitatives ne semble pas être habituellement utilisée dans ce type de travaux. La cohérence entre données floutées et données originelles a été vérifiée et les choix méthodologiques effectués ont également été validés en comparant les résultats obtenus avec la base anonymisée à ceux émanant de la base de données confidentielles.

Plus précisément, la validation de notre démarche consiste en un contrôle du niveau de confidentialité garanti par notre processus (k-anonymat, all-m et l-diversité). Le risque de ré-identification est également calculé. Enfin, les résultats obtenus avec la base anonymisée sont comparés à ceux produits (publiés ou non) avec la base de données confidentielles : statistiques descriptives, régressions, simulations à partir du modèle de microsimulation Autonomix [2]. Ces comparaisons montrent les avantages et limites en termes d'études et analyses réalisables à partir d'une telle base de données anonymisée. L'ensemble de ces éléments nous permet de juger de l'opportunité de la diffuser en *open data* et d'itérer ce processus sur d'autres bases de données individuelles de la Drees.

Abstract

This article describes the process implemented by Drees to anonymize a database of individual, quasi-exhaustive administrative data. It contains information on almost all the beneficiaries of the personalised autonomy allowance (APA), collected from the French departmental councils. The dissemination of this large sample in *open data* has several purposes, such as running the "Autonomix" microsimulation model, whose code has also been released on the GitLab platform. We implemented data blurring/removal methods to protect the file from three main risks : identity disclosure, attribute disclosure and probabilistic attack. Finally, the results obtained with the anonymized database are compared to those produced with the initial database. The comparisons show the advantages and limits of the analyses that can be carried out with such an anonymized database.

Introduction

La mise à disposition par la Drees de cette première base de données individuelles anonymisées s'inscrit dans l'optique de faciliter le travail préparatoire des chercheurs pendant la période parfois longue nécessaire pour obtenir l'avis du Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES) pour les données de santé ou bien l'accès aux bases de données confidentielles diffusées par le Centre d'accès sécurisé aux données (CASD). Mettre à disposition une base constituée des données aux valeurs modifiées mais conservant le lien entre les variables leur permettra de préparer leurs scripts et donc leurs travaux de recherche avant d'obtenir l'accès à la base d'origine non anonymisée et exhaustive.

Cette ouverture des données s'inscrit également dans une démarche globale qui vise à offrir aux utilisateurs potentiels un ensemble de bases de données individuelles en *open data*. En effet, suite au rapport Bothorel « Pour une politique publique de la donnée » [1], rendu public le 23 décembre 2020, le ministère des Solidarités et de la santé s'est engagé à ouvrir 7 jeux de données en 2021/2022. Au sein de la Drees, la sous-direction Observation de la Solidarité a entrepris en

2. Le modèle Autonomix est un modèle de microsimulation statique sur la prise en charge de la perte d'autonomie des personnes âgées.

2021 l'élaboration d'une version anonymisée d'une première base de données. C'est ainsi que dans le cadre de sa politique d'ouverture des données et des codes, la Drees a mis à disposition de nouvelles bases de données, dont les RI-APA « floutées » ainsi que les codes sources de plusieurs de ses outils sur les thématiques liées au vieillissement de la population.

Pour la diffusion, on choisit de nommer cette base de données « base floutée » plutôt que « base anonymisée ». Ce dernier terme pourrait en effet laisser penser à l'utilisateur - non expert de cette question de la confidentialité - que la base de données confidentielles disponible au CASD comprend des variables qui ne garantiraient pas l'anonymat (par exemple les noms ou encore les prénoms).

Cette politique d'ouverture des données en *open data* impose de bien respecter la contrainte de confidentialité et soulève des enjeux en termes d'anonymisation. En effet, les utilisateurs de la base mise à disposition pourraient souhaiter calculer, ou simplement vérifier, des agrégats globaux sur les variables fournies. La conservation de ces résultats agrégés n'est pas garantie par le processus d'anonymisation choisi. La méthode vise en premier lieu à minimiser le risque de ré-identification *via* un contrôle du niveau de confidentialité garanti par nos méthodes (k-anonymat et all-m). Cependant, nous avons également veillé à nous prémunir contre les risques de divulgation d'attribut et de révélation inférentielle (en traitant la l-diversité et en procédant à un échantillonnage au 1/2). Enfin, avant de décider de diffuser (ou non) la base floutée, nous avons vérifié que les résultats sont comparables à ceux de la base de données confidentielles diffusée au CASD, notamment pour être certains que les agrégats calculés concordent bien *via* des statistiques descriptives, mais aussi en allant plus loin et en comparant des modèles de régression et des simulations avec le modèle de microsimulation Autonomix [2].

Ces travaux ont été menés par une équipe pluri-disciplinaire, réunissant des profils « métiers » (indispensables de par leur connaissance de la base de données et plus généralement de la thématique dépendance), d'autres plutôt « méthodes » (pour la connaissance des méthodes statistiques et l'investissement sur les techniques d'anonymisation) ainsi qu'informatiques (pour la création des scripts et fonctions en R). Ces travaux ont donné lieu à a la mise en place d'un projet partagé sous GitLab dont le contenu et l'architecture ont évolués au fil du temps et des choix effectués. L'avantage de ce procédé est sa reproductibilité dans le temps, mais aussi son adaptabilité à d'autres sources de données. Nous sommes convaincus que des travaux d'anonymisation peuvent difficilement être réalisés à l'aide d'outils « clé en main » et que la connaissance métier est cruciale.

1 Préalable à l'anonymisation

1.1 Focus sur la base de données confidentielles : les RI-APA à domicile

L'article 74 de la loi d'adaptation de la société au vieillissement (loi ASV) du 28 décembre 2015 inscrit dans la loi plusieurs remontées de données individuelles dont les remontées d'informations sur les bénéficiaires de l'APA à domicile qui sont celles qui nous intéressent ici. Elles ont pour objectif la collecte de données administratives individuelles extraites des systèmes d'informations des collectivités territoriales en charge de l'APA, à savoir les conseils départementaux (CD).

Les RI-APA 2017 ont permis de collecter, en 2018-2019, les données de 96 départements – seuls 4 n'ont pas été en mesure de transmettre les informations requises, Mayotte n'ayant par ailleurs pas été incluse dans l'opération. Les données recueillies concernent l'APA et portent sur

les personnes vivant à domicile ayant eu un droit à cette prestation, ouvert au moins un jour au cours de l'année 2017, ou ayant fait l'objet d'une évaluation pour l'obtenir.

L'APA à domicile prend en charge une partie des dépenses d'aide pour la réalisation des activités de la vie quotidienne. Les différents besoins sont recensés dans un plan d'aide mensuel individualisé, valorisé en euros, qui est notifié à chaque demandeur éligible, qualifié de « montant notifié ». Ce montant notifié est ensuite partiellement ou intégralement utilisé, et donc « consommé », par le bénéficiaire (on parlera alors de « montant consommé »). La majeure partie des plans d'APA à domicile sont consacrés à des aides humaines : ils servent à rémunérer une personne qui aide le bénéficiaire à accomplir les actes de la vie quotidienne (faire le ménage, préparer les repas, se lever, se laver, s'habiller, etc.).

Le fichier confidentiel des bénéficiaires de l'APA vivant à domicile, payés en décembre 2017, comporte près de 618 000 individus (moins de 1 % de ces bénéficiaires sont décédés dans le courant du mois de décembre). La version diffusée au CASD comprend au total environ 110 variables. Parmi les données personnelles, telles que définies sur le site de la CNIL (cf. annexe 1, page 25), il ne comporte aucune variable directement identifiante (nom, prénom) ni même le NIR (ou un numéro de bénéficiaire par exemple) pour ce qui est des variables indirectement identifiantes.

En revanche, plusieurs variables repérées comme indirectement identifiantes ou sensibles doivent être examinées de près : est-ce qu'on les diffuse ? est-ce qu'on les diffuse mais de manière « partielle » (par exemple, en regroupant des modalités d'une variable qualitative, en diffusant des tranches plutôt que la variable quantitative) ou bien est-ce qu'on les « floute » ? En effet, ces variables croisées entre elles et/ou associées à des données tierces, feraient peser un risque trop élevé et donc une rupture de confidentialité, d'autant plus qu'il s'agit d'une remontée exhaustive d'informations administratives.

1.2 Risques encourus et principes généraux

Avant d'entamer un processus de protection des données, il convient de s'interroger sur certaines hypothèses : que savent les « attaquants » potentiels ? que recherchent-ils comme informations qu'il conviendrait de protéger ? [3], [4]. De telles personnes peuvent avoir diverses motivations : retrouver une personne connue et divulguer dans la presse les informations la concernant (son patrimoine, par exemple), rechercher une personne pour des raisons professionnelles (un établissement de crédit qui souhaiterait retrouver cette personne afin de vérifier les montants de revenus qu'elle a déclarés à l'administration fiscale), obtenir des informations sur le niveau de dépendance d'une potentielle victime de l'attaque, ou encore pour des motifs personnels. Bien sûr, tous les « attaquants » potentiels n'ont pas le même niveau d'information. Par exemple, le citoyen « lambda », sauf si la personne recherchée est un parent ou un voisin, disposera *a priori* de moins d'informations précises sur la personne recherchée que, par exemple, l'établissement bancaire ou encore la société de crédit à la consommation de cette dernière.

On a cherché à protéger le fichier de trois risques principaux [5]. Le premier est le risque de divulgation d'identité (*identity disclosure*). Ce risque existe dès lors qu'un bénéficiaire de l'APA peut être identifié de manière sûre et certaine dans le fichier diffusé en *open data*. Cependant, le fait de retrouver cette personne n'apporte pas forcément plus d'informations à l'attaquant que celles qu'il a déjà. Le deuxième risque est un risque de divulgation d'attribut (*attribute disclosure*). Il est dû au fait de reconnaître un bénéficiaire en particulier (*via* des variables quasi-identifiantes, généralement des éléments socio-démographiques connus : l'âge, le sexe, le lieu

d'habitation, etc.) et d'obtenir ainsi des informations sensibles le concernant (ses ressources, le fait qu'il fait l'objet d'une mesure de protection juridique, etc.). Par exemple, les données du fichier diffusé nous apprennent que toutes les femmes de 80 ans bénéficiaires de l'APA, habitant dans tel département, ont des ressources inférieures à Y euros par mois. L'attaquant sait que Madame M., âgée de 80 ans vit dans ce département et perçoit l'APA : alors il pourra en déduire qu'elle dispose de moins de Y euros par mois pour vivre (même s'il ne sait pas à quelle ligne du fichier correspond Madame M.). Dévoiler cette information sensible peut être dommageable pour la personne concernée. Le troisième est un risque de révélation inférentielle (*probabilistic attack*). Dans ce cas, on ne se focalise pas sur le fait de savoir quelle observation (ni même quel attribut considéré comme sensible) l'attaquant peut relier à une personne « cible » (Madame M. par exemple). Ce risque est lié au fait que l'attaquant arrive, en utilisant la base floutée, à inférer, avec une probabilité importante, une information sensible dont il ne disposait pas auparavant.

Nous nous sommes surtout focalisés sur la protection contre le risque de ré-identification. Nous verrons dans la partie 4.5 et la partie 5, pages 16 et 17, que les risques de révélation d'attribut et de divulgation inférentielle ont aussi été traités.

1.3 Variables quasi-identifiantes et variables sensibles

Nous avons, dans un premier temps, considéré que certaines variables ne pouvaient pas être diffusées, soit parce qu'elles pourraient permettre de ré-identifier des bénéficiaires, soit parce qu'elles révéleraient des informations sur certaines caractéristiques des bénéficiaires. Ainsi, ont été écartées d'office de la diffusion les variables suivantes : la date de naissance précise, le code postal du lieu de domicile habituel (en raison de la localisation assez fine permise par cette information), le montant annuel de la pension de retraite figurant sur l'avis d'imposition et la nature de la protection juridique de la personne (tutelle, curatelle simple/renforcée, etc.) notamment. Enfin, d'autres variables, pour la plupart de nature purement administratives [3], n'ont pas été jugées utiles pour les utilisateurs potentiels, aussi elles ont été supprimées avant traitement de la base de données (par exemple : la décision d'attribution de la première demande d'APA de la personne, la date du dernier dossier complet de demande d'APA, etc.).

Dans un second temps, nous nous sommes interrogés sur les variables que nous pouvons considérer comme étant quasi-identifiantes et sur celles que nous jugeons sensibles. Nous avons repéré quatre variables quasi-identifiantes : le département, le sexe du bénéficiaire, sa situation familiale (vie en couple ou non) et son âge. Chaque combinaison de ces quatre variables correspond à une clé complète d'identification. Notons cependant qu'un attaquant potentiel doit disposer d'une information supplémentaire pour savoir qu'une personne se trouve dans la base de données : le fait que cette personne est bénéficiaire de l'APA.

La clé d'identification $clé_i$, $i \in \{1, 2, 3, 4\}$ est la combinaison des modalités des variables quasi-identifiantes. Chaque bénéficiaire a ainsi une clé d'identification unique qui est la valeur prise pour chaque variable de la clé.

3. Il s'agit de variables de gestion présentes dans les systèmes d'information des conseils départementaux qui ne sont pas ou que très peu utilisées par les chargés d'études ou les chercheurs, mais peuvent être utiles aux redressements réalisés avant diffusion de la base de données confidentielles au CASD.

Nous avons jugé les variables suivantes comme étant sensibles :

• La variable synthétique, le GIR (Groupe Iso-Ressources), qui mesure la capacité des personnes âgées à effectuer différents actes de la vie quotidienne (Voir la grille AGGIR). Les bénéficiaires présents dans le fichier ont un GIR variant de GIR 1 à GIR 4. Les bénéficiaires de GIR 1 sont confrontés à une perte totale d'autonomie. Parmi les bénéficiaires vivant à domicile, ce sont les moins nombreux ;

• Les axes du GIR qui permettent d'attribuer le GIR. Sur la base d'observations et d'une série de questions, l'équipe médico-sociale évalue chacun des dix axes du GIR qui sont dix critères liés à la perte d'autonomie physique et psychique. La personne responsable de l'évaluation, l'observateur, attribue une note (A, B, C) à chacun des axes selon le degré d'autonomie constaté [4] ;

• La variable précisant la nature de la protection juridique le cas échéant, qui donne un niveau de détail important sur les facultés de la personne et ses capacités à défendre ses propres intérêts (sauvegarde de justice/curatelle simple/curatelle renforcée/tutelle/mandat de protection future) ;

• Le montant des ressources (au sens de l'APA) ;

• Le montant du plan d'aide APA, dans une moindre mesure [5].

Nous avons décidé de flouter ces deux dernières variables jugées sensibles : le montant des ressources et le montant des plan d'aide APA. Concernant l'information sur la protection juridique, nous avons pris la décision de ne pas diffuser le détail des mesures dont la personne pourrait faire l'objet (le révéler pourrait lui être préjudiciable), mais de publier uniquement une indicatrice précisant si le bénéficiaire fait (ou non) l'objet d'une protection juridique. Enfin, le GIR et les axes du GIR sont diffusés tels quels dans la mesure où il s'agit des principales variables d'intérêt de la base de données. De plus, cela n'aurait pas de sens d'attribuer un GIR 1, qui est donné aux personnes les plus dépendantes, à des bénéficiaires qui sont moins dépendants.

Nous conservons dans la base de données floutées 34 variables utiles pour décrire les caractéristiques sociodémographiques des bénéficiaires de l'APA à domicile, leur niveau de dépendance, leurs ressources, les montants de leurs plans d'aide ainsi que des informations sur le contenu de ces plans (montant et nombre d'heures d'aide humaine, etc.).

2 La démarche retenue : méthode perturbatrice et floutage « indirect par des règles »

Nous mettons en oeuvre deux types de traitements. Le premier type de traitement concerne deux variables quantitatives continues (le montant du plan d'aide APA et le montant des res-

4. Les axes : 1. Cohérence : parler ou se comporter raisonnablement ; 2. Orientation : s'orienter dans le temps et l'espace ; 3. WC : assurer son hygiène corporelle de façon autonome ; 4. Habille : se vêtir et se dévêtir en toute autonomie ; 5. Diète : manger des aliments préparés ; 6. Élimination : assumer l'hygiène urinaire et fécale ; 7. Transferts : se lever, s'allonger, s'asseoir ; 8. Déplacement intérieur : se déplacer à l'intérieur du lieu de vie ; 9. Déplacement vers l'extérieur : se déplacer en dehors du lieu de vie ; 10. Communication à distance : utiliser les médias, téléphone, sonnette, alarme.. Les notes ou cotations attribuées à chacun de ces axes sont les suivantes : A : actes accomplis seul spontanément, habituellement, totalement et correctement / B : actes partiellement accomplis / C : actes non réalisés.

5. Les montants du plan d'aide APA nous paraissent être des informations moins sensibles. En effet, les montants que peuvent percevoir les bénéficiaires sont plafonnés. On trouve aisément sur internet les montants de ces plafonds. Le barème dépend des revenus et du niveau de dépendance.

sources) ainsi que l'âge et la date d'ouverture des droits à l'APA. L'âge est une variable importante pour l'analyse et donc pour tout chercheur ou autre utilisateur de la base anonymisée. Aussi, il n'a pas été envisagé de diffuser un âge en tranches (aussi fines soient-elles), mais de conserver uniquement l'année de naissance qui est « floutée » de la même manière que le montant du plan et les ressources. De même, la date d'ouverture des droits à l'APA, variable d'analyse essentielle, a été transformée en une variable quantitative continue et elle est traitée de la même manière que les trois autres [6].

Le premier type de traitement consiste en une méthode qui fait partie de la famille des méthodes dites « perturbatrices », mais on ne procède ni par ajout d'un bruit additif ni par ajout d'un bruit multiplicatif [6]. La piste de la confidentialité différentielle [7] a également été examinée [7]. Nous avons testé l'ajout d'un bruit Laplacien de paramètre $\frac{1}{\epsilon}$, avec plusieurs valeurs de ϵ. Cependant, compte tenu des délais de diffusion relativement courts dont nous disposions, du fait qu'à notre connaissance il existe peu d'expériences sur la mise en oeuvre de cette approche et qu'en outre les premiers tests ne se sont pas avérés concluants, cette piste a été écartée pour l'instant.

Notre raisonnement a été le suivant. On dispose des valeurs précises pour nos variables quantitatives et nous cherchons à produire une approximation voisine de la valeur réelle qui soit en outre cohérente avec les valeurs des autres variables. Dans cette logique, on fixe des intervalles pour chaque variable à flouter « directement » et on applique un processus classique d'imputation par régression. On emploie la méthode d'imputation par résidus simulés. Cette dernière est habituellement utilisée lorsqu'on dispose de données en tranches pour une variable et que l'on souhaite en déduire une valeur quantitative pour celle-ci. Pour cela, on cherche à imputer une valeur spécifique au sein de la tranche conditionnellement à des variables auxiliaires. On transpose cette approche au processus de floutage des variables en question : notre objectif est de simuler une valeur différente de celle observée mais qui ne s'en éloigne pas trop et également de préserver la distribution observée sur les valeurs réelles. Cette méthode nous semble avoir l'avantage d'atteindre un compromis entre protection du fichier et perte d'information. En pratique pour les quatre variables concernées (ressources, montant du plan APA, âge et date d'ouverture des droits à l'APA) on impute donc par résidus simulés, ce qui préserve la distribution observée *via* une régression par intervalles.

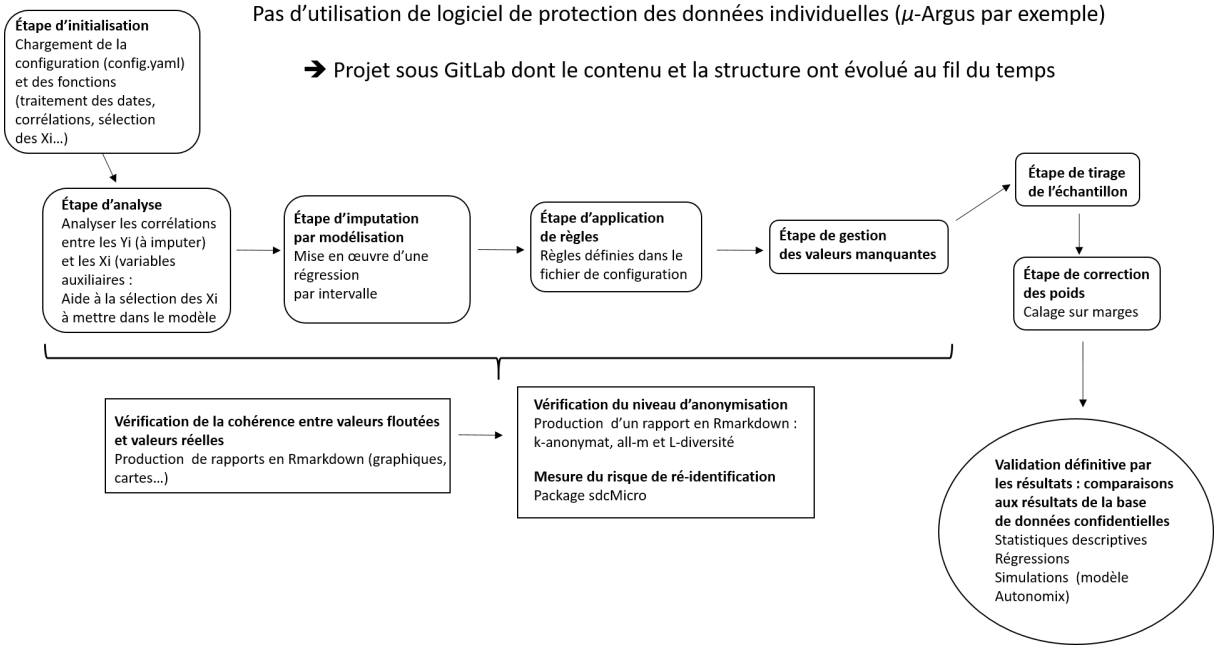
Le second type de traitement concerne une série d'autres variables, liées aux premières, et pour lesquelles on souhaite préserver les corrélations existant avec les variables floutées par la méthode perturbatrice. Ce second type de traitement consiste en un « floutage indirect par des règles ».

La figure 1 explicite les étapes de notre démarche que nous allons détailler par la suite.

6. La date, qui est un objet au format date, a été transformée, avant traitement, en un objet numérique qui est égal à la valeur de cette date moins une valeur d'origine (origine = « 1900-01-01 »).

7. Cela consiste à ajouter du bruit aléatoire aux données pour maintenir la confidentialité de ces données en minimisant les risques d'identification des entités qu'elle contient, si possible en maximisant la pertinence ou la précision des résultats . Il faut donc trouver une équation efficace entre la quantité du bruit à ajouter (le niveau de confidentialité) et la précision des résultats. Une quantité élevée du bruit conduit naturellement à une perte de la qualité des données, et donc à une faible précision des résultats.

FIGURE 1 – Schéma de la démarche générale adoptée



2.1 Méthode perturbatrice

2.1.1 Choix des intervalles en préalable au floutage

L'étape d'imputation présentée sur la figure 1 ci-dessus nécessite, préalablement à l'imputation par une régression, de constituer des intervalles cohérents par rapport à la distribution des valeurs réelles de la variable à flouter. Un intervalle est déterminé pour chaque variable selon une méthode fixée par l'utilisateur :
• soit par les quantiles [8] ;
• soit prédéterminé et fixé par l'utilisateur afin de refléter les spécificités de la distribution de la variable considérée. C'est sur ce point, notamment, que la connaissance métier est indispensable. Ainsi, par exemple, certains intervalles pour le montant de l'APA notifié on été implémentés afin de respecter la saturation observée des plans (c'est-à-dire le fait que les bénéficiaires ont des montants de plan APA massivement concentrés entre 96 % et 100 % du montant plafonné).

Les choix des intervalles appliqués aux variables floutées sont détaillés dans la partie 3, page 11.

2.1.2 La méthode employée pour flouter

Une fois les intervalles constitués, l'étape d'imputation est réalisée selon le processus suivant :

1. Une log-régression par intervalle est d'abord effectuée [9].

En effet, deux des variables numériques à flouter étant les ressources et le montant d'APA, on considère une forme log-normale de la distribution. La répartition des revenus dans la population

8. Le nombre de quantiles est paramétré dans le fichier de configuration. Cela rend les scripts reproductibles pour d'autres opérations et nous a permis de tester les quantiles les plus adaptés (vingtiles, déciles ?) avant d'arrêter notre choix.
9. Fonction Surv du package R survival.

est en effet généralement approchée par une loi log-normale.

Y est la variable dépendante, elle appartient à l'intervalle $[\log(\text{lower}_b)\ ;\ \log(\text{upper}_b)]$

L'équation s'écrit : $[\log(\text{lower}_b)\ ;\ \log(\text{upper}_b)] \sim X\gamma + \varepsilon$

b correspond aux bornes des intervalles : lower_b pour la borne inférieure de chaque intervalle et upper_b pour la borne supérieure de chaque intervalle.

Ce modèle correspond à un modèle linéaire paramétrique sur données censurées (données censurées à droite et à gauche qui correspond aux bornes de chaque intervalle).
La vraisemblance du modèle s'écrit :

$L_i = \phi\left(\frac{X_i\gamma - \text{upper}_b}{\sigma}\right) - \phi\left(\frac{X_i\gamma - \text{lower}_b}{\sigma}\right)$

Le modèle aboutit à une valeur prédite déterministe et à un écart-type des résidus, noté σ. Suite à l'estimation par maximum de vraisemblance, cette première prédiction du montant n'est pas nécessairement dans l'intervalle. La prédiction est notée : $\widehat{Y}$.

2. Pour que le montant prédit respecte les contraintes, c'est-à-dire qu'il soit dans l'intervalle, un résidu est simulé pour garantir l'appartenance de la prédiction du logarithme de la variable à l'intervalle souhaité. En effet, sans contrainte sur la variable estimée, la prédiction du logarithme de $\widehat{Y}$ serait égale à $\widehat{X\gamma}$. Toutefois, comme cette prédiction doit systématiquement être contenue dans un intervalle $[\log(a)\ ;\ \log(b)]$, on ajoute un résidu à cette prédiction. Ce résidu est simulé à partir de la loi normale tronquée issue de celle des résidus (obtenus à l'étape 1) telle que $u \sim N_{[\alpha;\beta]}(0, \sigma)$, avec respectivement $\alpha = X_i\gamma - a$ et $\beta = X_i\gamma - b$. On note ce résidu simulé $residu_s$.

3. La variable floutée se déduit finalement de l'exponentielle de la somme de la prédiction et du résidu simulé, soit $e^{\left(\widehat{Y} + residu_s\right)}$. Ce passage à l'exponentiel est en théorie inexact en raison de la non linéarité de l'exponentiel. Cependant, cette approximation de la prédiction de la variable reste en adéquation avec notre objectif de flouter la vraie valeur de la variable.

Nous remercions ici le Centre de Ressources interrégional sur les Enquêtes Ménages dans les DOM (CRIEM) qui nous a fourni, via le Département des Méthodes statistiques de l'Insee, les scripts R utilisés pour réaliser des imputations de ce type dans le cadre de variables déclarées en clair ou en tranches, par exemple dans le cadre du dispositif Statistiques sur les ressources et conditions de vie (SRCV).

L'âge[10], la date de la dernière évaluation GIR du bénéficiaire de l'APA ainsi que la date d'ouverture des droits à l'APA à domicile ont également été floutées selon cette méthode de régression par intervalles.

2.1.3 Le choix des variables auxiliaires

Les variables auxiliaires introduites dans les modèles ont fait l'objet d'une sélection automatique laquelle est modifiable par l'utilisateur. Des corrélations entre chaque variable à flouter directement et chaque X_i potentielle sont calculées. On utilise le coefficient de corrélation entre deux variables quantitative et le R^2 en régressant une variable quantitative sur une variable

10. L'âge est calculé en années révolues (cf. site de l'Insee pour la définition).

qualitative (régression linéaire simple). Sont sélectionnées automatiquement pour figurer dans le modèle quelques variables X_i qualitatives et quantitatives parmi celles qui sont les plus corrélées à Y_i, sans que la corrélation entre Y_i et un X_i ne soit supérieure à 80 %. En effet, on souhaite qu'une variable très corrélée au montant de l'APA, par exemple, le nombre d'heures d'aide humaine notifié [11], n'intervienne pas en tant que variable auxiliaire dans la régression du montant d'APA, mais soit au contraire flouté indirectement [12].

Enfin, lorsque le GIR ainsi que certains des axes du GIR ressortaient en même temps parmi les variables X_i qualitatives présélectionnées, l'une de ces deux dimensions n'a pas été retenue (ce choix est laissé à l'utilisateur lors de l'examen des X_i présélectionnées) et ce afin d'éviter d'introduire de la colinéarité entre les régresseurs. Lorsque le GIR et certains axes du GIR ressortaient de la sélection automatique pour une même variable Y_i, nous avons systématiquement conservé le GIR, plutôt que les axes (et ce afin de préserver la corrélation existant entre le montant d'APA notifié, par exemple, et le GIR).

2.2 Floutage « indirect par les règles »

Les autres variables liées au montant de l'APA notifié ont donc été floutées indirectement : en utilisant le montant d'APA flouté et, par exemple, le rapport observé entre les valeurs exactes de la variable à flouter indirectement et l'APA réellement observé. On conserve ainsi pour ces variables floutées les mêmes liens que ceux observés entre les valeurs exactes des variables (par exemple : une part similaire du montant relevant du CD dans le montant total, une part similaire de la participation du bénéficiaire dans le montant total, etc.). Par exemple, le montant de la participation financière du bénéficiaire ($NOTPB_F$) sera flouté indirectement de la manière suivante : $NOTPB_F = (NOTPB/APA) \times APA_F$, avec APA_F (montant APA) flouté *via* une log-régression par intervalles. Par ailleurs on évite d'introduire de la complexité, par exemple en développant un système d'équations simultanées (où Y_1 serait estimée en fonction des X_i , Y_2 en fonction des mêmes X_i – ou d'un sous-ensemble de ces X_i - et de Y_1, etc.). Ceci, qui serait sans doute plus rigoureux, conduirait à une formalisation complexe et des estimations rendues délicates par cet ajout de complexité.

Les méthodes et règles appliquées sont listées dans l'annexe 2.

Les étapes de floutage par l'imputation ou par les règles peuvent introduire des valeurs manquantes dans la variable cible alors que celle-ci avait une valeur non manquante dans la base initiale. En effet, ce cas peut se produire lors de l'imputation par régression lorsque certaines observations ont des variables auxiliaires quantitatives à valeurs manquantes ou bien lors de l'application des règles, lorsque des observations de la variable floutée utilisée dans la règle sont manquantes. Ainsi, le nombre de valeurs manquantes générées peut fortement augmenter par un effet « *boule de neige* ». Par exemple, cet effet se produit si la variable Y_1 a des valeurs manquantes alors que Y_1 permet de déduire par une règle Y_2 qui lui-même sert à obtenir Y_3 selon une autre règle.

C'est d'ailleurs ce que nous avons observé en première instance concernant la variable mesu-

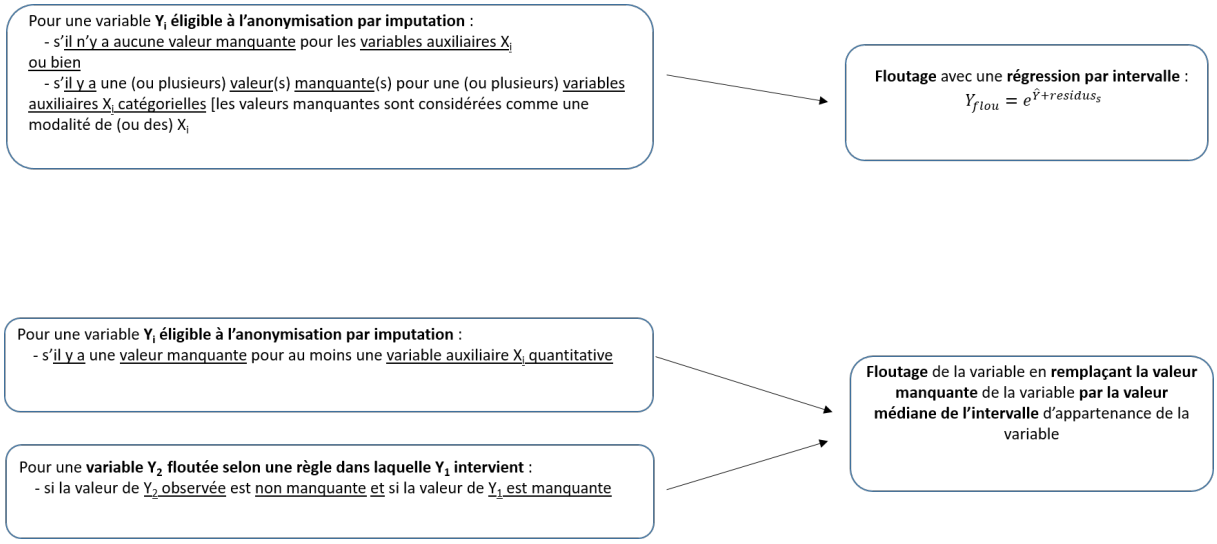
11. Le coefficient de corrélation entre ces deux variables est de 88 %.

12. Afin d'automatiser les scripts, les variables à flouter indirectement apparaissent parmi les variables quantitatives sélectionnées au départ si bien que les corrélations sont automatiquement calculées entre toutes les variables quantitatives, y compris celles que nous avons choisies, *in fine*, de flouter indirectement. L'objectif de ces scripts est en effet qu'ils soient applicables à d'autres bases de données pour lesquelles les calculs de corrélations pourront être nécessaires pour toutes les variables quantitatives sélectionnées.

rant le nombre d'heures d'aide humaine notifiée. 11 % des observations qui avaient des valeurs renseignées pour la variable originelle se sont vues attribuer des valeurs manquantes à l'issue du processus complet de floutage. En effet, notre procédé est le suivant : dans un premier temps on floute le montant mensuel d'aide humaine notifié via la règle suivante : montant mensuel d'aide humaine notifié flouté = (montant mensuel d'aide humaine notifié observé/montant du plan APA notifié observé) x montant du plan APA notifié flouté. Dans un second temps on floute le nombre d'heures d'aide humaine selon le même principe : nombre d'heures d'aide humaine notifié flouté = (nombre d'heures d'aide humaine notifié observé/montant mensuel d'aide humaine notifié observé) x montant mensuel d'aide humaine notifié flouté. Chacune des deux étapes a généré des valeurs manquantes pour la variable nombre d'heures d'aide humaine notifiée floutée. Ce 11 % d'observations devenues manquantes représente un nombre élevé d'observations (environ 70 000 observations). Ainsi, nous avons décidé d'imputer ces valeurs devenues manquantes suite au processus de floutage. Nous leur imputons la valeur médiane de la tranche dans laquelle les valeurs initiales observées se situent : par exemple le nombre d'heure d'aide humaine médian du 1er décile pour les observations devenues manquantes qui se situaient en dessous de la valeur-seuil du 1er décile. Ce traitement qui mobilise une méthode de type « micro-agrégation » permet un gain d'information non négligeable. En outre, l'ajout de ces valeurs médianes (au lieu de conserver les valeurs manquantes finales) ne déforme ni la distribution des valeurs floutées, ni l'adéquation des distributions floutées aux distributions réelles observées (cf. annexe 3, page 30).

La figure 2 synthétise les différents cas possibles et la décision prise pour le floutage de la variable cible.

FIGURE 2 – Schéma des cas observés et des décisions prises pour le floutage de la variable cible



3 Vérification de la cohérence

Une fois les opérations de floutage réalisées, nous procédons à une phase, importante, de vérification de la cohérence entre valeurs floutées et valeurs observées. Nous nous sommes appuyés sur la production d'un rapport en R Markdown. Il n'est pas produit uniquement à la fin du processus, il a au contraire permis de réitérer pour affiner nos configurations comme le choix des intervalles, la suppression de valeurs extrêmes ou encore la gestion des valeurs manquantes. Ce rapport présente notamment une comparaison des densités des valeurs avant et après floutage

pour chaque variable quantitative floutée.

Tout d'abord, pour les variables floutées directement (pour rappel : montants de ressources et de plan APA notifié, date d'évaluation de l'APA et âge des bénéficiaires), les choix de partitions en intervalles ont été effectués de la manière suivante. Grâce à l'examen des rapports de vérification de cohérence produits, nous avons été amenés à revoir certains paramètres ou certains choix.

Concernant le montant mensuel du plan d'APA notifié : une partition en quantiles a tout d'abord été testée. Mais une telle partition ne permettait pas de refléter les spécificités de la distribution de cette variable qui fait apparaître des points d'accumulation à proximité des montants des plafonds. Ces points d'accumulation représentent les plans saturés (cf. figure 3, page 13). Aussi les bornes des tranches ont été définies manuellement aux vues des données, afin de préserver après floutage ces points d'accumulation, c'est-à-dire que l'on considère comme tranches à part entière les intervalles suivants $[96\ \% * plafond\ ;\ plafond]$, par exemple [637 euros ; 664 euros] pour le 1er plafond [13]. La partition retenue comporte une quinzaine d'intervalles.

Pour le montant des ressources, une partition prédéterminée en déciles a été testée, mais elle ne convenait pas non plus : l'étendue du dernier décile apparaissait bien trop large. Ce partitionnement conduisait logiquement à imputer des valeurs trop élevées pour les montants floutés du dernier décile [14]. Ceci conduit à des montants moyens des ressources trop élevés pour la variable floutée, soit pour les hommes comme pour les femmes un écart non négligeable de + 100 à + 200 euros par rapport à des montants moyens de l'ordre de 1 400 euros pour les hommes et de 1 300 euros pour les femmes dans la base de données confidentielles. Nous avons en fait une cinquantaine d'individus qui ont des montants de ressources élevés que l'on considère comme aberrants. On suppose que ces montants erronés [15] correspondent à l'inclusion dans le montant des ressources d'éléments de patrimoine (par exemple, inclusion de la valorisation du patrimoine dormant). Ces observations extrêmes sont considérées comme des outliers et comme potentiellement à risque de ré-identification (bénéficiaires possédant un patrimoine important). Dans un premier temps nous avons envisagé de retirer ces observations de la base anonymisée. Finalement, nous avons considéré un seuil de diffusion des montants à moins de X euros et au dessus de ce seuil les montants de ressources des bénéficiaires ont été mis à valeurs manquantes (ce qui ne modifie pas la proportion de valeurs manquantes pour cette variable) et nous n'avons pas imputé de montants pour cette cinquantaine de bénéficiaires [16].

Ce seuil de diffusion de la base floutée à X euros maximum étant défini, le haut de la distribution des ressources a été découpé assez finement. De même, nous avons découpé finement les ressources dans le bas de la distribution, puisque la première comparaison des distributions de ressources, après et avant floutage, montrait des écarts trop importants (une valeur seuil du 1er décile de 40 % supérieure dans la base anonymisée pour les bénéficiaires célibataires, ceux ayant les plus faibles ressources). En effet, la méthode conduisait à imputer moins de petites valeurs proches de zéro et plus de valeurs entre 250 euros et 500 euros comparativement à ce qui est observé. Aussi, il a été décidé l'application d'un découpage plus fin, non seulement pour les raisons évoquées plus haut, mais également pour être en phase avec des tranches utilisées dans le modèle de microsimulation Autonomix. Au final, le partitionnement comporte une vingtaine

13. Le montant maximal du plan d'aide APA à domicile pour 2017 est de 664 euros pour les bénéficiaires de GIR 4 (les moins dépendants).

14. Nous n'avons pas imposé au modèle une valeur maximale de ressources à imputer.

15. Il s'agit de montants trop élevés pour la population concernée (à « dires d'experts »).

16. Le but est de ne pas diffuser dans la base anonymisée ces valeurs atypiques (des montants floutés de ressources ne sont donc pas imputés à ces individus), tout en conservant les individus concernés dont les valeurs d'autres variables d'intérêt sont renseignées.

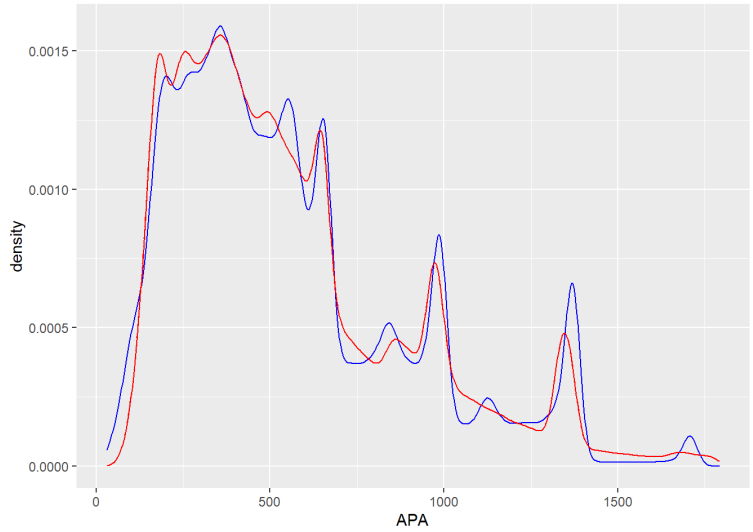
d'intervalles.

Pour la date d'ouverture des droits à l'APA un premier partitionnement en déciles a également été testé. Étant donné la forme de la distribution, on a finalement choisi des vingtiles dans le bas de la distribution, voire un découpage encore plus fin, et ensuite des déciles. Cela donne au final un découpage entre dix et quinze intervalles.

Concernant l'âge, nous avons dès le départ choisi des bornes, non pas selon un partitionnement statistique en quantiles, mais selon des tranches prédéterminées et ce compte tenu de la répartition par âge de la population des bénéficiaires de l'APA à domicile. Au départ, des tranches de deux ans en milieu de distribution ont été testées, mais ce découpage s'est avéré trop fin pour obtenir un âge suffisamment flouté. Des tranches un peu plus larges dans le milieu de la distribution ont finalement été appliquées. Au final, environ une douzaine d'intervalles ont été retenus.

Les graphiques de densité entre les variables initiales et les variables floutées sont présentés en annexe 3. Nous reprenons ici l'un d'eux pour montrer que nous simulons globalement assez bien la distribution du montant de l'APA notifié, une des principales variables d'intérêt de la base de données, en particulier les points d'accumulations dûs aux plans saturés (i.e. les pics élevés de la figure 3).

FIGURE 3 – Montant mensuel du plan d'aide APA notifié



Il ressort de l'ensemble de ces comparaisons graphiques que pour la quasi-totalité des variables la distribution des valeurs floutées est très proche de celle des valeurs observées. Seules les distributions de deux variables floutées s'ajustent un peu moins bien aux valeurs réelles. Pour le montant total de l'APA versé par le CD en décembre 2017 (figure 13 de l'annexe 2), on observe que le pic dans le bas de la distribution est légèrement moins élevé et un peu décalé vers la droite après floutage. Concernant la date de la dernière évaluation du GIR du bénéficiaire (figure 17 de l'annexe 2), on observe un pic un peu moins important qu'observé dans le haut de la distribution : on impute un peu moins de dates d'évaluation récentes. Cependant ces écarts sont minimes et ne nécessitent pas de retraitements particuliers au vu des résultats statistiques produits avec les deux bases (résultats qui seront présentés dans la partie 7, page 18).

4 Vérification du niveau de confidentialité

4.1 Traitement préalable sur les départements

Avant de contrôler le niveau de confidentialité, nous avons effectué un traitement de la variable département. Les départements comportant moins de 2 000 bénéficiaires sont conservés dans l'échantillon mais on leur affecte un code non-signifiant choisi aléatoirement. Seuls les départements 2A et 2B (Corse) sont regroupés dans le code 2A. Les départements suivants se voient donc affecter aléatoirement un code ayant une valeur comprise entre 100 et 107 : les Hautes-Alpes, la Creuse, le Lot, la Lozère, la Mayenne, les Deux-Sèvres et le Territoire-de-Belfort. Le premier département concerné par l'attribution d'un numéro aléatoire – soit les Hautes-Alpes - n'aura ainsi pas forcément la valeur 101.

4.2 Niveaux de confidentialité

Pour rappel, la clé d'identification $clé_i$, $i \in \{1, 2, 3, 4\}$ est la combinaison des modalités des quatre quasi-identifiants repérés (département, sexe, situation familiale et âge). Concernant l'âge nous prenons en compte pour la clé d'identification non pas l'âge en tant que tel puisqu'il est flouté, mais les mêmes classes d'âge que celles utilisées pour flouter l'âge (soit une douzaine d'intervalles environ).

La validation de notre démarche nécessite de contrôler le niveau de confidentialité garanti par notre processus. Pour cela, plusieurs comptages sont réalisés afin de vérifier le k-anonymat ainsi que le all-m anonymat. Le k-anonymat porte sur la clé complète d'identification et signifie que le seuil de diffusion est fixé à au moins k individus pour chaque croisement de la clé complète. On considère qu'au-dessus de ce seuil il n'est pas possible de « ré-identifier indirectement ces individus en mettant en œuvre des moyens raisonnables »(cf. annexe 1 sur les données à caractère personnel). Le all-m anonymat suit le même principe, mais sur des sous-combinaisons de quelques variables de la clé complète (trois variables, etc.). La l-diversité sera également examinée plus loin (cf. partie 4.5, page 16).

Un tableau de comptage par variable quasi-identifiante est produit et il permet de mettre en exergue les variables ayant des déséquilibres importants et qui seront potentiellement sources de problème de ré-identification [17]. Un second tableau met en évidence les problèmes de k-anonymat et de all-m anonymat. Ce tableau fournit, pour chaque niveau de croisement, le nombre de clés identifiantes pour lesquelles le nombre d'occurrence est inférieur ou égal à k ou à m.

La contrainte du k-anonymat est vérifiée si chaque clé complète d'identification est possédée par au moins k individus de la population. Nous fixons k = 3. En pratique des seuils de k = 10 ou k = 5 sont parfois utilisés, sans que des raisons objectives soient toujours avancées. Nous avons choisi d'être un peu moins restrictifs en choisissant un seuil de k = 3 ; nous minimisons donc un peu moins le risque de ré-identification. Le choix de ce seuil reste en tout état de cause quelque peu arbitraire et relève de l'expertise métier. La littérature, à notre connaissance, ne nous donne pas de règle pour fixer un k « optimal » [18]. La fait d'être moins restrictif ne nous semble pas problématique puisqu'au final ce n'est pas la base de données floutées complète qui sera diffusée mais un sous-échantillon ce qui contribue à réduire encore plus les risques (ré-identification, etc.).

17. Comme précédemment, sous la forme d'un rapport en R Markdown.

18. « L'estimation de la probabilité que ces informations donnent lieu à une divulgation par les personnes qui en ont eu connaissance est plus subjective (« à dires d'experts ») »[9].

Le « all-m », dont le principe dérive du k-anonymat consiste à effectuer les comptages sur des sous-ensembles de variables quasi-identifiantes. Nous fixons m = 5. Pour le all-m, la clé d'identification étant incomplète, le risque de ré-identification est *a fortiori* moins élevé (en l'absence d'une ou plusieurs variables informatives), nous acceptons donc un seuil plus élevé, donc plus restrictif.

4.3 Le k-anonymat

Après avoir flouté les départements à effectifs faibles, on recensait environ 5 000 croisements de la clé complète d'identification qui comportent de 1 à près de 3 000 individus. En moyenne, ces croisements comportent un peu plus de 100 individus ; la médiane se situant à 60 individus environ.

Parmi ces 5 000 croisements, environ 250 seulement ne respectent pas le 3-anonymat (soit seulement 5 % des croisements). Cependant, même si ces 250 croisements ne respectent pas le 3-anonymat, ceux qui comportent un département flouté et/ou une modalité = 9 (non connu) pour la situation de vie en couple ne s'avèrent de fait pas problématiques. Ces cas concernent au total 220 croisements environ. Pour les croisements ayant un département flouté, un attaquant potentiel ne saura pas dans quel département se situent les 3 individus en question. Il aura 1 chance sur 7 de déduire le « bon département » (puisqu'on a 7 catégories de départements floutés). De plus il ne connaîtra pas la situation familiale des bénéficiaires puisque tous les individus des croisements à moins de 3 individus dont le département est flouté ont une modalité « non connu »pour la situation familiale. Finalement les attaquants ne disposeront que de peu d'information : le sexe de la personne, son âge mais qui est flouté et 1 chance sur 7 de déduire le « bon » département (en outre ils doivent savoir que la personne est bénéficiaire de l'APA). Dans le cas où la situation familiale du bénéficiaire est « non connue », l'attaquant potentiel pourra inférer que les 3 individus en question ont 36 chances sur 100 d'être en couple (aux vues de la proportion de bénéficiaires en couple dans la base anonymisée ré-échantillonnée qui a été diffusée ; voir ci-dessous la partie concernant l'échantillonnage partie 5, page 17). Cette information ne lui permettra pas pour autant de ré-identifier de manière quasi-certaine les individus en question (même en connaissant leur département, leur sexe et leur âge qui est en outre flouté).

Il ne reste ainsi plus qu'une trentaine de croisements de la clé complète – dont tous les quasi-identifiants sont informatifs - pour lesquels on a 3 individus ou moins. Dans ces cas-là, l'attaquant potentiel dispose donc du département, du sexe, de la situation familiale et de l'âge qui est flouté. On traite tout de même ces cas même s'ils sont *a priori* moins à risque de ré-identification, l'âge vrai n'étant pas connu de l'attaquant potentiel, lequel ne connaît pas la méthode de floutage et ne sait pas forcément si les individus en question touchent l'APA. Ainsi, par mesure de protection supplémentaire, nous appliquons la méthode de suppression locale. Pour des bénéficiaires possédant une clé d'identification à risque, cette méthode consiste à supprimer les valeurs d'une(ou des) variable(s) quasi-identifiante(s) en les remplaçant par une valeur manquante [6]. Dans l'absolu, nous avons le choix entre trois variables sur lesquelles on pourrait appliquer cette méthode (département, sexe et situation familiale) ; nous ne prenons pas en compte l'âge parmi ces possibilités de choix puisque l'âge est déjà flouté. Il ne nous semble pas indispensable d'appliquer cette méthode à chacune des trois variables : ce serait une perte d'information trop importante, même si cela concerne peu d'individus. La base ne comporte aucune valeur manquante pour le département et le sexe, il ne serait pas judicieux d'en introduire [19]. Ainsi, nous choisissons d'agir uniquement sur la variable situation familiale, laquelle comporte déjà des observations à valeurs manquantes. On recode donc la situation de vie en couple à « non connu » pour les individus des

19. On tient à éviter que les chercheurs travaillant sur la base floutée imputent des valeurs à ces variables à valeurs manquantes, alors que dans la base de données confidentielles qu'ils seront amenés à utiliser par la suite ces variables sont complètement renseignées.

croisements de la clé complète comportant 3 bénéficiaires ou moins. Nous avons vérifié que ce traitement ne déforme pas la répartition des bénéficiaires selon leur situation de vie en couple [20] ce qui était prévisible étant donné le peu de cas retraités (une trentaine). Ainsi, plus aucun croisement informatif de la clé complète ne comporte 3 observations ou moins. Au regard des traitements réalisés, le fichier est donc 3-anonyme.

4.4 Le all-m

Afin de mieux contrôler l'anonymisation, le département des Méthodes statistiques de l'Insee, en particulier Maxime Beauté que nous remercions ici, nous a conseillé de s'attacher également à respecter le « all-m anonymat ». Cela consiste à vérifier que nous avons au minimum m individus pour les croisements comportant au moins une variable quasi-identifiante de moins que le nombre de variables composant la clé complète. Pour le all-m, étant donné que le nombre de quasi-identifiant est inférieur à celui de la clé complète, on définit un seuil un peu plus élevé que pour le k-anonymat, avec m = 5. Seuls les croisements avec trois variables - parmi lesquelles la tranche d'âge ne figure pas puisque l'âge est flouté - sont potentiellement sources de risques de ré-identification. Il s'agit au final du seul et unique croisement : département x sexe x situation familiale du bénéficiaire. En effet, si on ne considère que deux variables, comme le département et le sexe par exemple (ou bien la situation familiale), on estime qu'un utilisateur malveillant n'a pas assez d'information pour ré-identifier les individus en question. On suppose qu'il n'y a pas de risque de ré-identification ou du moins qu'il est très faible et on ne retraite pas ces cas-là.

Pour cette combinaison de la clé incomplète (département x sexe x situation familiale), on dénombre à peine 10 croisements comportant 5 individus ou moins (sur un peu plus de 400 croisements de cette clé recensés). Or pour ces croisements, la situation de vie en couple des individus concernés est toujours « non connue ». Ainsi l'attaquant potentiel ne dispose que du département et du sexe de la personne pour la ré-identifier. On considère que ces deux seules informations, ajoutées au fait que l'attaquant doit savoir que l'individu perçoit l'APA, sont insuffisantes pour qu'une ré-identification soit possible de manière certaine ou quasi-certaine. Aucun traitement n'est effectué pour ces cas que l'on ne considère pas comme étant à risque de ré-identification.

4.5 La l-diversité : prise en compte du GIR

On a défini une l-diversité = 2, c'est-à-dire que les croisements de la clé complète doivent comporter des individus ayant au moins deux GIR différents. Les croisements pour lesquels tous les individus ont le même GIR nous semblent problématiques. En effet, dans ce cas, les individus ne sont pas protégés des risques de divulgation d'attribut ou de révélation inférentielle. Un attaquant malveillant pourrait déduire, par exemple, que tous les individus du Cantal, bénéficiaires de l'APA, de sexe masculin, d'âge flouté à 80 ans (ayant potentiellement un âge véritable aux alentours de ces 80 ans), vivant seuls en 2017 sont tous de GIR 1, si seul ce GIR est représenté dans ce croisement de variables.

On décide donc de mettre en œuvre une protection supplémentaire en imposant non seulement le 3-anonymat mais également que les croisements de la clé complète d'identification soient associés à au moins deux GIR. Ces individus ayant la même clé complète d'identification et possédant le même GIR sont exclus de la base anonymisée diffusée. On exclut ainsi seulement 140 individus environ, soit très peu au regard de la taille de la base de départ (environ 618 000

20. Avant la mise en oeuvre de la suppression locale, la base comporte environ 23 000 valeurs manquantes pour la situation de vie en couple (3,7 % des observations).

bénéficiaires). On applique là une technique de sous-échantillonnage déterministe [21] où seules les observations ne présentant pas de risque de révélation d'attribut ou de révélation inférentielle tels que définis ici sont conservées.

Même sans ce traitement de la l-diversité imposée au GIR, dans la mesure où on échantillonne ensuite au 1/2, on disposera au final de croisements pour lesquelles il n'y aura qu'un GIR par croisement de variables quasi-identifiantes. L'utilisateur de la base floutée n'aura cependant aucun moyen de savoir si c'était déjà le cas dans la base de données confidentielles (auquel cas la base ne serait pas 2-diverse) ou bien si cela tient à l'échantillonnage. Il ne pourra donc qu'inférer éventuellement un résultat qui ne sera pas « quasi-certain ». Par exemple, il sera amené à supposer que tous les individus masculins de l'Ain, vivant en couple, ayant environ 75 ans (âge flouté) sont de GIR 3 parce que dans le sous-échantillon diffusé en *open data* les 6 individus (par exemple) de cette combinaison sont tous de GIR 3. Pourtant, avant l'échantionnage, il pourrait y avoir en fait 6 individus de GIR 3 (50 %) et 6 individus de GIR 4 (50 %). On estime donc que le risque d'inférence quasi-certaine (au sens de la CNIL) est minime.

5 Échantillonnage et calage avant diffusion

La base de données ainsi anonymisée, comportant des effectifs de taille importante, il est possible de n'en diffuser qu'une partie pour réduire plus encore les risques. Nous avons donc tiré un sous-échantillon aléatoire stratifié au 1/2 [22] afin de réduire encore plus le risque de ré-identification et *a fortiori* le risque de divulgation d'attribut ou de divulgation inférentielle [23].

On effectue ensuite un calage sur marges pour conserver les distributions observées avec les données confidentielles [24].

6 Risque de ré-identification

Afin d'apprécier le risque de ré-identification, on calcule le risque global de ré-identification [25] du fichier. Le risque est calculé sur la clé complète et sur le fichier complet (avant échantillonnage). On a exclu les cas où la l-diversité n'était pas respectée (140 individus exclus environ).

Concrètement, le risque est l'espérance de l'inverse des fréquences de la clé d'authentification complète dans l'échantillon. Ainsi, si 25 individus partagent la même clé d'authentification k ($f_k = 25$), chacun de ces individus aura un risque individuel de ré-identification de 1/25, soit 0,04. Le risque global du fichier est la somme pondérée des risques de ré-identification individuels, soit la somme des f_k/N, avec N le nombre d'observations du fichier.

La mesure du risque repose sur la probabilité de ré-identifications exactes d'individus dans notre échantillon (*via* des données exogènes comportant les mêmes quasi-identifiants). Plus un

21. Il existe d'autres techniques de sous-échantillonnage. Cette technique de sous-échantillonnage déterministe a été testée à l'Insee pour l'enquête Vols, violences et sécurité [4] et [6].

22. Pourquoi un échantillon au 1/2 ? Nous avons en fait testé plusieurs tailles de sous-échantillons : au 1/5e, au 1/3. Nous avons finalement opté pour le 1/2 ce qui permet d'avoir suffisamment d'observations et garantit une meilleure robustesse des résultats obtenus, en particulier pour la modélisation.

23. A la condition que la méthode de tirage ne soit pas divulguée à l'utilisateur potentiel.

24. Les marges sont les mêmes que celles utilisées pour pondérer la base de données confidentielles, à savoir les bénéficiaires payés de l'enquête Annuelle Aide Sociale produite par la Drees.

25. Se référer à Calcul de scores avec sdcMICRO.

individu est unique selon les variables quasi-identifiantes, plus il est à risque d'être ré-identifié de manière correcte ou certaine (c'est-à-dire que c'est le bon individu qui pourrait être ré-identifié). Le nombre probable de ré-identifications exactes de bénéficiaires est estimé à environ 5 000 individus, soit un risque global du fichier de 0,8 %. Ce nombre ne signifie pas pour autant que 5 000 individus déterminés de l'échantillon peuvent être ré-identifiés de façon certaine par un attaquant connaissant toutes les variables quasi-identifiantes. Il s'agit d'une espérance et non d'un taux de ré-identification, puisque toute tentative de ré-identification doit être pour partie aléatoire, du fait du k-anonymat : parmi les au moins k personnes qui partagent les mêmes quasi-identifiants, l'attaquant n'a d'autre solution que de tirer au sort pour essayer d'identifier les personnes.

Aux vues de ces résultats, nous considérons que le risque de ré-identification est suffisamment faible et ce d'autant plus que seul un sous-échantillon est diffusé.

7 Validation par les résultats

Avant de décider de diffuser cette base de données floutées, il nous a semblé indispensable de comparer les résultats qu'elle donne et de vérifier ainsi qu'elle délivre les mêmes messages que ceux résultant de la base de données confidentielles. Pour cela, nous procédons à la comparaison de statistiques descriptives, puis d'un modèle de régression et enfin de simulations *via* le modèle de microsimulation Autonomix.

7.1 Statistiques descriptives

L'ensemble des résultats des statistiques descriptives est reporté en annexe 4. Les écarts présentés portent toujours sur la comparaison de la base floutée par rapport à la base de données confidentielles (ou base au CASD).

Les distributions de l'âge selon le sexe des bénéficiaires sont quasiment identiques à 1 an révolu près dans le haut et le bas de la distribution, soit le P10 et le P90 pour les hommes comme pour les femmes (figure 18). La distribution des ressources floutées des bénéficiaires selon le sexe et la situation familiale est tout à fait comparable à celle observée avec la base au CASD (figure 19). Les écarts sont négligeables : au plus 1 % quel que soit l'indicateur de dispersion. La distribution des plans d'aides notifiés selon le GIR issue de la base floutée est tout à fait semblable à celle observée sur la base au CASD : la médiane du montant du plan s'élève à 1 309 euros pour les bénéficiaires du GIR 1 dans la base floutée, contre 1 316 euros dans la base exhaustive (figure 20). Quel que soit l'indicateur considéré (Q1, médiane, Q3, moyenne), les écarts de montants entre les deux bases ne dépassent pas 3 %, ce qui ne représente qu'au plus 7 euros mensuels d'écart. La part représentée par la participation financière des bénéficiaires au plan d'APA notifié, selon le GIR, est similaire dans les deux bases. Il en est de même pour la proportion de plans saturés [26], selon le GIR.

La distribution des montants du plan d'aide notifié selon les ressources des bénéficiaires dans la base floutée est similaire à celle observée dans la base au CASD (figure 21). On note au plus, une valeur-seuil du premier décile légèrement plus élevée dans la base floutée pour les bénéficiaires les plus aisés (tranche de ressources supérieures à 2 500 euros mensuels), mais l'écart est minime (+ 4 %, ce qui ne représente que 9 euros par mois). La distribution du montant versé par le conseil départemental (CD) selon le sexe et la situation familiale dans la base floutée est analogue à celle calculée avec la base au CASD (figure 22). Dans la base floutée, les valeurs seuils

26. Un plan est dit « saturé »(au seuil de 96 %) lorsque son montant est égal ou presque au plafond maximal d'aide, soit s'il est supérieur ou égal à 96 % du plafond.

sont légèrement plus élevées dans le bas de la distribution, mais les différences ne portent que sur 3 à 7 euros par mois environ. Un constat similaire ressort de la comparaison des distributions des montants versés par les CD, selon les ressources des bénéficiaires (figure 23). Seules les tranches basses et hautes de ressources font apparaître de légers écarts dans le bas de la distribution (premier décile). Mais ces écarts sont minimes, ils ne représentent qu'au plus une dizaine d'euros. Ces écarts sont considérés comme négligeables, puisque dans l'ensemble le message est le même : la part versée par le CD baisse à mesure que les ressources augmentent, quel que soit l'indicateur de dispersion. De même, les distributions des montants d'APA notifiés à la charge des bénéficiaires selon leurs caractéristiques (sexe et situation de vie en couple) sont analogues dans les deux bases de données (figure 24). Globalement, les distributions des montants d'APA notifiés à la charge des bénéficiaires selon les ressources calculées sur la base floutée sont proches de celles observées sur la base au CASD (figure 25). On note des écarts un peu plus importants que pour les autres variables dans le haut de la distribution pour les bénéficiaires aisés (les montants floutés étant inférieurs à ceux observés). Nous jugeons cependant ces différences acceptables : les montants en jeu sont relativement similaires et permettent de délivrer un message équivalent, à savoir que les montants notifiés restant à la charge des bénéficiaires augmentent à mesure que leurs ressources s'accroissent.

Même si on compte des effectifs de bénéficiaires ayant un droit à l'APA ouvert en 2017 un peu moins élevés dans la base floutée que dans la base au CASD, les répartitions des bénéficiaires selon la date d'ouverture de leurs droits à l'APA restent équivalentes (figure 26). Comme on a pu l'observer graphiquement, les bénéficiaires ayant des dates récentes de dernière évaluation du GIR (2016) sont un peu moins nombreux dans la base floutée que dans la base au CASD (figure 27). Pour autant les répartitions des bénéficiaires selon leur dernière date d'évaluation du GIR sont équivalentes d'une base à l'autre.

En conclusion, les statistiques descriptives obtenues sur la base floutée sont tout à fait comparables à celles obtenues sur la base de données confidentielles. Si quelques écarts ressortent parfois, ils restent généralement inférieurs à 10 % ou du moins portent sur des différences relativement faibles en montant (de l'ordre de 10 euros sur les petits montants, voire 100 euros pour les plus aisés).

7.2 Modèles de régression

La base de données floutée conduit donc à des résultats comparables à ceux obtenus avec la base de données confidentielle en termes de statistiques descriptives. Mais encore faut-il qu'un modèle de régression dont la variable à expliquer et certaines variables auxiliaires sont floutées conduisent aux mêmes messages qu'en utilisant les variables de la base de données confidentielles. Autrement dit, il ne faudrait pas que les résultats des régressions menées ne soient que le reflet du « bruit » introduit en floutant nos variables ou que les perturbations introduites déforment les corrélations entre variables et conduisent à des messages erronés.

Afin de vérifier ceci, nous avons répliqué l'estimation d'un modèle de régression censurée (modèle Tobit) dont les résultats ont été publiés en 2020 [10]. Ces régressions portent sur le montant d'aide humaine notifié selon les caractéristiques des bénéficiaires [27].

27. Les modèles de régressions ont été ré-estimés sur la base de données confidentielles, puisqu'une des variables auxiliaires utilisées dans les travaux publiés ne figure pas dans la base de données floutée (densité de population de la commune d'habitation) [9].

7.2.1 Modèle avec l'ensemble des bénéficiaires (tous GIR confondus)

Dans un premier temps le modèle est estimé sur l'ensemble des bénéficiaires quel que soit leur GIR. La comparaison des résultats des modèles montre que les deux bases de données permettent de délivrer les mêmes messages (figures 4 et 5) : un montant moyen d'aide humaine notifié qui augmente avec l'âge des bénéficiaires, à autres caractéristiques identiques, des montants moyens plus élevés pour les bénéficiaires sans conjoint, des montants qui augmentent quand les ressources sont plus élevées, avec une tendance qui s'inverse pour les bénéficiaires les plus aisés, et enfin des montants qui sont plus élevés pour les bénéficiaires les plus dépendants (bénéficiaires de GIR 1).

Plus précisément, par exemple, le montant d'aide humaine notifié dans le plan augmente avec l'âge des bénéficiaires, toutes choses égales par ailleurs. Par rapport aux bénéficiaires les plus jeunes, il est supérieur de 9 euros dans la base floutée pour les bénéficiaires âgés de 75 à 80 ans (12 euros avec la base CASD) et de 61 euros pour les bénéficiaires de 90 ans ou plus (63 euros avec la base CASD). Les personnes sans conjoint se voient notifier un plan d'un montant plus important que celui des bénéficiaires vivant en couple. Pour les hommes, comme pour les femmes, vivant seuls, le montant est supérieur de près de 130 euros (quelle que soit la base de données) à celui des hommes vivant en couple, à autres caractéristiques identiques. Le montant notifié diminue globalement lorsque le niveau de ressources augmente, même si la tendance s'inverse pour les bénéficiaires les plus aisés. À caractéristiques identiques, les équipes médico-sociales proposent un montant d'aide humaine plus faible de 85 euros avec la base floutée - 89 euros avec la base de données confidentielles - à un bénéficiaire dont les ressources sont comprises entre 2 000 et 2 500 euros qu'à un bénéficiaire dont les ressources sont inférieures à 740 euros. C'est pour le GIR, à autres caractéristiques identiques, que les effets sur le montant d'aide humaine notifié sont d'une ampleur légèrement moindre quand on compare les deux bases, c'est-à-dire un effet légèrement moins important du GIR 1 par rapport au GIR 4 dans la base floutée : + 813 euros contre + 828 euros pour la base CASD, mais cet écart de 15 euros ne représente qu'une différence minime (-1,8 %).

FIGURE 4 – Montant d'aide humaine notifiée selon les caractéristiques des bénéficiaires, base floutée

Variables	Montant moyen notifié d'aide humaine (en €)	Effet sur le montant, à caractéristiques identiques
Âge : [60 ; 75[	450	Ref.
Âge : [75 ; 80[	460	8.9***
Âge : [80 ; 85[	480	24.5***
Âge : [85 ; 90[	510	40.4***
Âge : 90 ou +	570	60.6***
Situation : Homme seul	510	128.2***
Situation : Homme en couple	460	Ref.
Situation : Femme seule	540	127***
Situation : Femme en couple	450	53.4***
Ressources (€ / mois) : [0 ; 739,8[	550	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	530	-10.4***
Ressources (€ / mois) : [1000 ; 1250[	510	-30.2***
Ressources (€ / mois) : [1250 ; 1500[	490	-47.7***
Ressources (€ / mois) : [1500 ; 2000[	470	-67.9***
Ressources (€ / mois) : [2000 ; 2500[	470	-84.5***
Ressources (€ / mois) : 2500 ou +	530	-80.5***
GIR 1	1 140	812.9***
GIR 2	870	538.8***
GIR 3	610	269.7***
GIR 4	340	Ref.

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.
Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire dont l'âge est compris entre 75 et 80 ans s'élève à 460 euros. À caractéristiques identiques, ce montant est supérieur de 8,9 euros à celui qui serait notifié à un bénéficiaire âgé de moins de 75 ans.
Champ > Bénéficiaires de l'APA à domicile recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base floutée, Drees.

FIGURE 5 – Montant d'aide humaine notifiée selon les caractéristiques des bénéficiaires, base au CASD

Variables	Montant moyen notifié d'aide humaine (en €)	Effet sur le montant, à caractéristiques identiques
Âge : [60 ; 75[	450	Ref.
Âge : [75 ; 80[	460	11.5***
Âge : [80 ; 85[	480	24.3***
Âge : [85 ; 90[	510	43.2***
Âge : 90 ou +	580	62.5***
Situation : Homme seul	510	131.2***
Situation : Homme en couple	450	Ref.
Situation : Femme seule	540	127.4***
Situation : Femme en couple	450	54.6***
Ressources (€ / mois) : [0 ; 739,8[	560	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	530	-12.9***
Ressources (€ / mois) : [1000 ; 1250[	510	-33.8***
Ressources (€ / mois) : [1250 ; 1500[	490	-52.4***
Ressources (€ / mois) : [1500 ; 2000[	470	-71.2***
Ressources (€ / mois) : [2000 ; 2500[	460	-88.5***
Ressources (€ / mois) : 2500 ou +	530	-86***
GIR 1	1 150	828.3***
GIR 2	870	543.8***
GIR 3	600	267***
GIR 4	340	Ref.

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.
Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire dont l'âge est compris entre 75 et 80 ans s'élève à 460 euros. À caractéristiques identiques, ce montant est supérieur de 11,5 euros à celui qui serait notifié à un bénéficiaire âgé de moins de 75 ans.
Champ > Bénéficiaires de l'APA à domicile recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base CASD, Drees.

7.2.2 Modèles séparés pour chaque niveau de dépendance des bénéficiaires

De la même manière on a examiné les comparaisons des résultats des mêmes modèles mais estimés cette fois indépendemment pour les bénéficiaires de chaque GIR. Les résultats figurent en annexe 5.

Globalement des effets du même type ressortent pour les bénéficiaires de GIR 1 (les plus dépendants) dans la base floutée et dans la base CASD (annexe 5, figure 28 et figure 29). Seul le coefficient d'une tranche de ressources ne ressort plus comme significativement différent de la situation prise en référence dans la base anonymisée, mais il reste de même signe (positif) que celui de la base CASD. Il est possible que cela tienne aux limites d'un tel exercice, les bénéficiaires de GIR 1 étant en effet les moins nombreux (7 000 individus environ). Pour les bénéficiaires de GIR 2, la base floutée fait ressortir les mêmes effets significatifs de chacune des caractéristiques que ceux estimés avec la base CASD (annexe 5, figure 30 et figure 31).

Concernant les bénéficiaires de GIR 3, les effets des caractéristiques des bénéficiaires sur le montant d'aide humaine notifiée sont similaires, l'ampleur est juste un peu différente pour les bénéficiaires les plus aisés (annexe 5, figure 32 et figure 33). La base CASD fait ressortir une baisse du montant d'aide humaine à mesure que les ressources augmentent, avec cependant une baisse qui ralentie pour les plus aisés : respectivement environ - 80 euros pour la tranche [1 500 ; 2 000[, - 95 euros pour la tranche [2 000 ; 2 500[et - 98 € pour la tranche à plus de 2 500 euros (comparativement aux bénéficiaires les moins aisés). Avec la base floutée, on estime aussi une baisse quand les ressources augmentent, mais elle est moins forte pour les plus aisés : – 76 euros pour la tranche [1 500 ; 2 000[, - 93 euros pour la tranche [2 000 ; 2 500[et - 87 euros pour la tranche à plus de 2 500 euros. Pour les bénéficiaires en GIR 4, les résultats sont tout à fait comparables (annexe 5, figure 34 et figure 35). Ainsi, chacune des variables (de même que leurs modalités) a les mêmes effets, d'une ampleur comparable sur le montant d'aide humaine notifié et ce, à autre caractéristiques égales.

7.3 Simulations avec le modèle Autonomix

La comparaison des résultats de simulations issus d'Autonomix porte sur les montants totaux annuels des plans notifiés par GIR et tranches fines de ressources (de même pour les montants mensuels moyens). Les plans « consommés » (montants totaux annuels et montants moyens mensuels) sont également examinés : en effet, les bénéfciaires de l'APA n'utilisent pas forcément tout le montant du plan qui leur est attribué [28].

Pour toutes les grandeurs examinées : les écarts entre les deux bases (bases floutée par rapport à la base de données confidentielles) concernant les montants par GIR sont limités. De faibles écarts ressortent sur les montants notifiés simulés pour les bénéficiaires de GIR 1 (en particulier pour le crédit d'impôt et le reste à charge, mais la différence ne dépasse pas 5 %). Les écarts varient de 5 % à 10 % pour le crédit d'impôt aide à domicile et le reste à charge calculés sur les plans consommés ce qui reste raisonnable. En effet, ces écarts reflètent des différences de montants totaux annuels faibles, par exemple 18 millions d'euros pour le reste à charge dans la base anonymisée contre 20 millions dans la base CASD.

De la même manière qu'on l'observait avec les statistiques descriptives et dans certaines sous-populations (les plus dépendants) pour les modèles de régressions, quelques différences ressortent toujours sur certaines tranches de ressources (parmi les plus élevées). On note, par exemple, un

28. Autonomix complète les données en entrée (RI APA, enquêtes CARE, etc.) en simulant d'autres éléments liés à la dépendance des personnes âgées, comme le « plan d'aide effectivement consommé », les réductions/crédits d'impôt liées à la dépendance et le montant restant à la charge des bénéficiaires.

écart en faveur de la base floutée (+ 10 %) pour les montants totaux annuels des plans d'APA notifiés dans la tranche de ressources de 2 800 à 3 200 euros, mais les sommes en jeu restent tout à fait du même ordre (11 millions d'euros de montant d'APA notifié pour la base floutée, contre 10 millions pour la base CASD).

Ventilés par tranches de ressources, on observe des écarts de montants totaux d'APA consommés sur toutes les tranches de ressources (en plus ou en moins dans la base floutée), mais qui restent dans des limites raisonnables (moins de 10 %). Même lorsque l'écart en pourcentage est un peu plus élevé, les montants simulés restent assez semblables que l'on utilise la base floutée ou que l'on mobilise la base de données confidentielles [29].

8 Conclusion et perspectives

En conclusion, les données de la base floutée permettent de produire des statistiques descriptives proches de celles issues de la base de données confidentielles pour les montants d'APA notifiés et ce même lorsque ces montants sont ventilés par GIR, sexe, situation familiale et tranches de ressources. De même, lorsque l'on modélise le montant du plan d'aide humaine notifié en fonction des caractéristiques des bénéficiaires, pour l'ensemble des GIR, comme pour chaque GIR pris séparément, les effets, à autres caractéristiques égales sont assez comparables d'une base à l'autre. Enfin, la base floutée conduit à des simulations de montants de plans notifiés et consommés qui restent proches de celles obtenues avec la base CASD.

Notons que lorsqu'il y a des écarts entre les deux sources, que nous jugeons cependant raisonnables, ils portent soit sur des sous-populations fines (les bénéficiaires les plus dépendants), soit sur certaines tranches de ressources. Aussi, peut-on entrevoir des pistes d'améliorations pour une nouvelle version de cette base floutée ou pour d'autres sources que l'on souhaiterait anonymiser :

* Envisager et implémenter d'autres méthodes de partitionnement, comme du clustering flou, pour construire les intervalles ;

* diffuser un sous-échantillon un peu plus important pour améliorer la robustesse, si la taille de la base de données de départ le permet ;

* tester d'autres méthodes perturbatrices pour les variables quantitatives, afin d'éviter les oscillations que l'on observe avec la méthode actuelle de simulations par intervalles et mieux gérer les écarts individuels entre valeur floutée et valeur observée [30].

Il n'est cependant pas acquis que même en changeant de méthodes on puisse atteindre « le » compromis optimal ou idéal entre une confidentialité totalement respectée et des informations floutées aussi proches que possible des valeurs exactes (soit une exactitude des messages délivrés).

29. Pour la tranche haute de ressources (3 200 euros et plus), on a 15 % d'effectifs en moins dans la base floutée, ce qui se répercute sur le montant annuel d'APA consommé, ainsi que sur les autres grandeurs telles que le reste à charge.

30. En particulier, la méthode actuelle ne garantit pas que la valeur simulée - pour les ressources par exemple - soit nulle, lorsque la valeur observée l'est.

Bibliographie

[1] Mission Bothorel, Pour une politique publique de la donnée, *Rapport de la mission confiée par le Premier ministre*, décembre 2020.

[2] Ouverture du code du modèle de microsimulation AUTONOMIX sous GitLab.

[3] Jess N., Bergeat M. et Dupont F., Création de fichiers anonymisés à partir d'une base médico-administrative (le PMSI) : un exemple pratique de mise en oeuvre des méthodes de protection des fichiers de données individuelles, Journées de Méthodologie Statistique, Insee, 2015.

[4] Bergeat M., Anonymisation de données individuelles : bien calées, bien protégées ?, Journées de Méthodologie Statistiques, Insee, 2015.

[5] Fung B.C.M, Wang K., Chen R., Yu P.S., Privacy preserving data publishing : a survey of recent developments, *ACM computing surveys*, 42, 4, article 14, 2010.

[6] Bergeat M., La gestion de la confidentialité pour les données individuelles, *Document de travail*, No M2016/07, Insee, 2016.

[7] Jachiet P.-A., Anonymisation : enjeux techniques et conséquences pratiques à l'heure des mégadonnées, Comité de direction élargi, Drees, 12 juillet 2019.

[8] Bergeat M., Un panorama de la protection des fichiers de données individuelles, Séminaire de méthodologie statistique, Insee, 24 juin 2014.

[9] Données de santé, anonymat et risque de ré-identification, *Dossiers Solidarité et santé*, n°64, Drees, juillet 2015.

[10] Arnault L. et Roy D.,Allocation personnalisée d'autonomie : en 2017, un bénéficiaire sur deux n'utilise pas l'intégralité du montant d'aide humaine notifié, *Études et résultats*, n°1153, Drees, juin 2020.

Annexe 1. Quelles sont les données à caractère personnel, les données identifiantes, directement ou indirectement ?

Qu'est-ce qu'une donnée à caractère personnel ?

Il s'agit d'une information que l'on peut relier à une personne physique reconnaissable immédiatement ou après une recherche d'une ampleur "raisonnable" (RGPD, article 4-1, RGPD considérant 26). Ceci est indépendant de la nature, de la sensibilité, de la "publicité" ou "notoriété" de la donnée (séminaire Confidentialité de l'Insee, 24-06-2019, Unité Affaires Juridiques et Contentieuses).
RGPD, article 4-1 : "on entend par données à caractère personnel toute information se rapportant à une personne physique identifiée ou identifiable [...] ; est réputée être une "personne physique identifiable" une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale".
RGPD, considérant 26 (extrait), "les moyens raisonnables" :" Pour déterminer si une personne physique est identifiable, il convient de prendre en considération l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne pour identifier la personne physique directement ou indirectement, tels que le ciblage. Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne, il convient de prendre en considération l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci".

On peut définir plus précisément ce que sont les données directement et indirectement identifiantes. Le site de la CNIL, en fourni d'ailleurs des exemples.

Les données directement identifiantes

Les données directement identifiantes sont celles qui indiquent (ou peuvent indiquer) clairement l'identité de la personne : le(s) nom(s) [nom de naissance, nom marital], le(s) prénom(s).

Les données indirectement identifiantes

Pour ce qui est des données indirectement identifiantes, on peut citer : le NIR, l'adresse (ou les adresses) postale(s) ou courriel (par exemple, un e-mail nominatif : prenom.nom@yahoo.fr), les coordonnées géographiques du lieu d'habitation, le(s) numéro(s) de téléphone (fixe, portable), la date de naissance précise (JJ/MM/AAAA), et plus généralement tout numéro d'identification accessible au public ou à un tiers (un numéro d'allocataire, un numéro de matricule, un identifiant fiscal, un numéro client, un numéro de plaque d'immatriculation, etc.). Rentrent aussi dans cette catégorie : la voix ou l'image. Ces données ne permettent pas isolément de savoir immédiatement à qui correspond ces informations. Il est cependant possible de retrouver l'identité de la personne soit en croisant toutes ces informations soit par association avec une autre base de donnée (détenue en interne ou par un tiers).

L'identification d'une personne physique peut être réalisée à partir d'une seule donnée (le nom) ou à partir du croisement d'un ensemble de données (par exemple : une femme vivant à telle adresse, née tel jour et membre dans telle association).

Annexe 2. Méthodes et règles appliquées selon les variables considérées

FIGURE 6 – Méthodes et règles appliquées aux variables floutées

Les variables floutées	Signification	Calcul ou précisions concernant les Xi	Mode de floutage
APA_FLOU	Montant mensuel du plan d'aide APA notifié	APA = NOTHPB + NOTPB APA = montant APA notifié relevant du CD + montant de la participation financière du bénéficiaire	**Floutage par régression** X_i = notamment l'âge, le GIR et d'autres variables socio-démographique
NOTHPB_FLOU	Montant mensuel de l'APA notifié relevant du CD	NOTHPB_FLOU = (NOTHPB/APA) * **APA_FLOU** On conserve après floutage la même part du montant relevant du CD dans le montant total.	*Floutage indirect*
NOTPB_FLOU	Montant mensuel de la participation financière notifiée du bénéficiaire APA	NOTPB_FLOU = (NOTPB/APA) * **APA_FLOU** On conserve après floutage la même part de participation financière dans le montant total.	*Floutage indirect*
DECHPB_APA_FLOU	Montant total de l'APA versé par le CD en décembre 2017 (ou dernier mois d'ouverture des droits)	DECHPB_APA_FLOU = (DECHPB_APA/NOTHPB) * NOTHPB_FLOU On conserve après floutage la même part du montant versé par le CD dans le au montant total.	*Floutage indirect*
AIDEHUM_HEURE_FLOU	Nombre d'heures d'aide humaine notifié	AIDEHUM_HEURE_FLOU = (AIDEHUM_HEURE/AIDEHUM_NOT) * AIDEHUM_NOT_FLOU On conserve après floutage la même part du nombre d'heures d'aides humaines par rapport au montant mensuel de l'aide humaine.	*Floutage indirect*
AIDEHUM_NOT_FLOU	Montant mensuel d'aide humaine notifié	AIDEHUM_NOT_FLOU = (AIDEHUM_NOT/APA) * APA_FLOU On conserve après floutage la même part du montant d'aide humaine dans le montant du plan d'aide notifié.	*Floutage indirect*
DECHUM_APA_FLOU	Montant versé par CD pour les aides humaines, au titre du mois de décembre 2017 ou au titre du dernier mois d'ouverture des droits	DECHUM_APA_FLOU = (DECHUM_APA/ AIDEHUM_NOT) * AIDEHUM_NOT_FLOU On conserve après floutage la même part du montant d'aide humaine versé en décembre dans le montant mensuel d'aide humaine.	*Floutage indirect*
RESSOURC_APA	Ressources au sens de l'APA	AIDEACC : Accueil de jour dans le plan d'aide APA notifié AIDETRANS : Aide au transport dans le plan d'aide APA notifié	**Floutage par régression** X_i = notamment l'âge, le département, etc.
DATE_APAD_FLOU	Date d'ouverture des droits à l'APA à domicile		**Floutage par régression** X_i = notamment le GIR, l'âge, etc.
DATEEVAL_APA_FLOU	Date de la dernière évaluation GIR du bénéficiaire de l'APA (quel que soit le type d'APA).	DATEEVAL_APA_FLOU = DATE_APAD_FLOU + (DATEEVAL_APA - DATE_APAD) On conserve le même écart de temps qu'observé entre la date d'évaluation et la date d'ouverture des droits	Floutage indirect
AGE_FLOU	Âge en années révolues		**Floutage par régression** X_i = notamment le département, le sexe, etc.

Annexe 3. Densités des variables initiales et des variables floutées

Via le rapport automatisé (R Markdown), nous produisons systématiquement les graphiques de densité entre les variables initiales (en bleu) et les variables floutées (en rouge). En rouge pointillé, il s'agit des variables floutées sans l'imputation des valeurs manquantes pour les observations qui avaient des valeurs non manquantes pour la variable initiale. Nous mentionnons la proportion d'observations concernées, pour les variables ayant fait l'objet d'imputation des valeurs manquantes générées par le floutage.

FIGURE 7 – Montant mensuel du plan d'aide APA notifié

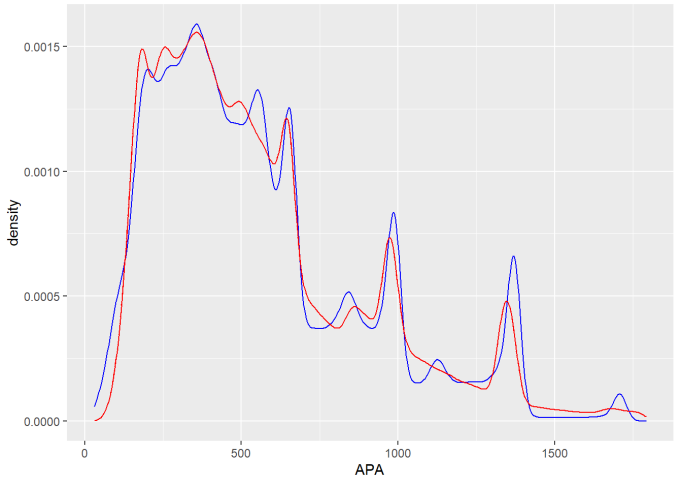


FIGURE 8 – Montant mensuel des ressources au sens de l'APA (plage 0 à 5 000 euros)

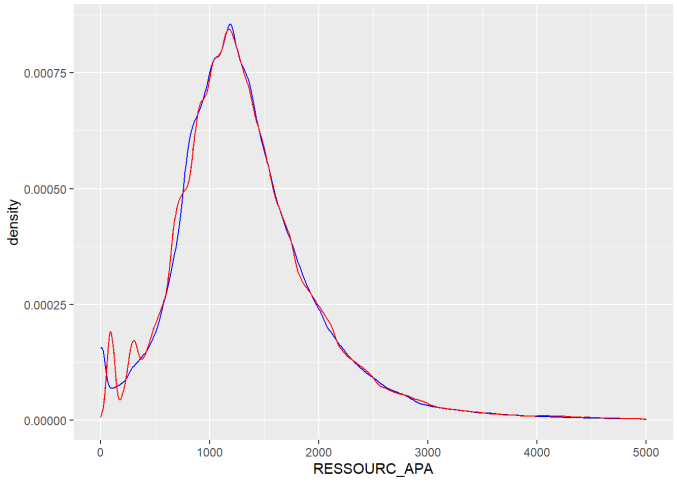


FIGURE 9 – Date d'ouverture des droits à l'APA

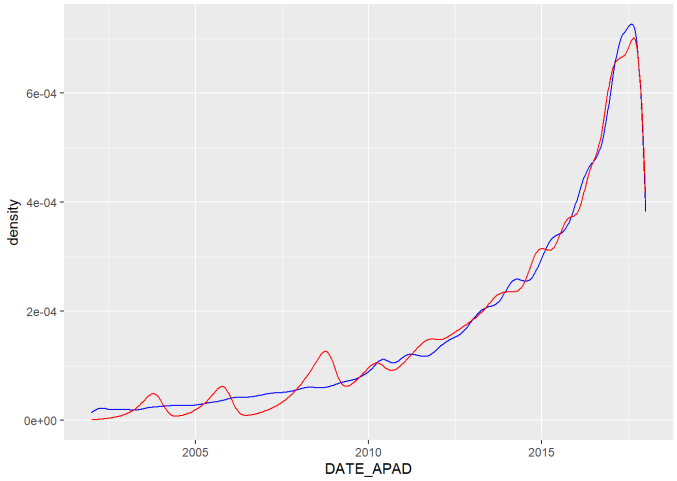


FIGURE 10 – Âge des bénéficiaires de l'APA (années révolues)

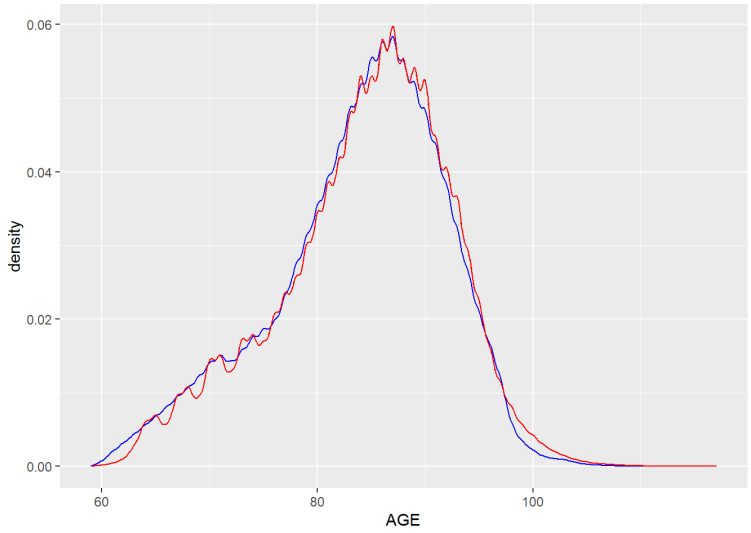


FIGURE 11 – Montant mensuel de l'APA notifié relevant du CD

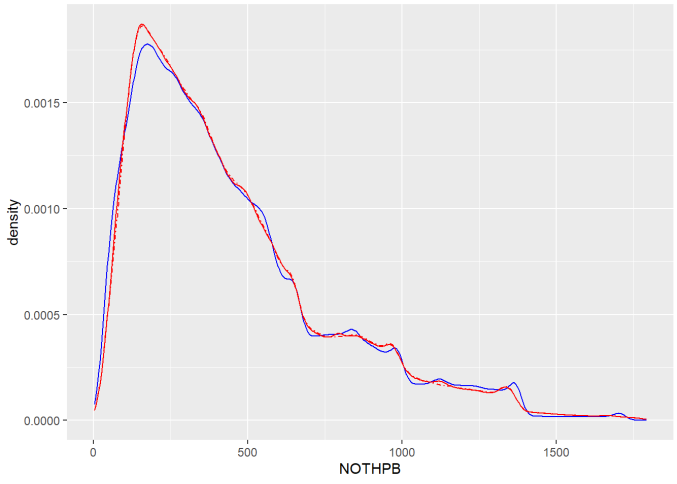


FIGURE 12 – Montant mensuel de la participation financière du bénéficiaire au plan APA notifié

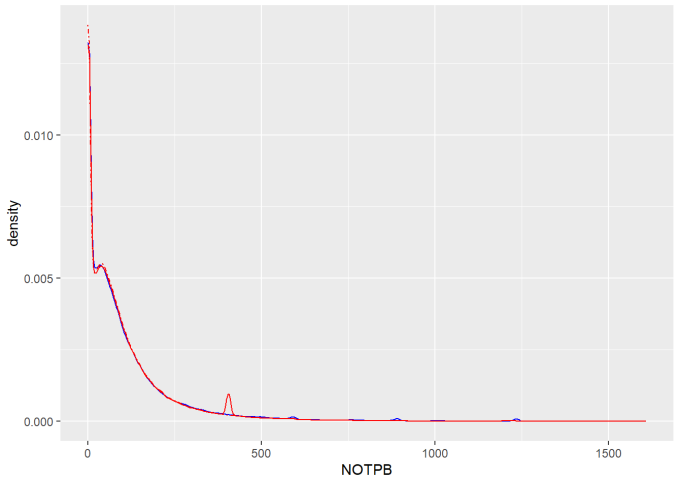


FIGURE 13 – Montant total de l'APA versé par le CD en décembre 2017

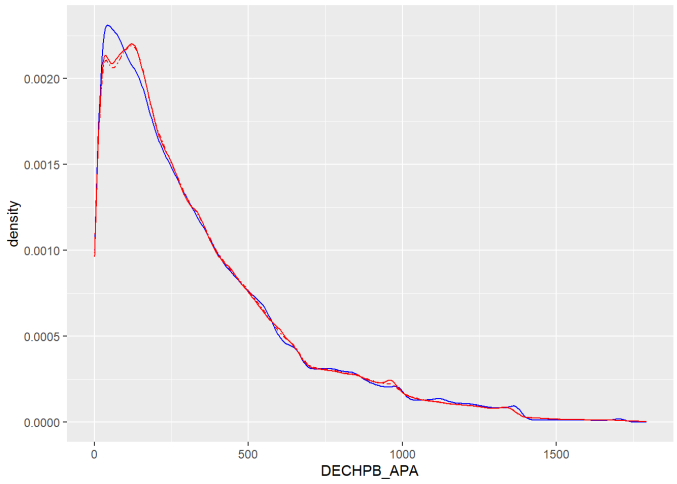


FIGURE 14 – Montant mensuel d'aide humaine notifiée

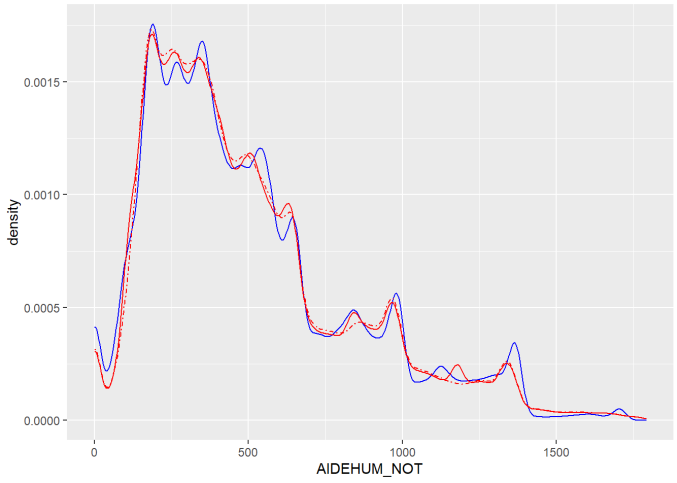


FIGURE 15 – Nombre d'heures d'aide humaine notifiée (plage 0 à 150 heures

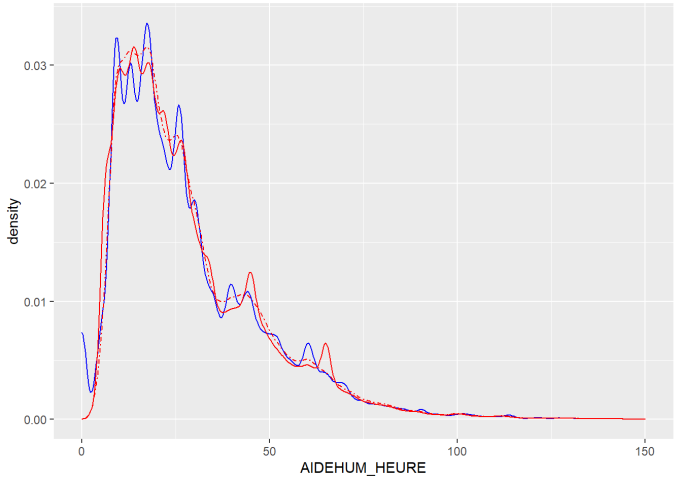


FIGURE 16 – Montant de l'APA versé par le CD pour l'aide humaine

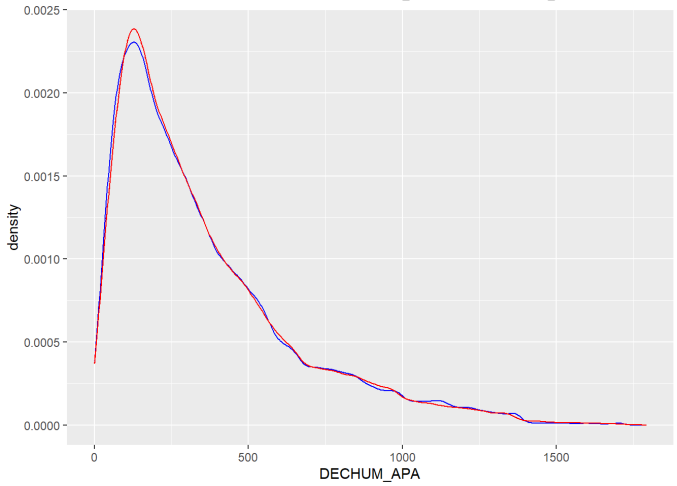
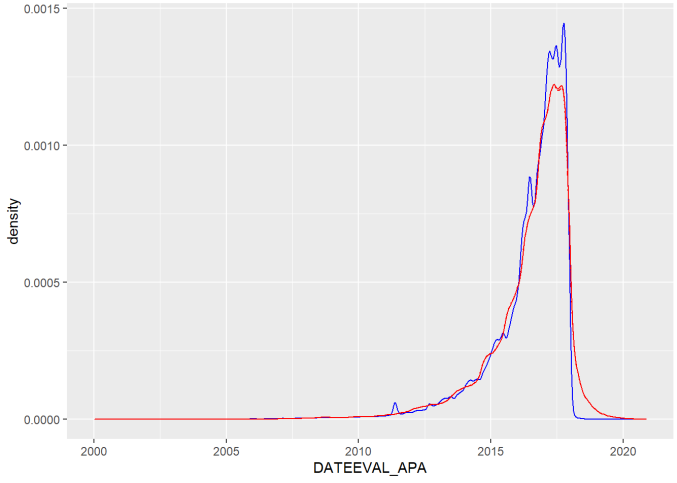


FIGURE 17 – Date de la dernière évaluation du GIR



La proportion d'estimation liée à l'imputation des NA (par la médiane de la tranche) parmi les observations qui avaient une valeur non-manquante dans la variable initiale est de :
* 1.3 % pour le montant mensuel de l'APA notifié relevant du CD ;
* 1.9 % pour le montant mensuel de la participation financière du bénéficiaire au plan APA notifié ;
* 1.5 % pour le montant total de l'APA versé par le CD en décembre 2017 ;
* 3.5 % pour le montant mensuel d'aide humaine notifiée ;
* 9.4 % pour le nombre d'heures d'aide humaine notifiée ;
* 0.4 % pour la date de la dernière évaluation du GIR.

Annexe 4. Statistiques descriptives comparatives : base floutée/base de données confidentielles

FIGURE 18 – Distribution de l'âge des bénéficiaires par sexe

Sexe	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Femme	RI floutée	74,0	80,0	86,0	90,0	94,0	84,9
écart en %		*1,4*	*0,0*	*1,2*	*0,0*	*1,1*	*0,6*
Homme	RI floutée	71,0	78,0	84,0	89,0	93,0	83,0
écart en %		*1,4*	*1,3*	*0,0*	*1,1*	*1,1*	*0,7*
Femme	Base CASD	73,0	80,0	85,0	90,0	93,0	84,4
Homme	Base CASD	70,0	77,0	84,0	88,0	92,0	82,4

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée

FIGURE 19 – Distribution des ressources, selon le sexe et la situation familiale

Caractéristiques	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Femme	RI floutée	642,0	918,0	1230,0	1610,0	2100,0	1329,7
écart en %		*-0,3*	*-0,2*	*-0,1*	*0,1*	*0,3*	*-0,4*
Homme	RI floutée	661,0	956,0	1299,0	1740,0	2294,0	1428,9
écart en %		*-0,4*	*-0,1*	*0,1*	*0,2*	*0,0*	*-0,4*
En couple	RI floutée	705,0	968,0	1255,0	1654,0	2165,0	1380,8
écart en %		*-1,4*	*-0,3*	*0,1*	*0,2*	*0,1*	*-0,3*
Seul	RI floutée	602,0	902,0	1243,0	1644,0	2145,0	1345,1
écart en %		*-0,2*	*-0,2*	*-0,2*	*0,0*	*-0,1*	*-0,4*
Femme	Base CASD	644,1	920,0	1231,0	1608,0	2094,7	1334,8
Homme	Base CASD	663,9	957,1	1298,0	1735,7	2293,0	1435,1
En couple	Base CASD	715,0	970,5	1253,5	1650,5	2162,4	1384,7
Seul	Base CASD	603,1	904,0	1245,0	1644,0	2147,0	1351,0

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 20 – Distribution du montant du plan d'aide notifié, selon le GIR (en euros mensuels)

GIR	Source	P10	Q1	Médiane	Q3	P90	Moyenne	% de participation du bénéficiaire	% de plans saturés à 96 %
GIR 1 (1)	RI floutée	573,0	954,1	1 309,0	1 656,0	1 737,0	1 237,0	16,0	26,7
écart en % ou en points (2)		***-1,2***	***-0,6***	***-0,5***	***-0,6***	***1,5***	***-0,4***	***0,0***	***0,0***
GIR 2	RI floutée	387,0	663,0	1 007,0	1274,0	1 360,0	941,7	19,4	21,9
écart en % ou en points (2)		***0,3***	***-0,3***	***-0,8***	***-1,0***	***-0,8***	***-0,2***	***0,0***	***0,0***
GIR 3	RI floutée	294,0	466,0	683,0	884,0	974,0	663,1	20,0	16,4
écart en % ou en points (2)		***2,1***	***0,2***	***0,1***	***1,1***	***-1,1***	***0,8***	***0,0***	***-0,1***
GIR 4	RI floutée	176,0	245,0	359,0	495,0	606,0	374,3	21,8	6,9
écart en % ou en points (2)		***2,9***	***0,8***	***-0,3***	***-0,6***	***1,0***	***0,9***	***0,0***	***0,0***
GIR 1	Base CASD	580,0	960,0	1 316,0	1 666,0	1 711,0	1 241,9	16,0	26,7
GIR 2	Base CASD	386,0	665,0	1 015,0	1 287,0	1 371,0	943,4	19,4	21,9
GIR 3	Base CASD	288,0	465,0	682,0	874,0	985,0	658,1	20,0	16,5
GIR 4	Base CASD	171,0	243,0	360,0	498,0	600,0	371,2	21,8	6,9

Note > (1) le GIR 1 correspond au plus fort niveau de perte d'autonomie ; (2) : les écarts de montants (quantiles, médiane et moyenne) sont calculés en pourcentage ; en revanche les écarts de répartitions sont calculés en point (2 colonnes de droite).
Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 21 – Distribution du montant du plan d'aide notifié, selon les ressources (en euros mensuels)

Tranches de ressources	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Inférieures à 803 €	RI floutée	231,0	346,0	515,0	753,0	1 066,0	589,6
écart en %		***1,3***	***-1,1***	***-1,9***	***-1,6***	***-1,7***	***-0,3***
Entre 803 € et 1 000 €	RI floutée	208,0	317,0	489,0	733,0	1 056,0	569,4
écart en %		***2,5***	***0,6***	***0,4***	***1,1***	***1,1***	***1,0***
Entre 1 000 € et 1 200 €	RI floutée	191,0	293,0	466,0	705,0	1 048,0	551,8
écart en %		***0,0***	***0,0***	***-0,2***	***0,6***	***0,8***	***0,3***
Entre 1 200 € et 1 400 €	RI floutée	187,0	283,0	452,0	687,0	1 029,0	541,1
écart en %		***1,6***	***1,1***	***1,1***	***3,2***	***3,4***	***1,1***
Entre 1 400 € et 1 600 €	RI floutée	184,0	273,0	438,0	662,0	993,0	526,3
écart en %		***2,2***	***0,7***	***0,2***	***-0,2***	***-0,1***	***0,6***
Entre 1 600 € et 1 800 €	RI floutée	183,0	271,0	436,0	660,0	989,0	524,5
écart en %		***2,8***	***-0,4***	***-0,2***	***-0,3***	***-0,5***	***0,4***
Entre 1 800 € et 2 000 €	RI floutée	185,0	272,0	442,0	663,0	992,0	529,2
écart en %		***3,4***	***0,0***	***0,0***	***0,0***	***-0,2***	***0,8***
Entre 2 000 € et 2 500 €	RI floutée	185,0	274,0	447,0	668,0	993,0	530,7
écart en %		***2,2***	***-0,4***	***-0,2***	***0,6***	***-0,1***	***0,1***
Supérieures à 2 500 €	RI floutée	223,0	341,0	528,0	804,0	11 56,6	607,4
écart en %		***4,2***	***0,6***	***-1,5***	***-1,3***	***1,5***	***0,3***
Inférieures à 803 €	Base CASD	228,0	350,0	525,0	765,0	1 084,0	591,4
Entre 803 € et 1 000 €	Base CASD	203,0	315,0	487,0	725,0	1 044,0	564,0
Entre 1 000 € et 1 200 €	Base CASD	191,0	293,0	467,0	701,0	1 040,0	549,9
Entre 1 200 € et 1 400 €	Base CASD	184,0	280,0	447,0	666,0	995,0	535,0
Entre 1 400 € et 1 600 €	Base CASD	180,0	271,0	437,0	663,0	994,0	523,4
Entre 1 600 € et 1 800 €	Base CASD	178,0	272,0	437,0	662,0	994,0	522,4
Entre 1 800 € et 2 000 €	Base CASD	179,0	272,0	442,0	663,0	994,0	524,8
Entre 2 000 € et 2 500 €	Base CASD	181,0	275,0	448,0	664,0	994,0	530,0
Supérieures à 2 500 €	Base CASD	214,0	339,0	536,0	815,0	1 140,0	605,4

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles 2017, base au CASD et base floutée.

FIGURE 22 – Distribution du montant de l'APA notifié versé par le CD, selon le sexe et la situation de vie en couple (en euros mensuels)

Caractéristiques	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Femme	RI floutée	48,4	123,6	262,6	490,8	805,4	352,4
écart en %		***5,2***	***4,8***	***0,2***	***0,1***	***0,3***	***0,6***
Homme	RI floutée	46,2	107,4	222,2	432,4	703,7	311,6
écart en %		***7,5***	***7,4***	***0,6***	***0,1***	***-0,3***	***0,7***
En couple	RI floutée	47,7	105,3	206,7	402,5	674,0	298,5
écart en %		***7,7***	***8,6***	***0,3***	***0,1***	***-0,3***	***0,6***
Seul	RI floutée	48,8	129,0	281,5	512,8	824,0	365,8
écart en %		***6,0***	***3,2***	***0,2***	***0,3***	***0,5***	***0,7***
Femme	Base CASD	46,0	118,0	262,0	490,0	803,0	350,3
Homme	Base CASD	43,0	100,0	221,0	432,0	706,0	309,3
En couple	Base CASD	44,3	97,0	206,0	402,0	676,0	296,7
Seul	Base CASD	46,0	125,0	281,0	511,1	820,0	363,4

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 23 – Distribution du montant de l'APA notifié versé par le CD, selon les ressources (en euros mensuels)

Tranches de ressources	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Inférieures à 803 €	RI floutée	112,1	239,0	418,0	638,5	961,0	481,8
écart en %		***11,0***	***-1,2***	***-0,5***	***-0,9***	***-1,3***	***-0,5***
Entre 803 € et 1 000 €	RI floutée	79,1	182,3	345,9	577,9	912,3	426,4
écart en %		***8,3***	***2,4***	***0,7***	***1,6***	***0,7***	***1,0***
Entre 1 000 € et 1 200 €	RI floutée	62,1	147,9	282,8	503,2	841,8	371,7
écart en %		***9,0***	***2,7***	***-0,2***	***0,0***	***0,1***	***0,4***
Entre 1 200 € et 1 400 €	RI floutée	53,8	127,7	243,4	445,0	771,1	331,6
écart en %		***5,6***	***5,4***	***1,0***	***0,0***	***1,1***	***1,4***
Entre 1 400 € et 1 600 €	RI floutée	47,4	109,9	205,1	390,8	673,9	290,3
écart en %		***5,3***	***4,6***	***-0,9***	***-0,5***	***-0,9***	***0,0***
Entre 1 600 € et 1 800 €	RI floutée	45,2	99,9	186,4	353,5	617,3	265,3
écart en %		***10,3***	***7,4***	***0,8***	***-0,8***	***-0,5***	***1,2***
Entre 1 800 € et 2 000 €	RI floutée	36,4	82,5	154,8	306,9	537,4	228,5
écart en %		***-1,7***	***0,6***	***-3,9***	***-3,2***	***-2,6***	***-2,1***
Entre 2 000 € et 2 500 €	RI floutée	32,9	69,5	131,9	259,7	451,8	195,4
écart en %		***9,8***	***10,4***	***4,7***	***4,4***	***3,9***	***5,8***
Supérieures à 2 500 €	RI floutée	16,0	32,0	63,2	119,0	201,7	92,5
écart en %		***14,2***	***-3,0***	***3,6***	***4,4***	***2,4***	***1,6***
Inférieures à 803 €	Base CASD	101,0	242,0	420,0	644,0	974,0	484,1
Entre 803 € et 1 000 €	Base CASD	73,0	178,0	343,7	569,0	906,0	422,0
Entre 1 000 € et 1 200 €	Base CASD	57,0	144,0	283,4	503,0	841,0	370,2
Entre 1 200 € et 1 400 €	Base CASD	51,0	121,2	241,0	445,0	763,0	327,1
Entre 1 400 € et 1 600 €	Base CASD	45,0	105,0	207,0	392,7	680,0	290,2
Entre 1 600 € et 1 800 €	Base CASD	41,0	93,0	185,0	356,4	620,5	262,2
Entre 1 800 € et 2 000 €	Base CASD	37,0	82,0	161,0	317,0	552,0	233,4
Entre 2 000 € et 2 500 €	Base CASD	30,0	63,0	126,0	248,7	435,0	184,7
Supérieures à 2 500 €	base CASD	14,0	33,0	61,0	114,0	197,0	91,0

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 24 – Distribution du montant d'APA notifié à la charge des bénéficiaires, selon le sexe et la situation de vie en couple (en euros mensuels)

Caractéristiques	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Femme	RI floutée	0,0	16,7	64,1	136,8	261,4	106,8
écart en %		***-***	***4,2***	***1,7***	***-0,2***	***-0,2***	***-2,2***
Homme	RI floutée	0,0	19,1	68,8	153,0	301,4	118,1
écart en %		***-***	***6,2***	***2,7***	***0,0***	***0,6***	***-2,2***
En couple	RI floutée	0,0	21,4	61,5	132,0	258,8	106,0
écart en %		***-***	***7,2***	***2,5***	***0,1***	***-0,4***	***-1,6***
Seul	RI floutée	0,0	14,0	68,0	147,2	282,9	113,0
écart en %		***-***	***0,2***	***1,5***	***0,1***	***0,3***	***-2,5***
Femme	Base CASD	0,0	16,0	63,0	137,0	262,0	109,2
Homme	Base CASD	0,0	18,0	67,0	153,0	299,6	120,8
En couple	Base CASD	0,0	20,0	60,0	131,8	260,0	107,7
Seul	Base CASD	0,0	14,0	67,0	147,0	282,0	115,9

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 25 – Distribution du montant d'APA notifié à la charge des bénéficiaires, selon les ressources (en euros mensuels)

Tranches de ressources	Source	P10	Q1	Médiane	Q3	P90	Moyenne
Inférieures à 803 €	RI floutée	0,0	0,0	0,0	0,0	4,9	4,1
écart en %		**-**	**-**	**-**	**-**	**-**	***8,6***
Entre 803 € et 1 000 €	RI floutée	0,0	7,6	16,2	26,7	36,6	19,3
écart en %		**-**	***-4,7***	***1,3***	***2,6***	***1,6***	***1,8***
Entre 1 000 € et 1 200 €	RI floutée	23,1	34,7	49,7	68,0	88,0	53,9
écart en %		***4,9***	***2,1***	***1,5***	***3,1***	***4,7***	***3,5***
Entre 1 200 € et 1 400 €	RI floutée	37,4	55,5	81,2	109,5	141,7	86,3
écart en %		***1,2***	***-1,0***	***-2,2***	***-0,5***	***1,2***	***0,3***
Entre 1 400 € et 1 600 €	RI floutée	52,1	77,7	116,4	159,1	208,3	125,0
écart en %		***4,2***	***0,9***	***-0,3***	***1,3***	***2,1***	***2,4***
Entre 1 600 € et 1 800 €	RI floutée	64,9	97,4	147,7	205,9	275,0	161,1
écart en %		***2,3***	***-1,6***	***-2,8***	***-1,0***	***-1,1***	***-0,2***
Entre 1 800 € et 2 000 €	RI floutée	82,9	123,8	193,7	277,5	380,2	213,8
écart en %		***7,7***	***3,1***	***3,0***	***5,9***	***6,3***	***5,2***
Entre 2 000 € et 2 500 €	RI floutée	95,9	145,1	229,1	343,6	484,6	264,0
écart en %		***-1,7***	***-5,5***	***-7,3***	***-5,3***	***-5,7***	***-4,9***
Supérieures à 2 500 €	RI floutée	169,9	292,6	404,7	548,1	847,5	457,6
écart en %		***1,8***	***1,6***	***-10,5***	***-19,6***	***-10,3***	***-10,6***
Inférieures à 803 €	base CASD	0,0	0,0	0,0	0,0	0,0	3,8
Entre 803 € et 1 000 €	base CASD	3,0	8,0	16,0	26,1	36,0	18,9
Entre 1 000 € et 1 200 €	base CASD	22,0	34,0	49,0	66,0	84,0	52,1
Entre 1 200 € et 1 400 €	base CASD	37,0	56,0	83,0	110,0	140,0	86,1
Entre 1 400 € et 1 600 €	base CASD	50,0	77,0	116,7	157,0	204,0	122,1
Entre 1 600 € et 1 800 €	base CASD	63,4	99,0	152,0	208,0	278,0	161,5
Entre 1 800 € et 2 000 €	base CASD	77,0	120,0	188,0	262,0	357,6	203,3
Entre 2 000 € et 2 500 €	base CASD	97,5	153,4	247,0	363,0	514,0	277,5
Supérieures à 2 500 €	base CASD	167,0	288,0	452,0	682,0	945,0	512,0

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 26 – Distribution de la date d'ouverture des droits à l'APA

Source	Année	Effectif	Répartition
RI floutée	2012	197 817	25,91
écart en % ou en point (pour la répartition)		***1,0***	***0,3***
RI floutée	2013	59 703	7,82
écart en % ou en point (pour la répartition)		***3,4***	***0,3***
RI floutée	2014	73 591	9,64
écart en % ou en point (pour la répartition)		***0,8***	***0,1***
RI floutée	2015	96 014	12,57
écart en % ou en point (pour la répartition)		***0,5***	***0,1***
RI floutée	2016	135 559	17,75
écart en % ou en point (pour la répartition)		***1,8***	***0,3***
RI floutée	2017	200 917	26,31
écart en % ou en point (pour la répartition)		***-3,5***	***-1,0***
Base CASD	2012	195 843	25,65
Base CASD	2013	57 768	7,56
Base CASD	2014	73 007	9,56
Base CASD	2015	95 564	12,51
Base CASD	2016	133 158	17,44
Base CASD	2017	208 305	27,28

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

FIGURE 27 – Distribution de la date de la dernière évaluation du GIR

Source	Année	Effectif	Répartition
RI floutée	2012	26 716	3,69
écart en % ou en point (pour la répartition)		***2,7***	***0,1***
RI floutée	2013	22 150	3,06
écart en % ou en point (pour la répartition)		***2,4***	***0,07***
RI floutée	2014	43 840	6,05
écart en % ou en point (pour la répartition)		***4,9***	***0,28***
RI floutée	2015	90 484	12,49
écart en % ou en point (pour la répartition)		***6,5***	***0,77***
RI floutée	2016	200 966	27,73
écart en % ou en point (pour la répartition)		***-4,3***	***-1,26***
RI floutée	2017	340 490	46,99
écart en % ou en point (pour la répartition)		***0,1***	***0,04***
Base CASD	2012	26 015	3,59
Base CASD	2013	21 625	2,98
Base CASD	2014	41 809	5,77
Base CASD	2015	84 928	11,72
Base CASD	2016	210 095	28,99
Base CASD	2017	340 238	46,95

Champ > Bénéficiaires de l'APA à domicile en France entière payés au titre de décembre 2017 (hors valeurs manquantes).
Sources > Remontées individuelles RI- APA 2017, base au CASD et base floutée.

Annexe 5. Modèles de régressions comparatifs : base floutée/base de données confidentielles

FIGURE 28 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 1 - base floutée

Variables	GIR1 - Montant moyen notifié (en €)	GIR1 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	1 030	Ref.
Âge : [75 ; 80[	1 040	7.6
Âge : [80 ; 85[	1 130	74.4***
Âge : [85 ; 90[	1 140	73.2***
Âge : 90 ou +	1 190	90.3***
Situation : Homme seul	1 140	136.2***
Situation : Homme en couple	1 020	Ref.
Situation : Femme seule	1 220	156.7***
Situation : Femme en couple	1 090	71.3***
Ressources (€ / mois) : [0 ; 739,8[	1 190	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	1 190	**5.7**
Ressources (€ / mois) : [1000 ; 1250[	1 170	-9.3
Ressources (€ / mois) : [1250 ; 1500[	1 150	-31.4*
Ressources (€ / mois) : [1500 ; 2000[	1 090	-99.9***
Ressources (€ / mois) : [2000 ; 2500[	1 030	-149.2***
Ressources (€ / mois) : 2500 ou +	930	-270.5***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 1 dont l'âge est compris entre 75 et 80 ans s'élève à 1 040 euros. À caractéristiques identiques, ce montant n'est pas significativement différent de celui qui serait notifié à un bénéficiaire de GIR 1 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 1 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base floutée, Drees.

FIGURE 29 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 1 - base CASD

Variables	GIR1 - Montant moyen notifié (en €)	GIR1 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	1 060	Ref.
Âge : [75 ; 80[	1 060	4.6
Âge : [80 ; 85[	1 130	49.8***
Âge : [85 ; 90[	1 170	77.3***
Âge : 90 ou +	1 200	82.4***
Situation : Homme seul	1 160	143.0***
Situation : Homme en couple	1 030	Ref.
Situation : Femme seule	1 230	160.0***
Situation : Femme en couple	1 110	85.2***
Ressources (€ / mois) : [0 ; 739,8[	1 200	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	1 220	26.6**
Ressources (€ / mois) : [1000 ; 1250[	1 170	-13.5
Ressources (€ / mois) : [1250 ; 1500[	1 160	-20.8
Ressources (€ / mois) : [1500 ; 2000[	1 090	-94.2***
Ressources (€ / mois) : [2000 ; 2500[	1 030	-157.0***
Ressources (€ / mois) : 2500 ou +	990	-218.4***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 1 dont l'âge est compris entre 75 et 80 ans s'élève à 1 060 euros. À caractéristiques identiques, ce montant n'est pas significativement différent de celui qui serait notifié à un bénéficiaire de GIR 1 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 1 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base CASD, Drees.

FIGURE 30 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 2 - base floutée

Variables	GIR2 - Montant moyen notifié (en €)	GIR2 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	780	Ref.
Âge : [75 ; 80[	780	18.0**
Âge : [80 ; 85[	830	51.2***
Âge : [85 ; 90[	880	75.6***
Âge : 90 ou +	930	91.3***
Situation : Homme seul	930	241.5***
Situation : Homme en couple	690	Ref.
Situation : Femme seule	970	242.8***
Situation : Femme en couple	770	81.9***
Ressources (€ / mois) : [0 ; 739,8[	910	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	920	5.3
Ressources (€ / mois) : [1000 ; 1250[	890	-13.9**
Ressources (€ / mois) : [1250 ; 1500[	880	-34.4***
Ressources (€ / mois) : [1500 ; 2000[	820	-79.9***
Ressources (€ / mois) : [2000 ; 2500[	780	-115.9***
Ressources (€ / mois) : 2500 ou +	790	-132.7***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 2 dont l'âge est compris entre 75 et 80 ans s'élève à 780 euros. À caractéristiques identiques, ce montant est supérieur de 18 euros à celui qui serait notifié à un bénéficiaire de GIR 2 âgé de moins de 75 ans.
Champ > Bénéficiaires de l'APA à domicile de GIR 2 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base floutée, Drees.

FIGURE 31 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 2 - base CASD

Variables	GIR2 - Montant moyen notifié (en €)	GIR2 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	780	Ref.
Âge : [75 ; 80[	790	26.5***
Âge : [80 ; 85[	830	54.9***
Âge : [85 ; 90[	890	79.7***
Âge : 90 ou +	940	93.4***
Situation : Homme seul	950	254.6***
Situation : Homme en couple	700	Ref.
Situation : Femme seule	970	242.5***
Situation : Femme en couple	780	84.0***
Ressources (€ / mois) : [0 ; 739,8[	920	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	910	-5.9
Ressources (€ / mois) : [1000 ; 1250[	890	-24.2***
Ressources (€ / mois) : [1250 ; 1500[	880	-49.3***
Ressources (€ / mois) : [1500 ; 2000[	830	-87.5***
Ressources (€ / mois) : [2000 ; 2500[	770	-131.9***
Ressources (€ / mois) : 2500 ou +	790	-143.4***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 2 dont l'âge est compris entre 75 et 80 ans s'élève à 790 euros. À caractéristiques identiques, ce montant est supérieur de 26,5 euros à celui qui serait notifié à un bénéficiaire de GIR 2 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 2 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base CASD, Drees.

FIGURE 32 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 3 - base floutée

Variables	GIR3 - Montant moyen notifié (en €)	GIR3 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	570	Ref.
Âge : [75 ; 80[	560	12.4**
Âge : [80 ; 85[	580	33.5***
Âge : [85 ; 90[	610	51.6***
Âge : 90 ou +	650	71.2***
Situation : Homme seul	650	166.4***
Situation : Homme en couple	480	Ref.
Situation : Femme seule	660	160.9***
Situation : Femme en couple	550	73.4***
Ressources (€ / mois) : [0 ; 739,8[	650	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	640	-16.8***
Ressources (€ / mois) : [1000 ; 1250[	610	-40.3***
Ressources (€ / mois) : [1250 ; 1500[	600	-56.8***
Ressources (€ / mois) : [1500 ; 2000[	580	-76.0***
Ressources (€ / mois) : [2000 ; 2500[	570	-93.0***
Ressources (€ / mois) : 2500 ou +	580	-86.5***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 3 dont l'âge est compris entre 75 et 80 ans s'élève à 560 euros. À caractéristiques identiques, ce montant est supérieur de 12,4 euros à celui qui serait notifié à un bénéficiaire de GIR 3 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 3 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base floutée, Drees.

FIGURE 33 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 3 - base CASD

Variables	GIR3 - Montant moyen notifié (en €)	GIR3 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	560	Ref.
Âge : [75 ; 80[	560	20.1***
Âge : [80 ; 85[	580	32.8***
Âge : [85 ; 90[	610	56.1***
Âge : 90 ou +	650	72.7***
Situation : Homme seul	640	165.8***
Situation : Homme en couple	480	Ref.
Situation : Femme seule	660	160.3***
Situation : Femme en couple	550	73.4***
Ressources (€ / mois) : [0 ; 739,8[	640	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	640	-15.5***
Ressources (€ / mois) : [1000 ; 1250[	610	-41.3***
Ressources (€ / mois) : [1250 ; 1500[	600	-58.9***
Ressources (€ / mois) : [1500 ; 2000[	580	-79.6***
Ressources (€ / mois) : [2000 ; 2500[	560	-95.1***
Ressources (€ / mois) : 2500 ou +	570	-97.8***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 3 dont l'âge est compris entre 75 et 80 ans s'élève à 560 euros. À caractéristiques identiques, ce montant est supérieur de 20,1 euros à celui qui serait notifié à un bénéficiaire de GIR 3 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 3 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base CASD, Drees.

FIGURE 34 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 4 - base floutée

Variables	GIR4 - Montant moyen notifié (en €)	GIR4 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	330	Ref.
Âge : [75 ; 80[	320	3.8**
Âge : [80 ; 85[	330	8.9***
Âge : [85 ; 90[	340	19.6***
Âge : 90 ou +	360	34.2***
Situation : Homme seul	360	61.9***
Situation : Homme en couple	290	Ref.
Situation : Femme seule	360	60.4***
Situation : Femme en couple	310	16.0***
Ressources (€ / mois) : [0 ; 739,8[	380	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	360	-12.8***
Ressources (€ / mois) : [1000 ; 1250[	340	-29.8***
Ressources (€ / mois) : [1250 ; 1500[	330	-44.7***
Ressources (€ / mois) : [1500 ; 2000[	320	-53.8***
Ressources (€ / mois) : [2000 ; 2500[	320	-60.3***
Ressources (€ / mois) : 2500 ou +	350	-33.8***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 4 dont l'âge est compris entre 75 et 80 ans s'élève à 320 euros. À caractéristiques identiques, ce montant est supérieur de 3,8 euros à celui qui serait notifié à un bénéficiaire de GIR 4 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 4 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base floutée, Drees.

FIGURE 35 – Montant d'aide humaine notifié selon les caractéristiques des bénéficiaires de GIR 4 - base CASD

Variables	GIR4 - Montant moyen notifié (en €)	GIR4 - Effet à caractéristiques identiques (en €)
Âge : [60 ; 75[	330	Ref.
Âge : [75 ; 80[	320	3.2**
Âge : [80 ; 85[	320	8.6***
Âge : [85 ; 90[	340	20.5***
Âge : 90 ou +	360	36.0***
Situation : Homme seul	360	63.6***
Situation : Homme en couple	290	Ref.
Situation : Femme seule	360	61.0***
Situation : Femme en couple	300	16.6***
Ressources (€ / mois) : [0 ; 739,8[	370	Ref.
Ressources (€ / mois) : [739,8 ; 1000[	360	-14.8***
Ressources (€ / mois) : [1000 ; 1250[	330	-32.4***
Ressources (€ / mois) : [1250 ; 1500[	320	-47.8***
Ressources (€ / mois) : [1500 ; 2000[	320	-56.6***
Ressources (€ / mois) : [2000 ; 2500[	310	-61.6***
Ressources (€ / mois) : 2500 ou +	350	-36.7***

Note > Les montants moyens notifiés d'aide humaine sont calculés sur données pondérées pour être représentatifs de l'ensemble des bénéficiaires de l'APA à domicile au niveau national. Les effets à caractéristiques identiques sont estimés sur données non pondérées à partir d'un modèle de régression censurée (modèle Tobit), où les seuils de censure correspondent aux valeurs des plafonds par GIR. Ces modèles incluent aussi des indicatrices départementales. * p < 0,10, ** p < 0,05, *** p < 0,001.

Lecture > En moyenne, le montant d'aide humaine notifié à un bénéficiaire de GIR 4 dont l'âge est compris entre 75 et 80 ans s'élève à 320 euros. À caractéristiques identiques, ce montant est supérieur de 3,2 euros à celui qui serait notifié à un bénéficiaire de GIR 4 âgé de moins de 75 ans.

Champ > Bénéficiaires de l'APA à domicile de GIR 4 recevant une aide humaine et payés au titre du mois de décembre 2017 (hors valeurs manquantes).
Sources > RI APA 2017, base CASD, Drees.