École nationale de la statistique
et de l'analyse de l'information

Mémoire de Master 2

Année 2014/2015

Amélioration du redressement de la non-réponse des communautés dans le recensement

Lise Reynaert

Membres du jury : Henri Bodet et Guillaume Chauvet

Sous la direction de : Nathalie Caron, Sylvie Rousseau et Frédéric Tallet

Résumé

En France, le recensement repose désormais sur une collecte annuelle qui concerne successivement toutes les communes au cours d'une période de cinq ans. Chaque année, environ 9 millions de personnes sont recensées, toutes catégories de population confondues. Ces catégories distinguent : la population des ménages, celle des habitations mobiles ou des sans-abris et celle des individus des communautés à laquelle s'intéresse ce mémoire.

Le recensement des communautés permet de dénombrer, exhaustivement sur un cycle de cinq ans, la population habitant en communauté et de la caractériser. Le redressement de la non-réponse de la collecte annuelle s'effectue par hot-deck séquentiel ; cette méthode impute chaque valeur manquante par la modalité de la variable prise par le répondant précédent « le plus proche » lorsque l'on parcourt le lot de saisie, les individus étant répartis dans une vingtaine de lots de saisie. Le « plus proche » donneur est choisi parmi les résidents des communautés, selon des contraintes propres à chaque question. Cette méthode limite les temps de traitement et permet ainsi de respecter la forte contrainte de délai à laquelle le recensement est soumis, de par son cadre législatif.

La population en communauté présente toutefois deux particularités qui engendrent quelques cas de redressements aberrants, certes limités mais fâcheux car visibles à un niveau communal : la première particularité est le lien entre le comportement de réponse et l'ordre du lot de saisie, occasionnant des redressements massifs de non-répondants successifs par un unique donneur ; la seconde est l'hétérogénéité des populations en communauté, entraînant le redressement de non-répondants par des donneurs très différents bien que proches dans le lot de saisie.

Dans une vision à court terme, nous proposerons des améliorations à la méthode actuelle, en restant dans la logique des traitements standards du recensement et de ses contraintes de délai : nous envisagerons notamment l'introduction d'un aléa d'imputation dans le hot-deck séquentiel ainsi que l'enrichissement des contraintes d'imputation. Ensuite, nous testerons et comparerons des méthodes de redressement alternatives, qui pourraient être mises en place en cas de refonte à moyen terme des applications du recensement. Enfin, nous conclurons quant à la méthode la plus adaptée aux particularités de la population des communautés et aux contraintes pratiques du recensement.

Abstract

The population census distinguishes several components of population : people residing in ordinary residences, those who live in mobile residences and homeless people, and those living in communities like boarding schools. For the latter, the nonresponse is adjusted by a sequential hotdeck, batch by batch and variable by variable in accordance with a pre-established order of importance. Although this method is time efficient, the population living in communities has two specific features that don't comply with it : the first feature is the strong link between the nonresponse behavior and the order of the file ; the second is the heterogeneity between different communities. We will first put forward improvements of the current method on a short-term vision ; we then compare other nonresponse adjustment methods like random imputation within classes that could be implemented on the long-term.

Mots clés

Recensement, correction de la non-réponse partielle

Remerciements

Je remercie tout particulièrement Nathalie Caron et Sylvie Rousseau qui ont dirigé ce stage de Master. Je les remercie pour l'aide très précieuse qu'elles m'ont apportée : leurs bons conseils, leurs nombreux encouragements et leur disponibilité m'ont été d'une aide inestimable.

Je suis très reconnaissante à Frédéric Tallet pour ses conseils, ses pistes d'amélioration et ses relectures attentives. Je le remercie également beaucoup d'avoir accepté d'assister à ma soutenance.

Je tiens à remercier David Haziza pour son cours de qualité sur le traitement de la non-réponse et pour avoir pris le temps de relire mon mémoire et de me faire part de ses pistes d'amélioration.

Merci également à Sébastien Hallépée d'avoir été disponible et de m'avoir conseillée tout au long de la constitution de ce mémoire. Je lui suis également très reconnaissante pour ses relectures.

Table des matières

Table des figures

Glossaire et abréviations

Communauté
Une communauté est définie comme un ensemble de locaux d'habitation relevant d'une même autorité gestionnaire et dont les habitants partagent à titre habituel un mode de vie commun. Les différentes catégories de communautés sont définies par le décret n° 2003-485 du 5 juin 2003 relatif au recensement de la population.

Classe d'imputation
Pour redresser la non-réponse partielle, les individus sont souvent regroupés dans des classes homogènes selon des critères définis au préalable. Les individus non-répondants d'une même classe sont alors redressés de la même manière. Les groupes ainsi formés sont appelés classes d'imputation.

Groupe de réponses homogènes (GRH)
Comme son nom l'indique, un groupe de réponses homogènes sert à regrouper des individus qui ont le même profil de réponse. Cette notion intervient au moment du redressement de la non-réponse totale par repondération.

Hot-deck
L'imputation par le hot-deck regroupe plusieurs méthodes basées sur le concept de « donneur ». La donnée manquante est remplacée par la valeur observée pour un individu répondant choisi au hasard.

Non-réponse
On distingue traditionnellement deux grands types de non-réponse : la non-réponse totale lorsque l'on obtient aucune réponse ou peu de réponses à l'ensemble du questionnaire relatif à un individu et la non-réponse partielle lorsqu'un individu ne répond pas à une partie plus ou moins importante du questionnaire.

RP
Recensement de la population.

V de Cramer
Le V de Cramer est une mesure d'association entre deux variables qualitatives comprise entre 0 et 1. Elle est basée sur la mesure du χ^2, mais a l'avantage de ne tenir compte ni de la taille de l'échantillon, ni du nombre de modalités des variables.

Variable auxiliaire
Une variable auxiliaire qualifie toute variable qui est connue pour l'ensemble des individus, qu'ils soient répondants ou non-répondants. Dans le cas présent, ce sont soit des variables de géographie, soit des variables provenant du répertoire des communautés.

ZEAT

Zone d'études et d'aménagement du territoire. Chaque ZEAT est composée d'un ensemble d'une ou plusieurs régions administratives.

1 Introduction

En France, le recensement de la population a changé de rythme depuis 2004, le dénombrement général ayant fait place à une méthode rénovée qui permet de produire des informations actualisées chaque année. En pratique, les enquêtes annuelles de recensement sont menées en janvier et février de chaque année. Elles couvrent successivement l'ensemble des communes au cours d'un cycle de cinq ans. Le cumul des cinq enquêtes les plus récentes conduisent aux résultats du recensement, datés du milieu du cycle quinquennal considéré. Ces résultats établis à un niveau communal, et même infracommunal pour certains, sont essentiels au fonctionnement de la société et, à ce titre, sont attendus avec une extrême exigence sur l'exactitude des chiffres à tout échelon géographique.

Le recensement distingue différentes catégories de population : celle des ménages (97,7 %), celle des habitations mobiles ou des sans-abris (0,2 %) et celle des individus des communautés (2,1 %) à laquelle s'intéresse ce mémoire. Une communauté est définie comme un ensemble de locaux d'habitation relevant d'une même autorité gestionnaire et dont les habitants partagent à titre habituel un mode de vie commun.

Les différentes sous-catégories de communautés sont :

- les maisons de retraite et les hospices (32,0 % de la population en communauté au RP2012) ;
- les internats, hors cités universitaires (25,7 %) ;
- les établissements sociaux de moyen ou de long séjour (18,8 %) ;
- les foyers de travailleurs (8,0 %) ;
- les cités universitaires (5,5 %) ;
- les établissements pénitentiaires (3,9 %) ;
- les établissements militaires (3,3 %) ;
- les communautés religieuses (2,0 %) ;
- les établissements sociaux de court séjour (0,5 %) ;
- les autres formes de communauté (0,3 %).

La population vivant en communauté est recensée de manière exhaustive sur un cycle de cinq ans, à raison d'un cinquième par an - soit environ 320 000 individus par an -.

Les opérations post-collecte s'effectuent simultanément pour les individus des communautés et ceux en ménage. A l'issue de la collecte, l'ensemble des questionnaires - environ 9 millions de bulletins - est réparti en une vingtaine de lots de saisie, structurés par commune. Le redressement de la non-réponse s'effectue par hot-deck séquentiel, une méthode qui impute chaque valeur manquante par la modalité de la variable prise par le répondant précédent le plus proche lorsque l'on parcourt le lot de saisie. Pour la non-réponse des communautés, le plus proche donneur est choisi parmi les résidents des communautés (la sienne ou une autre), selon des critères propres à chaque question, comme la tranche d'âge ou le sexe.

Cette méthode est robuste et efficace ; elle permet ainsi de respecter la forte contrainte de délai à laquelle le recensement est soumis, de par la loi relative à la démocratie de proximité du 27

février 2002 fixant la publication des populations légales actualisées en fin décembre de chaque année. La méthode se révèle globalement satisfaisante pour les ménages, dont la non-réponse est diffuse et de faible ampleur (2,3 % de logements non-répondants). Toutefois, elle comporte certaines faiblesses, particulièrement visibles pour les individus des communautés au niveau communal. En raison d'une population moins accessible et de contraintes administratives, le phénomène de non-réponse y est souvent bien plus concentré. Par ailleurs, la population de chaque communauté est assez spécifique et ne correspond pas forcément à celle d'une communauté voisine.

Ces particularités ont pour effet d'amplifier les deux inconvénients majeurs du hot-deck séquentiel :

- La méthode restreint drastiquement le champ des donneurs, ce qui combiné au lien entre le comportement de non-réponse et l'ordre du lot de saisie, provoque des distorsions dans la distribution des variables statistiques après imputation. En effet, dans le cas de la non-réponse en bloc d'une communauté, les réponses du dernier individu répondant d'une structure voisine seront imputées de manière déterministe à tous les non-répondants de la communauté, jusqu'à ce qu'un autre répondant soit rencontré.
- Le modèle d'imputation repose sur l'hypothèse, pas toujours valable, que l'ensemble des réponses d'un individu ainsi que son comportement de réponse sont similaires à ceux de l'individu précédent. Cette spécification du modèle d'imputation est susceptible de provoquer des biais d'imputation.

Ces particularités engendrent quelques cas de redressements aberrants visibles à un niveau communal qui s'avèrent particulièrement regrettables au vu de la portée des résultats du recensement et des attentes qu'ils soulèvent à un niveau fin. Ces redressements maladroits conduisent ainsi à des questions régulières d'utilisateurs. Ils sont d'autant plus gênants que les défauts constatés perdurent dans les résultats de cinq millésimes successifs, de par la méthode même du recensement.

L'objectif de ce stage de Master est de proposer des améliorations de la méthode de redressement afin d'éviter les inconvénients du hot-deck séquentiel actuel. Dans une vision à court terme, nous proposerons des alternatives à la méthode actuelle par hot-deck séquentiel, en restant dans la logique des traitements standards du RP et de ses contraintes de délai. Nous proposerons notamment la redéfinition des contraintes d'imputation et l'ajout d'un aléa d'imputation pour limiter le nombre de réplications. Ensuite, nous considérerons des méthodes de redressement alternatives, comme le hot-deck métrique, le hot-deck par classe et la repondération. Nous conclurons quant à la méthode la plus adaptée aux particularités de la population des communautés et aux contraintes pratiques du recensement.

2 Présentation du recensement

2.1 Introduction

Dans cette deuxième partie, nous présenterons l'environnement professionnel dans lequel s'est effectué ce stage, ainsi que les enjeux du recensement de la population. Nous nous intéresserons ensuite plus particulièrement à l'enquête de recensement dans les communautés, et notamment à la méthode actuelle de redressement de la non-réponse. Enfin, nous décrirons quelques cas de redressements aberrants d'individus en communauté, repérés par des utilisateurs de données communales.

2.2 Environnement professionnel

Ce stage a été effectué à la Division des Méthodes et Traitements des Recensements (DMTR) au sein du Département de la Démographie à la Direction des Statistiques Démographiques et Sociales (DSDS) de la Direction Générale (DG) de l'Insee. Les organigrammes de la DG et de la DSDS figurent en annexe A1 et A2.

La DMTR est responsable des méthodes de sondage et d'estimation en matière de recensement. Elle pilote l'ensemble des traitements qui aboutissent aux calculs des populations légales et à la constitution des bases de données statistiques du recensement servant à la diffusion sur Insee.fr. Par ailleurs, au delà des résultats du recensement obtenus par cumul de cinq collectes, la division met également à disposition des utilisateurs les bases de données relatives à la dernière enquête annuelle de recensement. Cette livraison est réalisée l'année de collecte.

2.3 Les objectifs et les contraintes du recensement

L'objectif du recensement est de déterminer la population légale de chaque collectivité territoriale et de chaque circonscription administrative et de décrire les caractéristiques démographiques et sociales de la population et des logements qu'elle occupe.

Confié à l'Insee par la loi, le dénombrement de la population doit être authentifié chaque fin d'année par décret ; il a donc un caractère officiel et s'impose pour l'application des multiples textes qui utilisent le chiffre de population pour la détermination d'un droit, notamment le montant de la dotation financière aux communes ou le nombre des membres du conseil municipal. Il en résulte de fait une attente extrêmement exigeante sur la qualité des chiffres produits jusqu'à un niveau local fin, que ce soit pour le chiffre de population ou pour la description détaillée des habitants, de leurs logements et familles.

Pour établir les populations légales, l'Insee dispose des informations collectées lors des enquêtes annuelles de recensement et de données non nominatives issues de sources administratives, comme les fichiers de la taxe d'habitation. La détermination des populations légales suppose qu'on ait exploité les données collectées en début d'année. La lourdeur de la collecte, la complexité de l'exploitation et le calendrier de production serré constituent un frein à la mise en place de changements dans les différentes chaînes de production. Il convient en particulier de prendre en compte le fait que les résultats du recensement se basent sur les cinq enquêtes an-

nuelles les plus récentes dont la dernière n'est disponible que tardivement.

Ci-dessous le calendrier des opérations qui suivent une collecte :

- La collecte s'effectue en janvier et février de l'année N.
- La réception en direction régionale et les contrôles de la collecte se déroulent de février à juin de l'année N.
- L'acquisition des données s'effectue de fin mars à septembre de l'année N.
- La codification automatique des libellés et les reprises manuelles très coûteuses en moyens se déroulent de mai à octobre de l'année N.
- Le redressement de la non-réponse s'effectue d'avril à octobre de l'année N.
- L'élaboration des populations légales se déroule d'avril à décembre de l'année N, pour la publication du décret avant la fin de l'année. Le nombre d'individus en communauté est donc fixé à ce moment là.
- Les résultats statistiques sont élaborés de janvier à avril de l'année N+1 et diffusés en juin N+1.

2.4 L'enquête de recensement dans les communautés

Le décret n° 2003-485 du 5 juin 2003 relatif au recensement de la population confie à l'Insee le recensement des personnes vivant dans des communautés. Cette disposition permet de faciliter l'organisation de la collecte. En effet, ce recensement nécessite des accords avec leurs autorités de tutelle (agences régionales de santé, conseils généraux ...) et confier à l'Insee la responsabilité de ces accords est plus efficace que de la disperser sur des milliers de communes.

Les modalités de collecte et le répertoire des communautés

La population vivant en communauté est recensée de manière exhaustive tous les cinq ans, à raison d'un cinquième par an - soit environ 320 000 individus par an -. Le nombre de communautés à recenser, une année donnée, est de l'ordre de 7 000 communautés. La charge de collecte est très variable d'une direction régionale de l'Insee à l'autre : près de 700 communautés sont recensées chaque année en Rhône-Alpes, une trentaine seulement en Corse.

Le répertoire des communautés recense toutes les communautés et maintient à jour les informations qui leurs sont associées. Il a été constitué à partir des informations du recensement exhaustif de 1999 et est mis à jour en continu grâce à des sources administratives (par exemple, le fichier FINESS des établissements sanitaires et sociaux) et des retours terrains. Chaque année, la liste des communautés présentes dans le répertoire est adressée pour expertise aux communes concernées l'année suivante par le recensement des communautés. Le répertoire renseigne en particulier la sous-catégorie de chaque structure ainsi que sa capacité d'accueil théorique.

La collecte des communautés s'effectue par dépôt-retrait de questionnaires en version papier sur une période de quatre semaines en janvier. Les principaux acteurs de cette collecte sont :

- Le pôle national « recensement des communautés » situé à la direction régionale de l'Insee de Haute-Normandie a la responsabilité complète de l'opération de recensement dans les communautés.
- L'enquêteur de l'Insee qui effectue la collecte ;
- La commune qui expertise la liste de ses communautés à collecter.

En cas de difficulté lors de la collecte, l'enquêteur a pour consigne de renseigner un bulletin sans omission ni double-compte pour chaque membre de la communauté, et - si possible - de renseigner le sexe et la date de naissance pour faciliter les redressements.

Habituellement, et sauf cas particulier, la collecte est réalisée dans toutes les communautés, même incomplète ou à partir d'une liste obtenue auprès du responsable de la structure. Toutefois, les difficultés de collecte se sont accentuées lors de la collecte 2015 manifestement suite à la mise en place des nouvelles conditions d'emploi des enquêteurs de l'Insee. En effet, la collecte s'est soldée par dix cas de communautés en non-réponse totale. Pour ces cas, des bulletins individuels sont créés en direction régionale à partir des instructions de la DMTR qui utilise les informations de la collecte précédente - si elle est disponible - ou d'une autre source, comme la capacité théorique issue du répertoire des communautés ou la collecte d'une communauté similaire.

La méthode actuelle de redressement de la non-réponse

Les opérations post-collecte s'effectuent simultanément pour les individus recensés des communautés et ceux en logements « ordinaires ». Chaque année, à l'issue de la collecte, l'ensemble des questionnaires - environ 9 millions de bulletins - revient à l'Insee. Ils sont répartis en une vingtaine de lots de saisie ; ces lots, structurés par commune, comprennent à la fois des bulletins d'individus en communauté et des bulletins d'individus en logements « ordinaires ». Les questionnaires sont numérisés par lecture optique, lot par lot, par un prestataire extérieur.

Le redressement de la non-réponse s'effectue par hot-deck séquentiel déterministe, une méthode d'imputation robuste et rapide qui permet de respecter les fortes contraintes de délai du recensement.

Pour chaque lot de saisie, le hot-deck séquentiel procède de la manière suivante :

- Les individus sont triés selon l'identifiant du recensement (département, commune, rang d'adresse et rang d'individu).
- Pour chaque valeur manquante, la valeur du répondant le plus proche lorsqu'on parcourt le lot de saisie est imputée. Pour la non-réponse des communautés, le plus proche donneur est choisi parmi les résidents des communautés (la sienne ou une autre), selon des critères propres à chaque question, comme la tranche d'âge ou le sexe (cf. tableau 1). Des valeurs ont été initialisées au cas où la première observation du lot de saisie est défaillante.
- Le redressement s'effectue variable par variable, dans un ordre logique prédéfini (cf. tableau 2). La première variable est imputée en utilisant potentiellement les variables auxiliaires, puis la variable imputée est utilisée pour imputer les variables suivantes et ainsi de suite.

L'Insee procède ensuite au codage des questions relatives à la profession et à l'activité ainsi qu'aux traitements statistiques nécessaires pour obtenir un fichier de données individuelles anonymes propre pour préparer la diffusion des résultats. Une fois ces traitements effectués, l'enquête annuelle participe à établir les populations légales et les résultats statistiques de cinq millésimes successifs.

Tableau 1 – Les critères d'imputation de la méthode actuelle par hot-deck séquentiel

Variables à imputer	Contraintes d'imputation de la méthode actuelle
Sexe	
Âge	
État matrimonial	Sexe
	Tranche d'âge
Vie en couple	Tranche d'âge
	État matrimonial
	Sexe
Indicatrice de nationalité	Pays de naissance
Inscription dans un étab. d'enseignement	Tranche d'âge
Situation Principale	Indicatrice « travaille actuellement »
	Sexe
	Tranche d'âge
Diplôme	Tranche d'âge
	Indicatrice de nationalité

Les exploitations statistiques principale et complémentaire

Les exploitations statistiques des questionnaires s'effectuent en deux temps de manière à mettre à disposition, le plus rapidement possible, l'essentiel des résultats.

L'exploitation « principale » porte sur l'ensemble des questionnaires collectés. Elle traite toutes les informations pouvant être codifiées aisément après la saisie des questionnaires. Les résultats statistiques issus de cette exploitation couvrent la plupart des critères d'étude permis par les questionnaires du recensement : âge, sexe, nationalité, pays de naissance, état matrimonial, scolarisation, diplôme, lieu de résidence un an plus tôt, etc.

La seconde exploitation statistique dite « complémentaire », plus complexe et lourde, se base sur un échantillon des questionnaires collectés. Elle permet d'affiner les caractéristiques de l'emploi et de procéder à l'analyse détaillée de la composition des ménages et des familles. Le profil des résidents des communautés est particulier sur les thèmes de l'exploitation complémentaire : d'une part, ils sont moins concernés que les ménages sur le thème « emploi » (notamment les élèves en internat ou les personnes âgées en maison de retraite) ; d'autre part, ils ne sont pas concernés par le thème « famille », qui s'applique uniquement aux ménages des logements ordinaires.

La diffusion des résultats statistiques

Les résultats statistiques du recensement sont diffusés gratuitement sur le site Insee.fr et mises à jour tous les ans. Le dispositif de diffusion des résultats a été conçu pour répondre à la demande

Tableau 2 – Ordre dans lequel les principales variables sont redressées

1	Sexe
2	Âge
3	État matrimonial
4	Indicatrice vie en couple
5	Nationalité
6	Indicatrice travaille actuellement
7	Inscription dans un établissement d'étude
8	Situation principale
9	Indicateur de résidence antérieure
10	Diplôme

de publics variés, aussi les résultats se présentent-ils sous plusieurs formes : pour le grand public, des données sont directement accessibles, sous une forme conviviale ; pour les spécialistes et les professionnels, des bases de données peuvent être téléchargées, nécessitant ensuite des manipulations par l'utilisateur.

Ci-dessous, les principaux supports d'information disponibles sur Insee.fr, par gamme croissante :

- Les chiffres clés par commune, à destination du grand public, fournissent les chiffres les plus importants pour toutes les communes.
- Les tableaux détaillés par commune, également à destination du grand public, complètent les chiffres clés sur toute zone de plus de 2 000 habitants. Ces tableaux portent sur tous les thèmes du recensement : sexe et âge, nationalité, immigration, diplômes, scolarité, mobilité résidentielle, situation vis-à-vis de l'emploi, profession, secteur d'activité, type d'emploi, lieu de travail, composition du ménage, taille et confort du logement.
- Les bases de données à un niveau communal et infracommunal reprennent les données des tableaux détaillés et permettent toutes les agrégations géographiques. Pour les grandes villes, des bases indicateurs au niveau des quartiers dits « Iris » sont également disponibles.
- Les fichiers détails anonymisés pondérés sont disponibles pour les utilisateurs les plus avertis et leur permettent des tabulations libres.

Ces supports répondent aux besoins de la majorité du public. Toutefois, pour les demandes particulières, l'Insee propose, contre paiement, des tabulations sur mesure. Par ailleurs, pour les organismes ayant une mission de service public (collectivités territoriales, services de l'État, etc.), l'Insee propose également un service payant de diffusion sur des zones à façon (DIAF-RP) dont les contours sont tracés par l'utilisateur.

2.5 Les redressements aberrants dans les communautés

Certains utilisateurs des données communales, comme les agences d'urbanismes, les mairies ou les conseils généraux, ont fait remonter à l'Insee quelques résultats aberrants visibles à un niveau communal ; il s'avère que ces cas proviennent du redressement maladroit de la non-réponse d'individus en communauté par la méthode par hot-deck séquentiel. Même s'ils sont peu nombreux, ces cas sont difficilement justifiables auprès des utilisateurs et s'avèrent regrettables au vu de

la portée des résultats du recensement et des attentes qu'ils soulèvent à un niveau communal.

Ces défauts sont d'autant plus problématiques qu'ils perdurent dans les résultats de cinq millésimes successifs, de par la méthode même du recensement. Il est très coûteux de corriger ces défauts dans la mesure où cela implique de modifier les résultats archivés de plusieurs millésimes antérieurs.

Ci-dessous quatre exemples de cas problématiques, parmi la dizaine de cas majeurs identifiés dans les enquêtes 2009 à 2013 par des agences d'urbanisme et des mairies :

- Cas des cadres nonagénaires dans une commune de plus de 10 000 habitants des Hauts-de-Seine : lors de la collecte 2009, une vingtaine d'hommes âgés de plus de 90 ans en maison de retraite qui avaient uniquement renseigné la date de naissance et le sexe ont été redressés par un cadre habitant en foyer de travailleurs de même sexe (masculin) et de même tranche d'âge (65 ans et plus), créant ainsi abusivement des individus cadres de plus de 90 ans. Ce cas a été repéré dans les fichiers détails anonymisés.

- Cas des déplacements domicile-travail dans une commune de moins de 2 000 habitants dans la Manche, entre deux communes distantes de 150 km : lors de l'enquête 2011, une centaine d'individus de trois communautés d'une même commune ont été redressés à partir d'un donneur travaillant dans une autre commune distante de 150 km, créant ainsi des déplacements domicile-travail entre deux communes lointaines.

- Cas des octogénaires dans une commune de moins de 5 000 habitants des Yvelines : lors de l'enquête 2010, une centaine d'individus en foyer de demandeurs d'asile n'ayant pas donné leur âge ont été redressés par un donneur unique de 86 ans en maison de retraite, déformant ainsi la pyramide des âges de la commune.

- Cas des octogénaires dans une commune de plus de 10 000 habitants du Val-de-Marne : lors de l'enquête 2009, 165 individus non-répondants d'un foyer de travailleurs n'ayant pas donné leur âge ont été redressés à partir d'un individu en maison de retraite né en 1920, déformant ainsi la pyramide des âges de la commune.

2.6 Conclusion

Les deux principaux objectifs du recensement sont d'une part de déterminer chaque année la population légale de chaque commune et d'autre part de décrire les caractéristiques de la population et des logements qu'elle occupe.

Les populations légales ont un caractère officiel et s'imposent pour l'application de nombreux textes de loi. Elles conditionnent notamment le montant de la dotation financière aux communes et le nombre des membres du conseil municipal. Cette utilisation constitue une très grande exigence sur l'exactitude des chiffres produits jusqu'à un niveau local fin. Par ailleurs, la lourdeur du processus de production combinée à la forte contrainte de délai influence les choix des méthodes de traitement utilisées.

L'enquête de recensement dans les communautés est indispensable pour établir les populations légales et les résultats statistiques diffusés à un niveau communal et infracommunal. Cette enquête, confiée à l'Insee, est exhaustive et la collecte est répartie sur cinq ans, à raison d'un

cinquième par an. Chaque année, 320 000 individus en communauté sont ainsi enquêtés par l'Insee.

Le redressement de la non-réponse s'effectue par hot-deck séquentiel, une méthode qui impute chaque valeur manquante par la modalité de la variable prise par le répondant précédent le plus proche lorsqu'on parcourt le lot de saisie. Même si elle a l'avantage d'être très efficace et robuste, cette méthode provoque des redressements maladroits à un niveau communal particulièrement regrettables au vu des attentes que les résultats du recensement soulèvent à un niveau fin.

3 Éléments préalables à l'étude des redressements

3.1 Introduction

Dans cette étape préalable aux redressements de la non-réponse, nous décrirons de façon synthétique la population résidant en communauté et nous mettrons en évidence les particularités de leur comportement de non-réponse. Nous ferons également le point sur les variables auxiliaires permettant de modéliser à la fois le comportement de non-réponse, mais aussi les réponses données au questionnaire.

Cette étape est importante dans la mesure où elle nous aidera à déterminer par la suite le traitement le plus adapté aux particularités de la population en communauté.

3.2 Champ de l'étude

Nous utilisons les réponses des 1 604 055 individus en communauté collectées entre 2009 et 2013. Cette période correspond à un cycle quinquennal complet, pendant lequel les communautés ont été enquêtées de manière exhaustive. Cette analyse est menée spécifiquement afin d'obtenir des estimations robustes. Les résultats, obtenus par simulation sur les cinq enquêtes cumulées, diffèrent ainsi des redressements actuels en production, effectués indépendamment par année de collecte.

Nous limitons cette étude aux dix variables figurant sur la première page du questionnaire, à savoir :

- Pour l'ensemble des individus : l'âge, l'année de naissance et la nationalité ;
- Pour l'ensemble des individus hors détenus [1] : l'inscription dans un établissement d'étude et le lieu d'habitation au 1er janvier ;
- Pour les individus de plus de 14 ans : l'état matrimonial, l'indicatrice vie en couple et le diplôme ;
- Pour les individus de plus de 14 ans hors détenus[1] : la situation principale et l'indicatrice du fait de travailler actuellement.

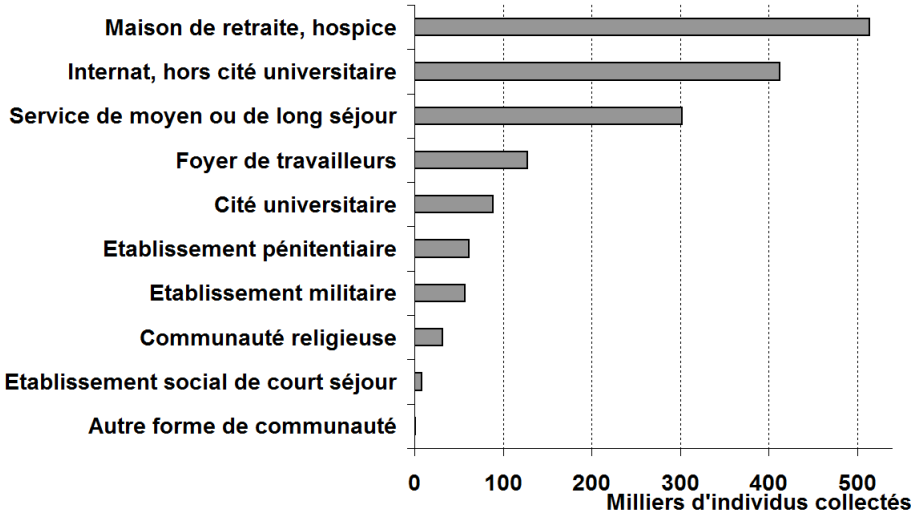
L'ensemble des variables du questionnaire et leurs modalités figurent dans les annexes A3 et A4.

3.3 Quelques éléments de description

La population habitant en communauté a des profils très variés. Elle comprend majoritairement les personnes âgées vivant en maison de retraite (514 000 personnes recensées entre 2009 et 2013, cf. tableau 3), les élèves et étudiants hébergés en internat (413 000 personnes) ou en cité universitaire (89 000 personnes) et les personnes hébergées dans un autre établissement sanitaire ou social de moyen ou long séjour (302 000 personnes). S'ajoutent 282 000 personnes vivant dans des communautés de types très divers : foyers de travailleurs, prisons, couvents...

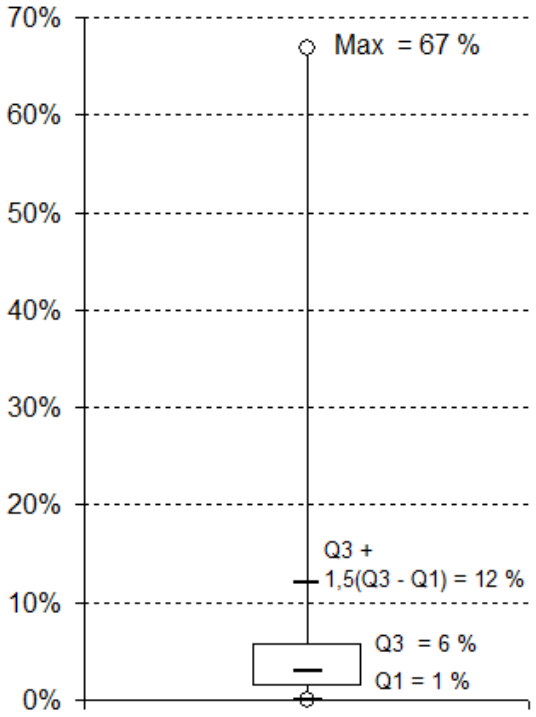
1. Ces informations ne sont pas collectées dans les centres de détention.

Tableau 3 – Répartition des individus collectés entre 2009 et 2013 par type de communauté



La population habitant en communauté est susceptible de représenter une part très importante de la population des communes (cf. tableau 4), allant jusque 67 % au RP2012 pour la commune de Taizé (code commune : 71532, population municipale au RP2012 : 181 habitants). Même si au niveau France entière, la population en communauté ne représente que 2,5 % de la population municipale, les communautés peuvent ainsi avoir un impact très visible dans les populations légales ou les résultats statistiques diffusés au niveau de certaines communes.

Tableau 4 – Part (%) de la population en communauté dans la population municipale communale



Les communautés peuvent être de taille très importante : lors des collectes 2009 à 2013, la plus grande communauté est la maison d'arrêt de Fleury-Mérogis en Ile-de-France avec 3 700 individus collectés en 2013 (cf. tableau 5). Cette communauté représente 40 % de la population municipale de la commune au RP2012.

Tableau 5 – Distribution de la taille des communautés en nombre d'individus collectés

Maximum	3 690
99 %	400
95 %	180
90 %	120
Q3 75 %	80
Médiane 50 %	40
Q1 25 %	15
Minimum	1

3.4 Particularité de la non-réponse dans les communautés

La première particularité de la non-réponse dans les communautés est le faible taux de non-réponse pour les variables « sexe » et « âge » (respectivement 0,4 % et 2 %) ; la non-réponse totale est donc quasi-inexistante (cf. tableau 6). Cela s'explique par les consignes de collecte évoquées précédemment : en cas de difficulté, l'enquêteur a pour consigne de renseigner un bulletin sans omission ni double-compte pour chaque membre de la communauté, et - si possible - de renseigner le sexe et la date de naissance pour faciliter les redressements et l'estimation de la pyramide des âges. En pratique, l'enquêteur obtient en général auprès du responsable de la structure une liste comportant a minima les nom, âge et sexe de l'ensemble des résidents. En revanche, la non-réponse pour les autres variables du questionnaire est nettement plus élevée dans les communautés que dans les ménages ; par exemple, la situation principale [2] n'est pas renseignée pour 28,9 % des individus en communauté, contre seulement 4,9 % des individus en ménage.

La deuxième particularité de la non-réponse dans les communautés est que la non-réponse est concentrée et corrélée entre les individus au sein d'une même communauté ; cela s'explique par un phénomène de non-réponse massive de communautés entières, de par des caractéristiques sociales intra-communautés similaires et du processus de collecte qui peut conduire à des réponses renseignées collectivement. Ainsi la probabilité empirique de non-réponse sachant que l'individu précédent selon l'identifiant du recensement (département, commune, rang d'adresse, rang d'individus) est non-répondant avoisine 0,9 pour l'ensemble des variables du questionnaire, sauf pour la variable sexe (cf. tableau 7).

2. Les modalités de la variable « situation principale » sont : emploi, apprentissage, études, chômage, retraite, homme ou femme au foyer ou autre situation.

Tableau 6 – Taux de non-réponse pour les variables d'intérêt en distinguant les individus en communauté des individus en ménage

Variables	Taux de non-réponse*** (en %)	
	Communautés	Ménages
Sexe	0,4 %	2,0 %
Année de naissance	2,0 %	2,3 %
Nationalité	17,3 %	3,8 %
Inscription dans un établissement d'étude	35,5 %	11,2 %
Vie en couple**	27,0 %	6,6 %
État matrimonial*	26,1 %	7,1 %
Diplôme*	53,0 %	10,4 %
Situation principale**	28,9 %	6,8 %
Indicatrice « travaille actuellement** »	39,7 %	13,0 %

* Champ : individus de plus de 14 ans
** Champ : individus de plus de 14 ans hors détenus
*** Champ : enquêtes 2009 à 2013

Tableau 7 – Probabilité empirique de non-réponse sachant que l'individu précédent est non-répondant pour la variable concernée

Variables	Probabilité empirique de non-réponse***	
	Communautés	Ménages
Sexe	0,27	0,14
Année de naissance	0,83	0,47
Nationalité	0,82	0,3
Inscription dans un établissement d'étude	0,92	0,39
Indicateur de résidence antérieure	0,87	0,26
Vie en couple**	0,87	0,08
État matrimonial*	0,84	0,10
Diplôme*	0,91	0,11
Situation principale**	0,88	0,08
Indicatrice « travaille actuellement** »	0,87	0,38

* Champ : individus de plus de 14 ans
** Champ : individus de plus de 14 ans hors détenus
*** Champ : enquêtes 2009 à 2013

3.5 Les variables auxiliaires disponibles

Les variables auxiliaires nous permettront de modéliser le comportement de non-réponse. Pour chaque individu, qu'il soit répondant ou non-répondant, nous connaissons :

- sa commune, son département et sa région/sa zone d'études et d'aménagement du territoire (ZEAT) ;
- la tranche d'aire urbaine 2010 de sa commune et la catégorie de cette aire urbaine ;

- la sous-catégorie de sa communauté [3] ;
- l'effectif de sa communauté découpé en 6 tranches [4] ;
- le fait que sa communauté soit enquêtée ou non pour la première fois.

Les modalités de ces variables auxiliaires se trouvent en annexe A5.

Il est important de souligner le pouvoir explicatif de la sous-catégorie de communauté, à la fois sur les réponses au questionnaire, mais aussi sur le comportement de non-réponse. Ce lien peut être aisément mis en évidence grâce à un test d'indépendance du χ^2. Pour toutes les variables du questionnaire, on rejette l'hypothèse nulle d'indépendance entre la sous-catégorie de communauté et le comportement de non-réponse (cf. tableau 8 pour la variable « nationalité » à titre d'exemple). La non-réponse est notamment très présente dans les foyers de travailleurs, les cités universitaires et les établissements sociaux de court séjour.

3.6 Conclusion

La population en communauté représente une part importante de la population de certaines communes et impacte donc fortement les populations légales et les résultats statistiques qui sont diffusés au niveau communal et infracommunal. Par ailleurs, les communautés sont susceptibles d'être de taille très importante, allant jusqu'à 3 700 individus pour les collectes 2009 à 2013.

La population en communauté possède plusieurs particularités. D'une part, elle a des profils très variés : elle comprend notamment les personnes âgées en maison de retraite, les élèves ou étudiants hébergés en internat ou en cité universitaire et les personnes hébergées dans un autre établissement sanitaire ou social de moyen ou long séjour. S'ajoutent les personnes vivant dans des communautés très diverses telles que les prisons, les couvents ou les casernes. D'autre part, la non-réponse dans les communautés est nettement plus élevée et concentrée que dans les ménages, avec un phénomène de non-réponse massive de communautés entières qui est amené à s'intensifier.

Les variables auxiliaires nous permettant de modéliser le comportement de non-réponse sont :
- la commune, le département et la région/la ZEAT ;
- la tranche et la catégorie d'aire urbaine 2010 de sa commune ;
- la sous-catégorie de sa communauté ;
- l'effectif de sa communauté découpé en 6 tranches ;
- le fait que sa communauté soit enquêtée ou non pour la première fois.

3. Les modalités de la variable sous-catégorie de la communauté sont : maison de retraite-hospice, foyer Adoma, autre foyer de travailleurs, service de moyen ou de long séjour, communauté religieuse, gendarmerie, quartier-base-camp militaire, cité universitaire, autre internat, établissement pénitentiaire, établissement social de court séjour, autre communauté

4. Cet effectif correspond au nombre de bulletins individuels :
   - soit collectés ;
   - soit renseignés à partir d'une liste obtenue auprès du responsable de la communauté ;
   - soit obtenus à partir d'une autre source (capacité théorique, collecte précédente etc.) pour les cas de non-réponse totale.

Tableau 8 – Dépendance entre la sous-catégorie de communauté et le fait de répondre à la question « nationalité »

Sous-catégorie de communauté	Effectif de population	Écart des effectifs observés par rapport à la situation hypothétique d'indépendance	
		Réponse	Non-réponse
Maison de retraite, hospice	513 994	2 %	-10 %
Foyer Adoma	55 901	-35 %	168 %
Autres foyers de travailleurs	71 554	-18 %	84 %
Services de moyen ou de long séjour	301 777	-1 %	6 %
Communauté religieuse	31 751	14 %	-68 %
Gendarmerie	3 755	16 %	-78 %
Quartier, base ou camp militaire	53 066	-7 %	36 %
Cité universitaire	88 589	-30 %	145 %
Autre internat	412 792	12 %	-60 %
Établissement pénitentiaire	61 952	1 %	-7 %
Établissement social de court séjour	7 762	-23 %	112 %
Autre communauté	763	-11 %	51 %

Effectif de la population : 1,6 million d'individus
χ^2 : 121 625
p-value $<$ 0,0001

La variable sous-catégorie de communauté a notamment un fort pouvoir explicatif, à la fois sur les réponses des individus et aussi sur leur comportement de non-réponse.

4 Mise en place de critères de qualité

4.1 Introduction

Avant de tester les méthodes de redressement de la non-réponse, il convient de préciser ce que nous attendons d'elles. Ces « qualités » nous permettront de comparer facilement les performances des différentes méthodes que nous allons mettre en œuvre, à savoir le hot-deck séquentiel, le hot-deck métrique, le hot-deck par classe et la repondération. Dans cette quatrième partie, nous allons définir les critères permettant d'appréhender la qualité des différentes méthodes de redressement que nous allons tester dans la suite de l'étude.

4.2 Définition des critères de qualité

Le principal objectif du redressement de la non-réponse est de corriger le biais induit par le fait que les non-répondants n'ont pas les mêmes caractéristiques que les répondants. L'enquête de recensement dans les communautés étant la seule source d'information sur cette population particulière, nous ne pouvons malheureusement pas estimer facilement l'efficacité des différents traitements pour corriger ce biais.

A défaut de comparer les réductions de biais des méthodes de redressement, nous avons défini les quatre critères empiriques suivants, par ordre d'importance :

- **Critère n° 1 :** La première qualité que nous attendons des méthodes testées est qu'elles n'engendrent pas de redressements aberrants visibles à un niveau communal et difficiles à justifier auprès des utilisateurs. En pratique, nous nous assurerons donc que ces méthodes résolvent les cas de redressements aberrants décrites dans la deuxième partie du mémoire, c'est-à-dire que les redressements donnent des caractéristiques imputées plus adaptées à la situation des non-répondants (par exemple, que des personnes âgées résidant en maison de retraite n'aient pas été redressées par des personnes en foyer de travailleur).

- **Critère n° 2 :** La deuxième qualité que nous attendons de la méthode de redressement est qu'elle ne déforme pas les distributions des variables à un niveau communal, et ce, même dans le cas de non-réponse massive de communautés entières. Pour ce critère, nous allons mesurer et comparer la distorsion induite par les différentes méthodes sur les variables tranche d'âge, état matrimonial, situation principale et diplôme en cas non-réponse sur des communautés entières.

  En pratique, nous générerons 50 fois une non-réponse massive dans 50 % des communautés de 100 communes « test », très différentes du point de vue de la population en communauté de par leur taille, leur situation géographique et leur type de communauté. Les communes « test » se trouvent en annexe A6.

  Pour quantifier la distorsion provoquée par le redressement sur nos quatre variables qualitatives, nous définissons une mesure correspondant à l'écart moyen par rapport à la distribution de la variable avant génération aléatoire de la non-réponse.

Soit une variable qualitative comportant J modalités, on note :
- T le nombre de communes « test » (T = 100) ;
- J le nombre de modalités de la variable d'intérêt ;
- $F_{tj\_repondants}$ la proportion de la modalité j parmi les répondants de la commune t ;
- $F_{tj\_imputation}$ la proportion de la modalité j parmi les répondants de la commune t après génération aléatoire de la non-réponse puis redressement par imputation.

On définit la distorsion D_t engendrée sur la commune test t comme l'écart moyen par rapport à la distribution de la variable avant imputation :

$$D_t = \frac{1}{J} \sum_{j=1}^{J} |F_{tj\_imputation} - F_{tj\_repondants}|$$

La mesure de distorsion D correspond à la moyenne de la distorsion observée sur l'ensemble des communes « test » :

$$D = \frac{1}{T} \sum_{t=1}^{T} D_t$$

Cette mesure nous servira à comparer les différentes méthodes, sachant que plus elle est petite et meilleure sera considérée la méthode. Même si nous ne l'avons pas mis en œuvre, il aurait été intéressant d'étudier la volatilité des D_t : les distorsions communales extrêmes auraient permis de détecter des redressements aberrants.

A noter également que le nombre de 50 itérations, qui peut sembler faible, a été limité par des contraintes de temps de traitement lourds. Toutefois, les résultats se montrent stables d'une itération à l'autre.

- **Critère n° 3 :** Nous souhaitons également éviter le redressement massif de communautés entières non-répondantes par un unique donneur. Nous recherchons donc que la méthode de redressement limite le nombre maximal de dons par individu répondant.

  Pour ce critère, aucune non-réponse supplémentaire n'est introduite dans les données, les données réelles de la base sont exploitées directement pour le choix des donneurs sur les non-réponses effectivement observées. Nous comparerons pour ce critère le nombre maximal de dons par individu, ainsi que les 99ème et 90ème percentiles, sachant que plus ces statistiques sont petites et meilleure sera considérée la méthode.

- **Critère n° 4 :** Nous souhaitons également que la méthode de redressement ait une bonne qualité prédictive en cas de non-réponse de faible ampleur et diffuse, c'est-à-dire répartie aléatoirement dans l'ensemble du lot de saisie.

  Pour évaluer la qualité prédictive, nous simulerons à multiples reprises une non-réponse diffuse dans le champ des répondants et nous analyserons le taux moyen de bien classés. En pratique, la variable d'intérêt est mise à blanc pour 5 % des répondants (non-réponse de faible ampleur) suivant un tirage aléatoire indépendant des caractéristiques des répondants (non-réponse diffuse), puis le redressement de la non-réponse est réalisé. Enfin, les variables d'intérêt avant et après l'imputation sont comparées. Nous calculons un taux

moyen de bien classés : il s'agit du pourcentage d'individus bien classés, c'est-à-dire ceux ayant la même modalité pour les variables observées et estimées. Cette opération est réalisée 50 fois, avec des résultats très stables d'une itération à l'autre.

4.3 Conclusion

nous avons défini quatre critères qui reflètent nos attentes quant à la qualité de la méthode de redressement :

- Critère n° 1 : Nous souhaitons tout d'abord que la méthode de redressement résolve les cas aberrants de la méthode actuelle.
- Critère n° 2 : La méthode doit également réduire la distorsion de la distribution des variables à un niveau communal en cas de non-réponse massive et concentrée.
- Critère n° 3 : Nous souhaitons réduire le nombre maximal de don par individu répondant.
- Critère n° 4 : Enfin, la méthode doit également avoir une bonne qualité prédictive en cas de non-réponse diffuse.

Toutefois, il aurait été intéressant de définir des critères supplémentaires. Par exemple, il aurait été souhaitable que la méthode de redressement minimise les distorsions communales extrêmes qui sont révélatrices de redressements aberrants.

Dans la suite du mémoire, nous appréhenderons la qualité des méthodes de redressement de la non-réponse à la lumière de ces critères.

5 Amélioration de la méthode actuelle par hot-deck séquentiel

5.1 Introduction

Dans cette cinquième partie, nous proposerons des améliorations à la méthode actuelle par hot-deck séquentiel en restant dans la logique des traitements standards du RP et en respectant ses fortes contraintes de délai. Nous nous inscrivons ainsi dans une perspective de mise en œuvre à court terme des propositions d'amélioration des redressements dans les chaînes RP.

5.2 Première proposition : regrouper les communautés dans un unique lot de saisie

A l'issue de la collecte, l'ensemble des questionnaires est réparti en une vingtaine de lots de saisie ; ces lots, structurés par commune, comprennent à la fois des bulletins d'individus en communauté et des bulletins d'individus en ménage. Une première proposition d'amélioration de la méthode actuelle consiste à regrouper l'ensemble des individus en communauté dans un unique lot de saisie ; cela permettrait d'améliorer la ressemblance entre les individus qui se suivent dans le lot.

Le fait que les individus d'un même département ou d'une même région soient potentiellement éclatés dans plusieurs lots de saisie est problématique. En effet, deux individus éloignés géographiquement et se situant à la suite dans le même lot de saisie sont susceptibles d'être très différents au regard de variables liées au lieu d'habitat ; l'hypothèse de base du hot-deck séquentiel n'est donc pas respectée. Nous pouvons illustrer ce problème par le cas d'un redressement lors de l'enquête 2008 - hors champ de l'étude - d'un foyer entier de travailleurs non-répondants à Boulogne-Billancourt par un donneur moine-agriculteur vivant dans une communauté religieuse du Vaucluse. Ce dernier précédait directement le foyer de travailleurs dans le lot de saisie.

Résultats du regroupement dans un unique lot de saisie

Nous évaluons le résultat de cette première proposition à la lumière de nos critères de qualité.

Critère n° 1 (cas des redressements aberrants) : pour les quatre cas de redressements aberrants considérés, le fait de regrouper les individus dans un seul lot de saisie permet de résoudre le cas des déplacements domicile-travail dans la Manche, car les non-répondants ont été redressés par des individus plus proches géographiquement. En revanche, cela n'améliore pas les trois autres cas de redressements aberrants où la dimension géographique importe peu.

Critère n° 2 (simulation d'une non-réponse massive et concentrée dans les communes « test ») : le regroupement des individus dans un unique lot de saisie permet de diminuer légèrement la distorsion de la distribution des variables (cf. tableau 9).

Critère n° 3 (nombre maximal de dons) : cette proposition ne permet pas de limiter le nombre de dons par individu. Cela s'explique par le fait que nous n'ayons pas introduit d'aléa dans le hot-deck.

Tableau 9 – Mesure de distorsion du hot-deck séquentiel en regroupant les individus dans un unique lot de saisie

Variables à imputer	Hot-deck séquentiel - **unique** lot de saisie	Hot-deck séquentiel - **plusieurs** lots de saisie (méthode actuelle)
Âge	14,9	**15,2**
État matrimonial	14,4	**15,5**
Situation principale	13,1	**14,0**
Diplôme	10,6	**11,3**

Critère n° 4 (simulation diffuse de non-réponse) : le taux de bien classés est identique à celui de la méthode actuelle. Cela s'explique par le fait que la non-réponse étant diffuse, les individus ont été majoritairement redressés par des donneurs provenant de la même communauté dans les deux méthodes (cf. tableau 10).

Tableau 10 – Taux moyen de bien classés du hot-deck séquentiel en regroupant les individus dans un unique lot de saisie

	Taux de bien classés	
Variables à imputer	Hot-deck séquentiel - **unique** lot de saisie	Hot-deck séquentiel - **plusieurs** lots de saisie (méthode actuelle)
Sexe	69,2 %	**69,1 %**
Âge	74,9 %	**74,7 %**
État matrimonial	72,7 %	**72,7 %**
Vie en couple	91,7 %	**91,7 %**
Inscription dans un étab. enseis.	95,2 %	**95,2 %**
Situation principale	86,1 %	**86,2 %**
Indicateur résidence antérieure	67,4 %	**67,2** %
Diplôme	50,0 %	**49,7 %**

Cette première proposition permet donc de résoudre certains cas de redressements aberrants, particulièrement lorsqu'ils mettent en jeu des informations de nature géographique. Par ailleurs, en regroupant l'ensemble des individus, on diminue les chances que le premier individu du fichier soit défaillant et qu'il soit redressé par une valeur initiale déterminée a priori.

5.3 Deuxième proposition : ajout de nouvelles contraintes d'imputation

Une deuxième amélioration de la méthode actuelle consiste à redéfinir les contraintes d'imputation, à défaut de pouvoir modifier l'ordre de tri du fichier pour chaque variable à imputer. Cette proposition a été testée en regroupant l'ensemble des individus dans un unique lot de saisie.

Sélection des contraintes d'imputation

Pour cela, il faut tout d'abord déterminer les variables qui « expliquent » le mieux les variables à imputer parmi toutes celles qui apportent de l'information sur l'ensemble des répondants et des non-répondants (c'est-à-dire les variables auxiliaires et les variables renseignées du questionnaire). On utilise pour ce faire les observations correspondant aux répondants et un modèle polytomique non ordonné.

Soit n le nombre d'individus, répartis en J catégories distinctes (i.e. les modalités de la variable à expliquer). Chaque individu i appartient à une catégorie j parmi les J modalités de la variable à expliquer. Il est décrit par un ensemble de K caractéristiques x_{i1}, x_{i2},..., x_{ik} (par exemple, la sous-catégorie de sa communauté, sa tranche d'âge, son sexe, son diplôme, etc).
La probabilité d'observer, pour l'individu i, compte tenu de ses caractéristiques x_{ik}, la modalité j de la variable à expliquer s'écrit :

$$P(j/x_i) = \frac{exp(x_i\beta_j)}{\sum_{h=1}^{J} exp(x_i\beta_h)} \qquad pour j = 1, 2, ..., J$$

où J est le nombre de modalités de la variable à expliquer, les x_i sont les caractéristiques de l'individu i prises en compte dans le modèle et les β_i et β_j les paramètres estimés.

Or ce modèle n'est pas identifiable, car il possède un nombre trop élevé de paramètres. En effet, supposons que l'on ajoute un terme quelconque θ_0 aux J paramètres β_{0j}, un terme θ_1 aux J paramètres β_{1j},..., un terme θ_k aux J paramètres β_{1j}. On obtient alors, en notant $\theta = (\theta_0, \theta_1, ..., \theta_K)$:

$$\frac{exp(x_i(\beta_j + \theta))}{\sum_{h=1}^{J} exp(x_i(\beta_h + \theta))} = \frac{exp(x_i\beta_j)exp(x_i\theta)}{\sum_{h=1}^{J} exp(x_i\beta_h)exp(x_i\theta)} = \frac{exp(x_i\beta_j)}{\sum_{h=1}^{J} exp(x_i\beta_h)}$$

Une infinité de β_j conduit à une même valeur de probabilité. Nous imposons donc la nullité de tous les paramètres relatifs à une catégorie donnée (appelée catégorie de référence), ce qui permet l'identification du modèle.

Avec cette condition identifiante et en écrivant le modèle sous une forme plus facile à manier, on obtient :

$$Ln(\frac{P(j/x_i)}{P(J/x_i)}) = x_i\beta_j \qquad pour \;\; j = 1, 2, ..., J$$

Ainsi, les paramètres du modèle s'interprètent comme des écarts au référentiel et les paramètres associés à la modalité de référence sont normalisés à 0.

En pratique, pour chaque variable à imputer, on sélectionne les variables qui l'expliquent le mieux à partir d'une procédure automatique de sélection de type « stepwise ». Un niveau usuel de significativité de 5 % est nécessaire pour permettre à la fois l'ajout d'une nouvelle variable dans le modèle et pour garder une variable dans le modèle.

En procédant pas à pas, on construit, pour chaque variable du questionnaire, un modèle polytomique non ordonné (cf. tableau 11), et qui s'écrit par exemple pour le diplôme, en prenant

la modalité 3 « Pas de diplôme, mais scolarité au-delà du collège » par rapport à la modalité 1 « Vous n'avez jamais été à l'école ou vous l'avez quittée avant la fin du primaire » :

$$Ln(\frac{P(dipl=3/x_i)}{P(dipl=1/x_i)}) = \beta_{01} + \beta_{11}.dep\_code_i + \beta_{21}.id\_scat_i + \beta_{31}.tranche\_age_i + \beta_{41}.matr_i + \beta_{51}.inscr_i$$

où id_scat, tranche_age, matr et inscr sont respectivement les variables « sous-catégorie de communauté », « tranche d'âge », « état matrimonial » et « inscription dans un établissement scolaire ».

Les contraintes d'imputation sont différentes pour chaque variable à imputer (cf. tableau 11). Seule la sous-catégorie de communauté explique fortement l'ensemble des variables à imputer.

Tableau 11 – Sélection des variables à utiliser comme contrainte d'imputation

Variables à imputer	Contraintes d'imputation	
	Hot-deck séquentiel avec redéfinition des contraintes	**Hot-deck actuel**
Sexe	Sous-catégorie de communauté	
	Effectif de la communauté	
Âge	Sous-catégorie de communauté	
	Effectif de la communauté	
	Sexe	
État matrimonial	Sous-catégorie de communauté	**Sexe**
	Effectif de la communauté	**Tranche d'âge**
	Sexe	
	Tranche d'âge	
Vie en couple	Sous-catégorie de communauté	**Tranche d'âge**
	Effectif de la communauté	**État matrimonial**
	Tranche d'âge	**Sexe**
	État matrimonial	
Indicatrice de nationalité	Sous-catégorie de communauté	**Pays de naissance**
	Effectif de la communauté	
	Tranche d'âge	
Inscription dans un étab. d'enseignement	Sous-catégorie de communauté	**Tranche d'âge**
	Tranche d'âge	
Situation principale	Sous-catégorie de communauté	**Indicatrice « travaille actuellement »**
	Inscription dans un étab. d'ens.	**Sexe**
	Tranche d'âge	**Tranche d'âge**
Diplôme	Sous-catégorie de communauté	**Tranche d'âge**
	Tranche d'âge	**Indicatrice de nationalité**

Résultats du hot-deck séquentiel avec redéfinition des contraintes

Nous évaluons le résultat du hot-deck séquentiel avec redéfinition des contraintes d'imputation en utilisant les 4 critères définis précédemment.

Critère n° 1 (cas des redressements aberrants) : pour les quatre cas de redressements aberrants considérés, la redéfinition des contraintes d'imputation permet de résoudre le problème. En effet, l'ajout de la sous-catégorie de communauté dans les contraintes d'imputation permet de trouver des donneurs beaucoup plus proches du non-répondant en terme de caractéristiques (notamment d'âge).

Critère n° 2 (simulation d'une non-réponse massive et concentrée dans les communes « test ») : le hot-deck avec redéfinition des contraintes diminue fortement la distorsion de la distribution des variables, par rapport au hot-deck actuel (cf. tableau 12).

Tableau 12 – Mesure de distorsion pour le hot-deck séquentiel avec redéfinition des contraintes

Variables à imputer	Hot-deck séquentiel avec redéfinition des contraintes	**Hot-deck séquentiel actuel**
Âge	9,5	**14,9**
État matrimonial	8,2	**14,4**
Situation principale	6,6	**13,1**
Diplôme	7,5	**10,6**

Critère n° 3 (nombre maximal de dons) : le nombre de dons par donneur est susceptible d'être très élevé, tant pour le hot-deck séquentiel actuel que celui avec redéfinition des contraintes d'imputation (cf. tableau 16). Ce type de redressement déterministe, qu'il utilise les contraintes actuelles ou redéfinies, est potentiellement massif pour certains donneurs très sollicités. Cela contribue à distordre la distribution des variables redressées. Par exemple, dans le hot-deck actuel, la réponse d'un individu à la variable « diplôme » a servi à redresser plus de 2 300 individus de 7 communautés non-répondantes dans les Bouches-du-Rhône qui se situaient à la suite dans le lot de saisie. Dans le hot-deck séquentiel avec redéfinition des contraintes, la réponse d'un individu a servi à redresser plus de 2 900 individus de 8 cités universitaires se situant à la suite dans le lot de saisie.

Critère n° 4 (simulation diffuse de non-réponse) : le taux de bien classés de la nouvelle méthode n'est pas meilleur que celui de la méthode actuelle (cf. tableau 13). Cela s'explique par le fait que la non-réponse étant diffuse, les individus ont été majoritairement redressés par des donneurs provenant de la même communauté dans les deux méthodes, ce qui gomme l'apport des nouvelles contraintes d'imputation.

Tableau 13 – Taux moyen de bien classés du hot-deck séquentiel avec redéfinition des contraintes

	Taux de bien classés	
Variables à imputer	Hot-deck séquentiel - redéfinition des contraintes d'imputation	**Hot-deck séquentiel actuel**
Sexe	69,2 %	**69,2 %**
Âge	75,2 %	**74,9 %**
État matrimonial	72,8 %	**72,7 %**
Vie en couple	91,7 %	**91,7 %**
Inscription dans un étab. enseis.	95,4 %	**95,2 %**
Situation principale	86,1 %	**86,1 %**
Indicateur résidence antérieure	67,3 %	**67,4 %**
Diplôme	50,2 %	**50,0 %**

5.4 Troisième proposition : ajout d'un aléa d'imputation

En plus du regroupement des individus en communauté dans un unique lot de saisie et de la redéfinition des contraintes d'imputation, nous proposons d'ajouter un aléa dans le processus d'imputation afin de limiter le nombre de dons par individu. Pour ce faire, nous adaptons la méthode de hot-deck séquentiel de cette manière : pour chaque non-répondant, nous tirons aléatoirement son donneur parmi les trois répondants qui précèdent dans le fichier et qui correspondent aux contraintes d'imputation de la méthode avec redéfinition des contraintes.

Les résultats de l'imputation pour les critères n° 1, 2 et 4 sont similaires à ceux du hot-deck séquentiel avec redéfinition des contraintes. L'ajout de l'aléa d'imputation ne réduit pas la distorsion induite par le redressement de la non-réponse dans le cas de non-réponse concentrée par rapport au hot-deck séquentiel avec redéfinition des contraintes (critère n° 2, cf. tableau 14). Le taux de bien classés en présence de non-réponse diffuse (critère n° 4) est globalement similaire à celui du hot-deck séquentiel actuel et à celui avec redéfinition des contraintes (cf. tableau 15). Il permet en revanche de limiter légèrement le nombre maximum de dons par individu, même si ce dernier reste tout de même très élevé (cf. tableau 16) : pour la variable diplôme, la réponse d'un individu a servi à redresser plus de 1 550 individus (contre 2 900 avec le hot-deck séquentiel avec redéfinition des contraintes).

Tableau 14 – Mesure de distorsion des variables pour le hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation

Variables à imputer	Hot-deck séquentiel avec redéfinition des contraintes et aléa	Hot-deck séquentiel avec redéfinition des contraintes	**Hot-deck séquentiel actuel**
Âge	7,9	9,5	**14,9**
État matrimonial	7,8	8,2	**14,4**
Situation principale	6,5	6,6	**13,1**
Diplôme	7,1	7,5	**10,6**

Tableau 15 – Taux moyen de bien classés du hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation

	Taux de bien classés	
Variables à imputer	Hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation	**Hot-deck actuel**
Sexe	69,3 %	**69,2 %**
Âge	75,2 %	**74,9 %**
État matrimonial	72,9 %	**72,7 %**
Vie en couple	91,6 %	**91,7 %**
Inscription dans un étab. enseis.	95,3 %	**95,2 %**
Situation principale	86,1 %	**86,1 %**
Indicateur résidence antérieure	67,6 %	**67,4 %**
Diplôme	50,1 %	**50 %**

5.5 Conclusion

La méthode actuelle peut-être améliorée sensiblement du point de vue des critères de qualité que nous avons retenus grâce à trois propositions :

- **Proposition n° 1 :** regrouper l'ensemble des individus en communauté dans un unique lot de saisie. Cette proposition permettrait de limiter les différences entre des individus consécutifs dans le lot de saisie (critère n° 1). Elle permet également de diminuer les chances pour que le premier individu du fichier soit défaillant et qu'il soit redressé par une valeur initiale déterminée a priori.
- **Proposition n° 2 :** redéfinir les contraintes d'imputation, et notamment en ajoutant la variable sous-catégorie de communauté qui explique l'ensemble des variables du questionnaire (critères n° 1 et 2).
- **Proposition n° 3 :** ajouter un aléa d'imputation permettant de limiter les redressements massifs de communautés entières par un unique individu (critère n° 3).

Tableau 16 – Distribution du nombre de dons par individu pour les différents hot-decks séquentiels testés

Variables	Effectifs des non-répondants	Distribution	Nombre de dons		
			Hot-deck séquentiel avec redéfinition des contraintes et aléa	Hot-deck séquentiel avec redéfinition des contraintes	**Hot-deck séquentiel actuel**
Sexe	6 106	Maximum	26	56	**56**
		99ème quantile	5	6	**6**
		90ème quantile	3	3	**3**
Âge	32 259	Maximum	436	872	**502**
		99ème quantile	58	103	**120**
		90ème quantile	18	21	**20**
État matrimonial	429 077	Maximum	1 596	3 265	**2 706**
		99ème quantile	61	115	**146**
		90ème quantile	21	31	**29**
Vie en couple	422 649	Maximum	2 543	4 640	**3 839**
		99ème quantile	71	112	**188**
		90ème quantile	25	34	**47**
Indicatrice de nationalité	276 904	Maximum	1 546	2 975	**2 703**
		99ème quantile	55	81	**112**
		90ème quantile	18	19	**14**
Inscription dans un étab. d'enseig.	507 355	Maximum	1 589	2 959	**2 370**
		99ème quantile	124	187	**194**
		90ème quantile	47	57	**62**
Situation principale	454 578	Maximum	1 964	3 683	**1 985**
		99ème quantile	102	153	**138**
		90ème quantile	37	45	**43**
Indicateur de résidence antérieure	466 565	Maximum	2 160	3 184	**2 370**
		99ème quantile	60	90	**128**
		90ème quantile	19	22	**27**
Diplôme	830 023	Maximum	1 550	2 911	**2 375**
		99ème quantile	104	170	**187**
		90ème quantile	39	25	**23**

Nous allons maintenant tester et comparer différentes méthodes de redressement qui pourraient être mises en place en cas de refonte à moyen terme des applications du recensement : le hot-deck par classe, le hot-deck métrique et la méthode par repondération.

6 Quelle méthode à adopter en cas de refonte des applications du recensement ?

6.1 Introduction

Dans cette sixième partie, nous mettrons en évidence qu'il serait intéressant d'envisager le hot-deck par classe et le hot-deck métrique en cas de refonte à moyen terme des applications du recensement. Nous expliquerons aussi pourquoi la méthode par repondération n'est pas adaptée à une diffusion au niveau communal.

6.2 Le hot-deck par classe

6.2.1 Le principe du hot-deck par classe

Le hot-deck par classe est une méthode d'imputation qui consiste à remplacer la donnée manquante par celle observée pour un individu répondant choisi aléatoirement à l'intérieur d'une classe d'imputation. Les classes sont construites à partir des variables auxiliaires et sont définies de manière à être homogènes par rapport à la probabilité de réponse et par rapport à la réponse donnée à la variable à imputer. Le choix du nombre de classes résulte d'une concession entre, d'une part, augmenter le nombre de classes pour assurer une plus grande homogénéité à l'intérieur des classes, et, d'autre part, diminuer le nombre de classes pour avoir davantage de répondants afin de gagner en robustesse.

Comme actuellement, le redressement s'effectue variable par variable, dans un ordre logique prédéfini. La première variable est imputée en utilisant potentiellement les variables auxiliaires, puis la variable imputée est utilisée comme variable auxiliaire pour imputer les variables suivantes et ainsi de suite.

6.2.2 Création de classes homogènes d'imputation

Nous utiliserons la méthode du score pour former nos classes d'imputation. Cette méthode nous conduit, pour chaque variable, à partitionner nos individus de telle sorte qu'à l'intérieur des classes, les individus soient homogènes à la fois du point de vue de la réponse à la variable, mais aussi du comportement de non-réponse. Cette double spécification permet de donner plus de robustesse au modèle.

Soit n le nombre d'individus et J le nombre de modalités de la variable à imputer. Chaque individu i est décrit par un ensemble de K caractéristiques x_{i1}, x_{i2},..., x_{ik}, par exemple la sous-catégorie de la communauté d'appartenance, sa tranche d'âge ou son sexe.

Pour chaque individu i, il y a deux cas possibles : soit l'individu a répondu à la variable à imputer et il appartient à une catégorie j parmi les J modalités de la variable à expliquer, soit il n'a pas répondu à la variable à imputer.

Pour chaque variable du questionnaire, nous procédons ainsi :

- nous estimons les probabilités de réponse p_i en utilisant les K caractéristiques connues de

l'individu x_{i1}, x_{i2},..., x_{ik} grâce à un modèle logistique. Ce modèle s'écrit :

$$Logit(p_i) = log(\frac{p_i}{(1-p_i)}) = \beta x_i$$

où p_i est la probabilité de réponse de l'individu i, β le vecteur des coefficients de la régression logistique, x_i le vecteur des variables explicatives du comportement de réponse et ϵ_i le résidu entre l'estimation $\hat{p_i} = \beta x_i$ et la valeur réelle p_i. Une fois le vecteur des coefficients β estimés par la méthode du maximum de vraisemblance, la probabilité de réponse estimée $\hat{p_i}$ peut être déduite pour chaque individu grâce à la fonction exponentielle.

- nous estimons les odds ratio (les probabilités relatives) r_{ij} pour chaque modalité j de la variable à imputer par rapport à une modalité de référence O. Cette estimation s'effectue grâce à un modèle polytomique non-ordonné en utilisant les caractéristiques de l'individu x_{i1}, x_{i2},..., x_{ik} :

$$Ln(r_{ij}) = Ln(\frac{P(j/x_i)}{P(O/x_i)}) = \beta_j x_i \qquad pour j = 1, 2, ..., J \; avec \; j \neq O$$

  Une fois le vecteur des coefficients β_j estimés par la méthode du maximum de vraisemblance, les odds ratio estimés de chaque modalité $\hat{r_{ij}}$ peuvent être déduits pour chaque individu grâce à la fonction exponentielle.

- nous utilisons un algorithme de classification de type « k-means » afin d'obtenir des classes homogènes par rapport aux estimations $\hat{p_i}$ et $\hat{r_{ij}}$.

La classification se fait sur la base du critère des plus proches voisins : chaque individu est affecté à une classe s'il est très proche de son centre de gravité. Cette méthode a l'avantage d'être efficace et très rapide.

Appelons k le nombre de classes homogènes que nous souhaitons obtenir. L'algorithme de la méthode de classification « k-means » est le suivant :

- *Etape 0 : Initialisation de l'algorithme.* On tire au hasard k individus appartenant à la population (appelés k centres initiaux) : C_1^0, C_2^0...C_k^0.
- *Etape 1 : Constitution de classes.* On répartit l'ensemble des individus en k classes Γ_1^0, Γ_2^0... Γ_k^0 regroupant autour de chaque centre C_i^0 pour i = 1,2,...,n l'ensemble des individus qui sont plus proches C_i^0 que des autres centres C_j^0 pour $j \neq i$.
- *Etape 2 : Calcul des nouveaux centres.* On détermine les centres de gravité G_1, G_2,..., G_k des k classes ainsi obtenues et on désigne ces points comme les nouveaux centres $C_1^1 = G_1, C_2^1 = G_2, ..., C_k^1 = G_k$

On répète les étapes 1 et 2 jusqu'à la stabilisation de l'algorithme, c'est-à-dire jusqu'à ce que le découpage en classes obtenu ne soit presque plus modifié par une itération supplémentaire.

Pour déterminer le nombre optimal de classes, nous utilisons un critère empirique qui se base sur le coefficient de détermination (D. Haziza et J.F. Beaumont, 2007). Le coefficient de détermination est le carré du coefficient de corrélation résultant d'une analyse de la variance entre la variable dépendante p_i ou r_i et l'identifiant de la classe. A mesure que le nombre de

classes augmente, ces dernières deviennent de plus en plus homogènes et le R^2 tend vers 1. Nous recherchons le plus petit nombre de classes tel que R^2 soit supérieur à un seuil relativement élevé (fixé arbitrairement à 0,95). Ainsi, pour la variable « situation principale », nous obtenons des classes relativement homogènes à partir de 30 classes d'imputation (cf. tableau 18)

Tableau 17 – Coefficient de corrélation résultant d'une analyse de la variance entre le score et la variable de classe pour la situation principale

Nombre de classes	R^2
10	0,85
15	0,92
20	0,92
25	0,94
30	0,95

Résultats du hot-deck par classe

Comme pour le hot-deck séquentiel, nous évaluons le résultat du hot-deck par classe en utilisant les 4 critères définis précédemment.

Critère n° 1 (cas des redressements aberrants) : pour les quatre cas de redressements aberrants, le hot-deck par classe permet de résoudre le problème, les individus en foyer de travailleurs sont notamment redressés par d'autres individus en foyer de travailleurs.

Critère n° 2 (simulation d'une non-réponse massive et concentrée dans les communes « test ») : le hot-deck par classe réduit fortement la distorsion par rapport à la méthode actuelle et également par rapport au hot-deck séquentiel avec redéfinition des contraintes (cf. tableau 18).

Tableau 18 – Mesure de distorsion des variables du hot-deck par classe

Variables à imputer	Hot-deck par classe	Hot-deck séquentiel avec redéfinition des contraintes et aléa	**Hot-deck séquentiel actuel**
Âge	7,0	7,9	**14,9**
État matrimonial	7,3	7,8	**14,4**
Situation principale	4,5	6,5	**13,1**
Diplôme	6,7	7,1	**10,6**

Critère n° 3 (nombre maximal de dons) : le hot-deck par classe permet d'éviter les donneurs multiples, de par le tirage aléatoire au sein de classes d'imputation vastes.

Critère n° 4 (simulation diffuse de non-réponse) : le taux de bien classés du hot-deck par classe est inférieur à celui de la méthode actuelle pour l'ensemble des variables du questionnaire (cf. tableau 19). Ce faible taux moyen de bien-classés provient essentiellement d'individus en service de moyen ou de long séjour (sous-catégorie n° 14) ou en foyer de travailleurs (sous-catégories

n° 12 et 13). Cela s'explique par le fait que ces sous-catégories contiennent des individus aux caractéristiques très différentes : par exemple, les services de moyen ou de long séjour regroupent à la fois des enfants et des adultes handicapés ou nécessitant des soins médicaux.

Tableau 19 – Taux moyen de bien classés du hot-deck par classe

	Taux de bien classés	
Variables à imputer	Hot-deck par classe	**Hot-deck actuel**
Sexe	61,3 %	**69,2 %**
Âge	66,0 %	**74,9 %**
État matrimonial	68,2 %	**72,7 %**
Vie en couple	72,5 %	**91,7 %**
Inscription dans un étab. enseis.	89,3 %	**95,2 %**
Situation principale	70,0 %	**86,1 %**
Indicateur résidence antérieure	53,0 %	**67,4 %**
Diplôme	42,3 %	**50,0 %**

6.2.3 Conclusion

Au vu de nos critères de qualité, le bilan du hot-deck par classe est relativement bon. Il permet en effet de résoudre le cas des redressements aberrants (critère n° 1) et de limiter le nombre de dons multiples (critère n° 3). Il permet de réduire la distorsion des variables dans le cas d'une non-réponse massive et concentrée par rapport au hot-deck séquentiel actuel (critère n° 2). Il améliore plus modérément la distorsion par rapport au hot-deck séquentiel avec redéfinition des contraintes.

Toutefois, le taux de bien classés en cas de non-réponse diffuse est inférieure à celle du hot-deck séquentiel (critère n° 4) ; cela est dû au fait que certaines sous-catégories de communautés contiennent des individus aux caractéristiques très hétérogènes.

A noter qu'à partir de la collecte 2016, la sous-catégorie « services de moyen ou de long séjour » qui regroupe 8,8 % de la population en communauté sera ventilée en cinq postes plus homogènes :

- les structures pour personnes nécessitant des soins médicaux (enfants, adultes) ;
- les structures pour les enfants handicapés ;
- les structures pour les adultes handicapés ;
- les structures d'aide sociale à l'enfance et de protection judiciaire pour enfants et jeunes majeurs ;
- les structures adultes et familles nécessitant un accompagnement social et psychologique.

Cette décision de redécouper la sous-catégorie « services de long et moyen séjour » est susceptible d'améliorer les résultats de ce hot-deck.

6.3 Le hot-deck métrique

Nous avons également envisagé un hot-deck métrique basé sur le V de Cramer ; les modalités de la mise en œuvre ainsi que les résultats de cette méthode sont décrits ci-dessous.

6.3.1 Le principe du hot-deck métrique

Le hot-deck métrique est une méthode d'imputation qui consiste à remplacer la valeur manquante pour un receveur par la valeur observée pour le donneur le plus proche, au sens d'une mesure de similarité. Cette mesure est calculée à partir des informations auxiliaires et celles renseignées par les individus ; elle doit être définie de façon à respecter la corrélation entre ces variables et la variable d'intérêt à imputer, en accordant plus d'importance aux variables les plus liées à la variable d'intérêt. L'objectif est de tirer parti des informations disponibles (à la fois les variables auxiliaires et les réponses de l'individu). Comme pour les précédentes méthodes, nous redressons les variables les unes après les autres dans un ordre pré-établi, en utilisant les informations précédemment imputées. Ainsi, pour chaque variable, le champ des donneurs potentiels correspond à l'ensemble des individus ayant répondu à cette variable.

Afin d'ajouter un aléa à l'imputation et donc de limiter le nombre de réplications, le donneur est tiré au sort parmi ceux qui maximisent la distance de similarité.

6.3.2 Une mesure de similarité basée sur le V de Cramer

Calcul des pondérations affectées aux différentes variables auxiliaires

Les pondérations affectées aux variables auxiliaires sont d'autant plus fortes que leurs corrélations avec les variables à imputer sont importantes. Nous choisissons ces pondérations comme la somme normalisée des V de Cramer calculées sur l'ensemble des 10 variables à imputer. Contrairement au χ^2, le V de Cramer a l'avantage de ne tenir compte ni de la taille de l'échantillon, ni du nombre de modalités des variables auxiliaires.

Le V de Cramer entre deux variables qualitatives X et Z est défini ainsi :

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{inf(t-1, k-1)}}$$

où t est le nombre de modalités de X, k le nombre de modalités de Y, n est la taille de la population et χ^2 la mesure du Khi-deux entre X et Y :

$$\chi^2 = \sum_{i=1}^{t} \sum_{j=1}^{k} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{ni.n.j}{n}}$$

avec :

- n_{ij} l'effectif des individus ayant la modalité i de la variable X et la modalité j de la variable Y ;
- $n_{i.}$ l'effectif des individus ayant la modalité i de la variable X ;
- $n_{.j}$ l'effectif des individus ayant la modalité j de la variable Y.

La mesure de similarité entre le receveur r et le donneur potentiel d s'écrit :

$$S = \frac{\sum_{j=1}^{p} \omega_j \delta_{j,rd}}{\sum_{j=1}^{p} \omega_j}$$

où :

- p est le nombre de variables auxiliaires ;
- w_j le poids de la variable auxiliaire x_j au sens du V de Cramer ;
- $d_{j,rd}$ vaut 0 si la variable auxiliaire x_j prend la même modalité pour le receveur r et le donneur d, c'est-à-dire si $x_{j,r} = x_{j,d}$, et 1 sinon.

Cette méthode présente l'inconvénient de ne pas prendre en compte les corrélations possibles entre les variables auxiliaires, ce qui peut conduire à surestimer le poids attribué à celles-ci : l'effet régional s'expliquant en partie par une répartition inégale des sous-catégories de communautés sur le territoire, il est à la fois pris en compte par la pondération affectée à la sous-catégorie mais aussi à celle de la variable région. Nous contournons ce problème en prenant la ZEAT au lieu de la région.

Tableau 20 – V de Cramer entre les variables à imputer et les variables auxiliaires

Variables	Sous-catégorie de communauté	ZEAT	Effectif de la communauté	Tranche d'aire urbaine	Catégorie d'aire urbaine	Indicatrice « première collecte »
Sexe	0,46	0,09	0,23	0,09	0,09	0,04
Âge	0,46	0,08	0,18	0,10	0,10	0,08
État matrimonial	0,48	0,07	0,17	0,08	0,10	0,09
Vie en couple	0,57	0,08	0,13	0,09	0,06	0,08
Nationalité	0,37	0,18	0,13	0,18	0,12	0,02
Indicatrice « travaille actuellement »	0,53	0,09	0,20	0,08	0,06	0,03
Inscription étab. d'enseig.	0,55	0,04	0,16	0,07	0,10	0,08
Situation principale	0,47	0,06	0,18	0,07	0,09	0,12
Résidence antérieure	0,39	0,11	0,17	0,10	0,10	0,14
Diplôme	0,34	0,09	0,22	0,12	0,13	0,11

Prise en compte de l'information fournie par les variables potentiellement imputables

Parmi l'ensemble des 10 variables imputables, les réponses à un certain nombre d'entre elles sont connues, et d'autres ont été imputées. Pour prendre en compte les corrélations des variables manquantes et celles déjà présentes (connues ou imputées), on introduit dans le calcul de la

distance les 9 variables pondérées selon le même principe que les variables auxiliaires (somme normalisée des V de Cramer avec les autres variables).

Finalement, on obtient la mesure de similarité suivante entre un donneur potentiel et un receveur :

$$S = \frac{\sum_{j=1}^{p} \omega_j \delta_{j,rd} + \sum_{k=1}^{q} \omega_k \delta_{k,rd}}{\sum_{j=1}^{p} \omega_j + \sum_{k=1}^{q} \omega_k}$$

où :

- p est le nombre de variables auxiliaires ;
- w_j le poids de la variable auxiliaire j au sens du V de Cramer ;
- w_k le poids de la variable imputable k ;
- $\delta_{j,rd}$ (respectivement $\delta_{k,rd}$) vaut 0 si la variable j (resp. k) prend la même modalité pour le receveur r et le donneur d, et 1 sinon.

Les donneurs potentiels sont ceux qui sont les plus proches du receveur, i.e. ceux qui maximisent cette mesure de similarité. Le donneur est choisi aléatoirement parmi l'ensemble de ces donneurs potentiels.

Tableau 21 – Pondérations affectées aux variables auxiliaires et aux variables du questionnaire pour la variable « situation principale »

Variable	Pondération
Sous-catégorie	0,11
ZEAT	0,01
Effectif de la communauté	0,04
Tranche d'aire urbaine	0,02
Catégorie d'aire urbaine	0,02
Indicatrice « première collecte »	0,03
Sexe	0,07
Âge	0,12
Nationalité	0,02
Inscription dans un étab. d'enseignement	0,18
Indicateur de résidence antérieure	0,02
Vie en couple	0,03
État matrimonial	0,09
Diplôme	0,07
Indicatrice « travaille actuellement »	0,17

A priori, l'inconvénient pratique majeur de cette méthode est son coût très lourd en temps de traitement à cause de sa procédure itérative : pour chaque non-répondant il faut calculer sa distance avec chaque donneur potentiel avant de tirer aléatoirement dans l'ensemble de ceux qui maximisent la mesure de similarité. Ainsi, avec une non-réponse de l'ordre de 300 000 non-répondants pour la variable état matrimonial et le redressement d'un non-répondant en 1

seconde, le temps de traitement total s'élève à 83 heures (pour une variable seulement !).

Le procédé alternatif que nous avons utilisé consiste à contourner l'aspect itératif du hot-deck métrique de cette manière : nous rassemblons l'ensemble des non-répondants et l'ensemble de leurs donneurs potentiels dans un fichier, et nous procédons au redressement de l'ensemble des non-répondants en une seule étape. Toutefois, pour limiter le nombre d'observations dans le fichier, il s'agit de présélectionner les donneurs potentiels : nous n'utilisons que les répondants de même sous-catégorie de communauté et de même département. Ce procédé permet de limiter le temps de traitement pour une variable à une dizaine de minutes.

Une piste intéressante pour implémenter le hot-deck métrique serait de tester le système gratuit CANCEIS (acronyme pour CANadian Census Edit and Imputation System), développé par Statistics Canada. Ce système, extrêmement rapide et flexible, permet l'imputation par le plus proche voisin avec composante aléatoire.

Résultats du hot-deck métrique

Comme précédemment, nous évaluons le résultat du hot-deck métrique à la lumière de nos 4 critères de qualité.

Critère n° 1 (cas des redressements aberrants) : le hot-deck métrique permet de résoudre le problème pour les cas de redressements aberrants, grâce à l'importance de la sous-catégorie de communauté dans la mesure de similarité.

Critère n° 2 (simulation d'une non-réponse massive et concentrée dans les communes « test ») : le hot-deck métrique réduit fortement la distorsion par rapport à la méthode actuelle (cf. tableau 22). Il n'améliore que pour certaines variables la distorsion par rapport au hot-deck séquentiel avec redéfinition des contraintes.

Tableau 22 – Mesure de distorsion du hot-deck métrique

Variables à imputer	Hot-deck métrique	Hot-deck séquentiel avec redéfinition des contraintes et aléa	**Hot-deck séquentiel actuel**
Âge	6,1	7,9	**14,9**
État matrimonial	7,9	7,8	**14,4**
Situation principale	4,2	6,5	**13,1**
Diplôme	6,4	7,1	**10,6**

Critère n° 3 (nombre maximal de dons) : le hot-deck métrique permet de réduire le nombre de dons maximum par rapport à la méthode actuelle (cf. tableau 23).

Critère n° 4 (simulation diffuse de non-réponse) : le taux de bien classés du hot-deck métrique est globalement similaire à la méthode actuelle pour l'ensemble des variables du questionnaire (cf. tableau 24).

Tableau 23 – Distribution du nombre de dons par individu pour le hot-deck métrique

Variables	Effectifs des non-répondants	Distribution	Nombre de dons	
			Hot-deck métrique	**Hot-deck séquentiel actuel**
Sexe	6 106	Maximum	31	**56**
		99ème quantile	5	**6**
		90ème quantile	4	**3**
Âge	32 259	Maximum	61	**502**
		99ème quantile	6	**120**
		90ème quantile	3	**20**
État matrimonial	429 077	Maximum	745	**2 706**
		99ème quantile	18	**146**
		90ème quantile	8	**29**
Couple	422 649	Maximum	591	**3 839**
		99ème quantile	19	**188**
		90ème quantile	8	**47**
Indicatrice de nationalité	276 904	Maximum	2 125	**2 703**
		99ème quantile	22	**112**
		90ème quantile	6	**14**
Inscription dans un étab. d'enseig.	507 355	Maximum	1 307	**2 370**
		99ème quantile	91	**194**
		90ème quantile	23	**62**
Situation principale	454 578	Maximum	1 246	**1 985**
		99ème quantile	185	**138**
		90ème quantile	58	**43**
Indicateur de résidence antérieure	466 565	Maximum	1462	**2 370**
		99ème quantile	75	**128**
		90ème quantile	21	**27**
Diplôme	830 023	Maximum	1 516	**2 375**
		99ème quantile	106	**187**
		90ème quantile	29	**23**

6.3.3 Conclusion

Le hot-deck métrique permet de résoudre le cas des redressements aberrants (critère n° 1) et de limiter le nombre de dons multiples (critère n° 3). Le taux de bien classés du hot-deck métrique est globalement similaire à la méthode actuelle pour l'ensemble des variables du questionnaire (critère n° 4). Le hot-deck métrique permet de réduire sensiblement la distorsion des variables dans le cas d'une non-réponse massive et concentrée par rapport au hot-deck séquentiel actuel et au hot-deck séquentiel avec redéfinition des contraintes (critère n° 2).

Tableau 24 – Taux moyen de bien classés du hot-deck métrique

Variables à imputer	Taux de bien classés	
	Hot-deck métrique	**Hot-deck actuel**
Sexe	69,2 %	**69,2 %**
Âge	72,0 %	**74,9 %**
État matrimonial	76,6 %	**72,7 %**
Vie en couple	90,4 %	**91,7 %**
Inscription dans un étab. enseis.	96,2 %	**95,2 %**
Situation principale	95,8 %	**86,1 %**
Indicateur résidence antérieure	65,0 %	**67,4 %**
Diplôme	46,7 %	**50,0 %**

6.4 Un redressement de la non-réponse par repondération ?

Après la mise en œuvre des méthodes d'imputation, on peut s'interroger sur la pertinence d'un traitement de la non-réponse par repondération. Après avoir rappelé les principes de ce traitement, nous expliquerons pourquoi il ne peut pas être mis en place dans le cadre du recensement des communautés.

Le principe de la méthode par repondération consiste à modifier le poids des unités répondantes pour compenser la présence de non-réponse totale afin d'extrapoler les résultats obtenus à la population de référence. Le poids initial de chaque unité répondante est ainsi augmenté par l'inverse de sa probabilité de réponse, quantité inconnue qu'il faut estimer. Dans ce but, une méthode consiste à supposer le mécanisme de réponse homogène à l'intérieur de sous-populations. Cette approche repose sur l'hypothèse qu'à l'intérieur de sous-populations particulières les individus possèdent tous la même probabilité de répondre et que leurs comportements de réponse sont indépendants. Au sein d'un groupe donné, la probabilité de réponse est estimée en rapportant le nombre d'unités répondantes à l'effectif collecté.

Toutefois, l'utilisation de la méthode de repondération n'est pertinente que si nous étudions les communautés à un niveau national ou encore à un niveau régional. En revanche, elle est problématique lorsque l'on fait des études à un niveau communal car les groupes de réponse homogènes (GRH) ne peuvent pas être définies au sein de chaque commune, étant donné :

- le faible nombre d'individus en communauté dans certaines communes ;
- l'absence de répondants (ou le faible taux de répondants) dans certaines communes.

Les statistiques et les populations légales étant élaborées à un niveau communal, la méthode par repondération n'est pas adaptée au redressement de la non-réponse dans les communautés. Illustrons le problème engendré par la repondération à un niveau communal sur un cas simple : on considère le cas où nous avons un seul groupe homogène et une variable commune dont les deux modalités A et B ne correspondent pas avec le découpage en GRH puisqu'elles sont toutes les deux présentes dans le groupe.

N n° obs	Commune	Poids initial	Variable d'intérêt : sexe
1	A	1	H
2	A	1	F
3	A	1	
4	B	1	

La correction de la non-réponse par repondération est adaptée si l'on cherche à calculer un estimateur au niveau global (en ne distinguant pas les deux communes) ; le résultat sera identique à la méthode par imputation : dans le cas de la repondération, on augmente le poids des observations 1 et 2 afin d'avoir un total de 4 individus ; dans le cas de l'imputation, on utilisera les réponses de 1 ou 2 pour compléter les réponses des individus 3 et 4, et aboutir à 4 individus.

En revanche, si on souhaite obtenir des estimateurs au niveau communal, la correction par imputation donnera ici 3 individus dans la commune A et 1 dans la commune B, alors que la correction par repondération donnera 4 individus dans la région A : avec la repondération, on donne plus de poids à la commune A et on perd la commune B.

La repondération ne peut donc être envisagée dans le cadre du recensement, car elle conduit à une perte d'information au niveau communal. Cette perte n'est pas acceptable dans la mesure où :

- les populations légales fixent le nombre d'individus avant les traitements ;
- les résultats du recensement doivent refléter fidèlement la population de chaque commune.

7 Enseignements, propositions et préconisations

Dans ce chapitre, nous récapitulerons tout d'abord les enseignements tirés de l'étape préalable au redressement. Nous proposerons ensuite des améliorations à moindre coût de la méthode actuelle par hot-deck séquentiel. Enfin, nous préconiserons des méthodes alternatives de redressement qui pourraient être mises en place en cas de refonte à moyen terme des applications du recensement.

Constats et enseignements tirés des étapes préalables au redressement

L'étape préalable au redressement nous a permis de caractériser la population en communauté et d'analyser son comportement de non-réponse. Les personnes en communauté sont ainsi très différentes des ménages :

- La non-réponse y est beaucoup plus forte et concentrée, avec un phénomène de non-réponse massive de communautés entières. Cette non-réponse pourrait être amenée à s'intensifier en raison des nouvelles conditions d'emploi des enquêteurs Insee.
- La population de chaque communauté est très spécifique et ne ressemble pas forcément à celle d'une communauté voisine.

Ces particularités des communautés amplifient deux inconvénients de la méthode actuelle de redressement par hot-deck séquentiel. Premièrement, le comportement de non-réponse est fortement lié à l'ordre du lot de saisie. En cas de non-réponse massive de communautés entières, les réponses du dernier individu répondant d'une structure voisine sont ainsi imputées de manière déterministe à tous les non-répondants de la communauté. Deuxièmement, le hot-deck séquentiel repose sur l'hypothèse pas toujours vérifiée que les réponses d'un individu ainsi que son comportement de réponse sont similaires à ceux de l'individu précédent.

Ces particularités, combinées au fait que les communautés soient susceptibles d'être de taille très importante, engendrent quelques cas de redressements aberrants visibles à un niveau communal. Ces cas s'avèrent regrettables au vu des attentes que les résultats du recensement soulèvent à un niveau communal et infracommunal.

Afin de comparer aisément les performances des différentes méthodes de redressement testées, nous avons défini ce que nous attendons de la méthode de redressement :

- Nous souhaitons qu'elle résolve les cas de redressements aberrants de la méthode actuelle.
- Nous souhaitons qu'elle réduise la distorsion moyenne de la distribution des variables en cas de non-réponse massive et concentrée.
- La méthode doit également permettre de réduire le nombre maximal d'utilisation de la réponse d'un individu répondant pour imputer des individus non-répondants.
- La méthode doit conserver une bonne qualité prédictive en cas de non-réponse diffuse.

Nous aurions également pu définir des critères supplémentaires. Par exemple, nous aurions pu considérer les distorsions maximales de la distribution des variables induites par le redressement de la non-réponse à un niveau communal. Cette mesure aurait permis de repérer les redressements maladroits des différentes méthodes envisagées.

Propositions et préconisations à court terme tirées du redressement de l'enquête

La méthode actuelle de redressement par hot-deck séquentiel est susceptible d'être considérablement améliorée à moindre coût, principalement grâce à deux propositions. La première consiste à regrouper l'ensemble des individus en communauté dans un unique lot de saisie, permettant ainsi de limiter les différences entre des individus successifs dans le lot de saisie. Cette proposition permettrait d'éviter certains redressements aberrants mettant en jeu des informations de nature géographique. Elle sera implémentée pour la collecte 2016, à l'occasion de la fin du projet de modernisation de la collecte « Homere ».

La deuxième proposition consiste à redéfinir les contraintes d'imputation du hot-deck séquentiel, et notamment en ajoutant la variable sous-catégorie de communauté qui explique l'ensemble des variables du questionnaire. Un moyen d'implémenter cette proposition à partir de la collecte 2016 ou 2017 serait de trier le lot de saisie par sous-catégorie de communauté avant le passage du hot-deck séquentiel. Cela sera rendu possible par le regroupement de l'ensemble des individus en communauté dans un unique lot de saisie. Par ailleurs, à partir de la collecte 2016, le pouvoir explicatif de cette variable sera par ailleurs vraisemblablement améliorée grâce à la ventilation de la sous-catégorie « services de moyen ou de long séjour » en cinq postes plus homogènes.

Une troisième proposition, qui semble plus difficile à mettre en place à court terme, consiste à ajouter un aléa dans le processus d'imputation. En pratique, il s'agit pour chaque non-répondant de tirer son donneur parmi les trois (ou plus) répondants qui précèdent dans le fichier. Cette proposition permet de limiter quelque peu les redressements massifs de communautés entières par un unique individu.

Propositions et préconisations à moyen terme tirées du redressement de l'enquête

Dans une vision à moyen terme, nous avons considéré des méthodes de redressement alternatives que sont le hot-deck par classe, le hot-deck métrique et la repondération. Le hot-deck métrique semble être a priori une méthode de redressement particulièrement adaptée à l'enquête de recensement des communautés.

En effet, le hot-deck métrique permet d'améliorer l'ensemble de nos critères de qualité. Il permet ainsi de résoudre les cas de redressements aberrants, de limiter les dons multiples et de réduire la distorsion des variables par rapport au hot-deck séquentiel. Quant à sa qualité prédictive en cas de non-réponse diffuse, elle est globalement similaire à la méthode actuelle. Il serait intéressant de tester le hot-deck métrique en utilisant le système gratuit CANCEIS, développé par Statistics Canada, qui permet l'imputation par le plus proche voisin avec composante aléatoire. L'avantage de ce système est qu'il est extrêmement rapide et offre de la flexibilité. Il pourrait être intéressant de recourir à ce système.

Le hot-deck par classe donne de bons résultats : il permet de résoudre les cas de redressements aberrants, de limiter les dons multiples et de réduire la distorsion des variables par rapport au hot-deck séquentiel. En revanche, sa qualité prédictive en cas de non-réponse diffuse est inférieure à celle du hot-deck séquentiel.

Enfin, la méthode de redressement par repondération n'est pas envisageable car elle engendre une perte d'information à un niveau communal en raison du faible nombre d'individus en communauté ou encore de l'absence de répondants (ou le faible taux de répondants) dans quelques communes. Cette perte d'information n'est pas admissible dans la mesure où les populations légales et les résultats du recensement sont attendus pour refléter fidèlement la population de chaque commune.

8 Conclusion

Le recensement de la population permet principalement de déterminer les populations légales et de décrire les structures démographiques et sociales de la population et des logements qu'elle occupe. Ces résultats, établis à un niveau communal, voir infracommunal pour les plus localisés, sont essentiels au fonctionnement de la société et sont attendus avec une grande exigence sur leur qualité à tout échelon géographique.

Certains utilisateurs se sont toutefois étonnés de quelques résultats communaux aberrants, comme la présence d'une vingtaine de cadres nonagénaires dans une commune des Hauts-de-Seine. Il s'avère que ces cas fâcheux sont dus au redressement maladroit de non-réponses d'individus en communauté. Ces cas ne concernent pas les ménages : chez ces derniers, l'impact du redressement de la non-réponse est moins apparent du fait d'un volume plus important d'individus, et d'une non-réponse plus diffuse et de plus faible ampleur.

Le redressement de la non-réponse s'effectue par un hot-deck séquentiel. Cette méthode a l'avantage principal d'être robuste et efficace, permettant de respecter la forte contrainte de délai à laquelle le recensement est soumis. Toutefois, cette méthode comporte certaines faiblesses, particulièrement impactantes pour la population des communautés, du fait d'une non-réponse bien plus concentrée et de caractéristiques très spécifiques à chaque communauté.

A moindre coût en matière de production courante et d'évolution des processus, nous pouvons limiter l'impact de ces faiblesses de deux manières :

- d'une part en regroupant l'ensemble des individus en communauté dans un unique lot de saisie. Cette proposition permettrait de limiter les différences au regard des variables liées au lieu d'habitat entre des individus consécutifs dans le lot de saisie.
- d'autre part en redéfinissant les contraintes d'imputation, et notamment en ajoutant la variable sous-catégorie de communauté qui explique l'ensemble des variables du questionnaire (une solution équivalente consiste à trier les individus selon la sous-catégorie de communauté avant de procéder au redressement).

Parmi les méthodes de redressement que nous avons testées, le hot-deck métrique est celle qui obtient les meilleurs résultats pour la population des communautés. Elle a tout d'abord l'avantage de réduire sensiblement la distorsion de la distribution des variables en cas de non-réponse massive et concentrée. Elle possède également un bon comportement prédictif en cas de non-réponse diffuse. Enfin, elle permet également d'éviter les redressements aberrants de la méthode actuelle et de limiter le nombre de donneurs multiples. Cette méthode pourrait être testée en utilisant le système CANCEIS développé par Statistics Canada.

9 Bibliographie

[1] **CARON N. (2005)**, La correction de la non-réponse par repondération et par imputation, document de travail Insee n° M0502

[2] **GODINOT A. (2005)**, Pour comprendre le recensement de la population, Insee méthode hors série

[3] **HAZIZA D. (2012)**, Redressement d'échantillon et traitement de la non-réponse support de cours Master de statistique publique

[4] **HAZIZA, D. et BEAUMONT, J-F. (2007)**, On the construction of imputation classes in surveys, International Statistical Review, 75, 25-43

[5] **REYNAERT L. (2015)**, Amélioration du redressement de la non-réponse des communautés dans le recensement, journées de la méthodologie statistique

[6] **PIROU D., POULLAIN N, ROCHELLE S. (2013)**, « La vie en communauté : 1,6 million de personnes en France », Insee Première n° 1434

[7] **Documentation Insee.fr** sur les redressements dans le recensement

10 Annexes

A1 : Organigramme simplifié de la Direction Générale de l'Insee (août 2015)

A2 : Organigramme de la Direction des Statistiques Démographiques et Sociales (août 2015)

A3 : Bulletin individuel communauté - hors centres pénitentiaires

A4 : Bulletin individuel pour les centres pénitentiaires

A5 : Signification des modalités des variables auxiliaires

Tableau A1 – Organigramme simplifié de la Direction Générale de l'Insee (août 2015)

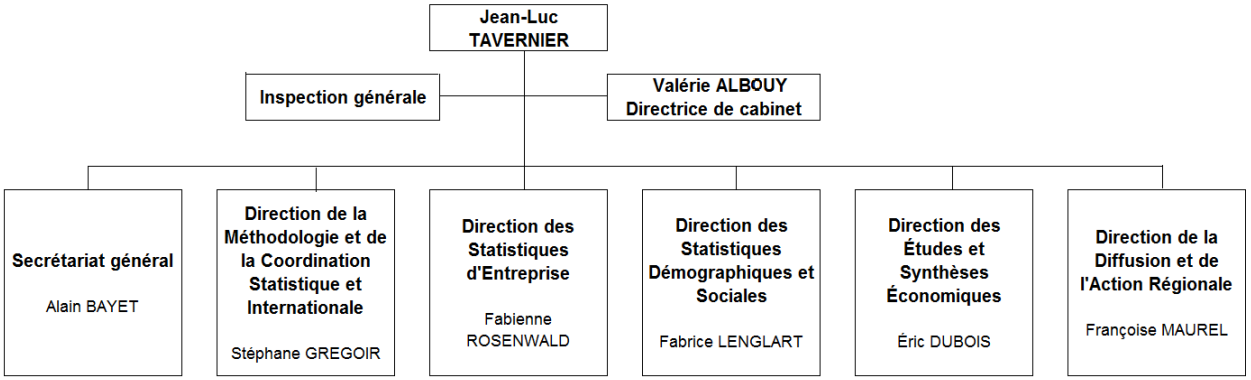


Tableau A2 – Organigramme de la Direction des Statistiques Démographiques et Sociales (DSDS, août 2015)

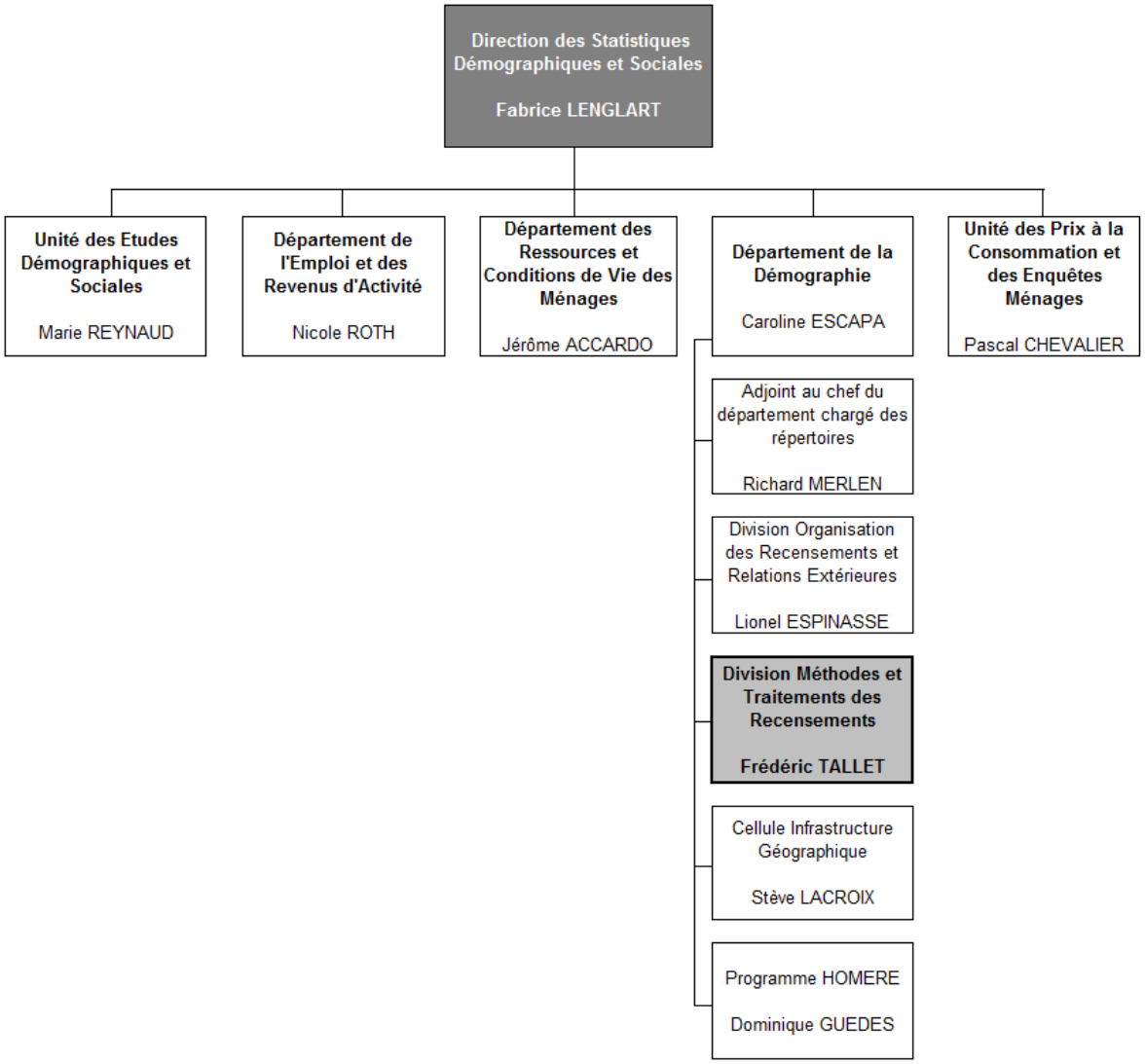


Tableau A3 – Bulletin individuel communauté - hors centres pénitentiaires

145141_5EP_imprime_7_specimen_210x297 20/05/2014 11:12 Page1

Recensement de la population - 2015

Bulletin individuel – Communauté

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Imprimé n° 7

Exemple : DUPAS, épouse MAURIN

Nom :

Prénom :

Identifiant de la communauté

Cadre à remplir par l'enquêteur

Avez-vous une résidence personnelle dans une autre commune ? (exemple : adresse des parents pour un élève interne)

Non ☐ 1 Oui ☐ 2 ⇨ *Si oui, précisez où :*

commune (et arrondissement pour Paris, Lyon, Marseille) département n° DOM Pays pour l'étranger, territoire pour les TOM

1 Sexe Masculin ☐ 1 Féminin ☐ 2

2 Date et lieu de naissance

Né(e) le : jour mois année

à : commune (et arrondissement pour Paris, Lyon, Marseille)

département n° DOM pays pour l'étranger, territoire pour les TOM

3 Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ?

année

4 Quelle est votre nationalité ?

- **Française**
  - Vous êtes **né(e) français(e)** ☐ 1
  - Vous êtes **devenu(e) français(e)** (par exemple : par naturalisation, par déclaration, à votre majorité) ☐ 2
    - Indiquez votre nationalité à la naissance :
- **Étrangère** ☐ 3
  - Indiquez votre nationalité :

5 Êtes-vous inscrit(e) dans un établissement d'enseignement pour l'année scolaire en cours ?

***Y compris** apprentissage ou études supérieures.*

Oui ☐ 1 Non ☐ 2

- **Si oui, où est situé cet établissement d'enseignement ?**
- Dans la **commune où vous résidez** (ou dans le même arrondissement pour Paris, Lyon, Marseille) ☐ 1
- Dans une **autre commune** (ou un autre arrondissement) ☐ 2
  - Indiquez cette autre commune :

commune (et arrondissement pour Paris, Lyon, Marseille) département n° DOM

6 Où habitiez-vous le 1er janvier 2014 ?

Les enfants nés après cette date ne sont pas concernés.

- Dans le **même logement** que maintenant ☐ 1
- Dans un **autre logement** de la **même commune** (ou du même arrondissement pour Paris, Lyon, Marseille) ☐ 2
- Dans une **autre commune** (ou un autre arrondissement pour Paris, Lyon, Marseille) ☐ 3
  - Indiquez cette autre commune :

commune (et arrondissement pour Paris, Lyon, Marseille)

département n° DOM pays pour l'étranger, territoire pour les TOM

7 La suite du questionnaire s'adresse aux personnes de 14 ans ou plus.

8 Vivez-vous en couple ? Oui ☐ 1 Non ☐ 2

9 Êtes-vous ?

- Marié(e) ☐ 1
- Pacsé(e) ☐ 2
- En concubinage ou union libre ☐ 3
- Veuf(ve) ☐ 4
- Divorcé(e) ☐ 5
- Célibataire ☐ 6

10 Quel(s) diplôme(s) avez-vous ?

- Vous n'avez jamais été à l'école ou vous l'avez quittée avant la fin du primaire ☐ 01
- Aucun diplôme et scolarité interrompue à la fin du primaire ou avant la fin du collège ☐ 02
- Aucun diplôme et scolarité jusqu'à la fin du collège ou au-delà ☐ 03
- CEP (certificat d'études primaires) ☐ 11
- BEPC, brevet élémentaire, brevet des collèges, DNB ☐ 12
- CAP, BEP ou diplôme de niveau équivalent ☐ 13
- Baccalauréat général ou technologique, brevet supérieur, capacité en droit, DAEU, ESEU ☐ 14
- Baccalauréat professionnel, brevet professionnel, de technicien ou d'enseignement, diplôme équivalent ☐ 15
- BTS, DUT, Deug, Deust, diplôme de la santé ou du social de niveau bac+2, diplôme équivalent ☐ 16
- Licence, licence pro, maîtrise, diplôme équivalent de niveau bac+3 ou bac+4 ☐ 17
- Master, DEA, DESS, diplôme grande école niveau bac+5, doctorat de santé ☐ 18
- Doctorat de recherche (hors santé) ☐ 19

11 Quelle est votre situation principale ?

*Ne cochez qu'**une seule case**.*

- **Emploi** (salarié ou à votre compte, y compris aide d'une personne dans son travail) ⇨ *cochez puis passez en* 18 ☐ 1
- **Apprentissage** sous contrat ou **stage rémunéré** ⇨ *cochez puis passez en* 18 ☐ 2
- **Études** (élève, étudiant) ou **stage non rémunéré** ☐ 3
- **Chômage** (inscrit ou non au pôle emploi) ☐ 4
- **Retraite** ou **préretraite** (ancien salarié ou ancien indépendant) ☐ 5
- **Femme ou homme au foyer** ☐ 6
- **Autre situation** ☐ 7

12 Travaillez-vous actuellement ?

Si vous avez un emploi occasionnel ou de très courte durée, ou si vous êtes en apprentissage ou en stage rémunéré, cochez « Oui ». Si vous êtes en congé maladie ou de maternité, cochez « Oui ».

- Oui ⇨ *cochez puis passez en* 18 ☐ 1
- Non ⇨ *cochez puis passez en* 13 ☐ 2

Continuez page suivante et n'oubliez pas de signer →

13 Si vous ne travaillez pas actuellement, répondez aux questions 14 à 17.

14 Avez-vous déjà travaillé ?
- Oui ☐ 1
- Non ⇨ *cochez puis passez à la question* 17 ☐ 2

15 Étiez-vous :
- salarié(e) ou stagiaire rémunéré ? ☐ 1
- indépendant ou à votre compte ? ☐ 2
- Vous aidiez une personne dans son travail sans être rémunéré(e) ☐ 3

16 Quelle était votre profession principale ?

17 Cherchez-vous un emploi ?
- **Oui,** depuis moins d'un an. ☐ 1
- **Oui,** depuis un an ou plus ☐ 2
- **Non** ☐ 3

18 La suite du questionnaire s'adresse aux personnes qui travaillent actuellement.
*Si vous exercez plusieurs emplois, décrivez uniquement votre emploi **principal** aux questions 19 à 31.*

19 Quel est le nom de l'établissement qui vous emploie ou que vous dirigez ?
*Si vous êtes **intérimaire**, précisez le nom de l'établissement où vous faites votre mission. Si vous êtes **à votre compte**, inscrivez le nom de l'entreprise ou votre nom.*

20 Quelle est l'activité de cet établissement ?
***Soyez très précis** (par exemple : « RÉPARATION AUTOMOBILE »). S'il s'agit d'une **exploitation agricole**, précisez également l'orientation des productions (vigne, élevage de volailles, etc.).*

21 Quelle est l'adresse de votre lieu de travail ?
*Indiquez **l'endroit où vous commencez habituellement votre travail** (exemple : 18, boulevard Pasteur).*
*Si cet endroit n'est pas fixe, notez « **variable** ».*
*Si vous travaillez à votre domicile, notez « **à domicile** ».*
*Si vous travaillez chez un particulier, notez « **particulier** ».*

Est-ce dans la commune où vous résidez ?
(ou dans l'arrondissement pour Paris, Lyon, Marseille)
Oui ☐ 1 Non ☐ 2

Si non, indiquez la commune où vous travaillez :

commune (et arrondissement pour Paris, Lyon, Marseille)

département n° DOM pays pour l'étranger

22 Quel mode de transport principal utilisez-vous le plus souvent pour aller travailler ?
- Pas de déplacement ☐ 1
- Marche à pied (ou rollers, patinette) ☐ 2
- Vélo (y compris à assistance électrique) ☐ 3
- Deux-roues motorisé ☐ 4
- Voiture, camion ou fourgonnette ☐ 5
- Transports en commun ☐ 6

23 Occupez-vous votre emploi :
à temps complet ? ☐ 1 à temps partiel ? ☐ 2

24 Êtes-vous :
- indépendant ou à votre compte ? ☐ 1
- chef d'entreprise salarié, PDG, gérant(e) minoritaire de SARL ? ☐ 2
- salarié(e) ? ⇨ *cochez puis passez en* 27 ☐ 3
- Vous aidez une personne dans son travail sans être rémunéré(e) ☐ 4

25 Si vous êtes à votre compte ou chef d'entreprise combien de salariés employez-vous ?
Aucun ☐ 0 1 à 9 ☐ 1 10 ou plus ☐ 2

26 Si vous n'êtes pas salarié, quelle est votre profession ?
Soyez précis. Par exemple : « FLEURISTE » (et non « COMMERÇANT »).

27 La suite du questionnaire s'adresse aux salariés.

28 Quel est votre type de contrat ou d'emploi ?
- Emploi sans limite de durée, CDI (contrat à durée indéterminée), titulaire de la fonction publique ☐ 1
- Contrat d'apprentissage et de professionnalisation ☐ 2
- Placé par une agence d'intérim ☐ 3
- Stage rémunéré en entreprise ☐ 4
- Emploi aidé (contrat unique d'insertion, d'initiative emploi, d'accompagnement dans l'emploi, avenir, etc.) ☐ 5
- Autre emploi à durée limitée, CDD (contrat à durée déterminée), contrat court, saisonnier, vacataire, etc. ☐ 6

29 Dans votre emploi, êtes-vous :
- manœuvre, ouvrier spécialisé ? ☐ 1
- ouvrier qualifié ou hautement qualifié, technicien d'atelier ? ☐ 2
- technicien (non cadre) ? ☐ 3
- agent de catégorie B de la fonction publique ? ☐ 4
- agent de maîtrise, maîtrise administrative ou commerciale, VRP ? ☐ 5
- agent de catégorie A de la fonction publique ? ☐ 6
- ingénieur, cadre d'entreprise ? ☐ 7
- agent de catégorie C de la fonction publique ? ☐ 8
- employé (par exemple : de bureau, de commerce, de la restauration, de maison) ? ☐ 9

30 Quelle est votre profession principale ?
***Soyez précis.** Par exemple : « CAISSIÈRE » (et non « EMPLOYÉE »), « CHEF DE SERVICE CLIENTÈLE » (et non « CADRE »).*
***Si vous êtes agent de la fonction publique** d'État, territoriale ou hospitalière, indiquez votre grade (corps, catégorie, etc.).*

31 Dans votre emploi, quelle est votre fonction principale ?
- Production, exploitation, chantier ☐ 1
- Installation, réparation, maintenance ☐ 2
- Gestion, comptabilité ☐ 3
- Études, recherche ☐ 4
- Autre : *commerciale, secrétariat, logistique, etc.* ☐ 5

Merci pour votre participation

IMPRIMERIE NATIONALE 145 141
PEFC 10-31-1646 Cette entreprise a fait certifier ses chantiers de contrôle : papiforum.org

Vu l'avis favorable du Conseil national de l'information statistique, et en application de la loi n° 51-711 du 7 juin 1951 modifiée, cette enquête, reconnue d'**intérêt général et de qualité statistique**, est **obligatoire**. Les réponses sont protégées par le secret statistique et destinées à l'élaboration de statistiques sur la population et les logements.
Visa n° 2015A001EC du Ministre des finances et des comptes publics et du Ministre de l'économie, du redressement productif et du numérique, **valable** pour les années **2015 à 2019**
En application de la loi n° 2002-276 du 27 février 2002, l'enquête de recensement est placée sous la responsabilité de l'Insee et des communes ou des établissements publics de coopération intercommunale.
La loi n° 78-17 du 6 janvier 1978 modifiée garantit aux personnes enquêtées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des directions régionales de l'Insee.

Date : .....
Signature :

Tableau A4 – Bulletin individuel pour les centres pénitentiaires

145142_4EP_imprime_8_specimen_210x297 20/05/2014 10:30 Page1

Recensement de la population - 2015

Bulletin individuel – Établissement pénitentiaire

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Imprimé n° 8

Exemple : DUPAS, épouse MAURIN

Nom :

Prénom :

Identifiant de l'établissement

Cadre à remplir par l'enquêteur

1 Sexe Masculin ☐ 1 Féminin ☐ 2

2 Date et lieu de naissance

Né(e) le : jour mois année

à : commune (et arrondissement pour Paris, Lyon, Marseille)

département n° DOM pays pour l'étranger, territoire pour les TOM

3 Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ?

année

4 Quelle est votre nationalité ?

- **Française**
  - Vous êtes **né(e) français(e)** ☐ 1
  - Vous êtes **devenu(e) français(e)** (par exemple : par naturalisation, par déclaration, à votre majorité) ☐ 2
    - Indiquez votre nationalité à la naissance :
- **Étrangère** ☐ 3
  - Indiquez votre nationalité :

5 Êtes-vous ?

- Marié(e) ☐ 1
- Pacsé(e) ☐ 2
- En concubinage ou union libre ☐ 3
- Veuf(ve) ☐ 4
- Divorcé(e) ☐ 5
- Célibataire ☐ 6

6 Quel(s) diplôme(s) avez-vous ?

- Vous n'avez jamais été à l'école ou vous l'avez quittée avant la fin du primaire ☐ 01
- Aucun diplôme et scolarité interrompue à la fin du primaire ou avant la fin du collège ☐ 02
- Aucun diplôme et scolarité jusqu'à la fin du collège ou au-delà ☐ 03
- CEP (certificat d'études primaires) ☐ 11
- BEPC, brevet élémentaire, brevet des collèges, DNB ☐ 12
- CAP, BEP ou diplôme de niveau équivalent ☐ 13
- Baccalauréat général ou technologique, brevet supérieur, capacité en droit, DAEU, ESEU ☐ 14
- Baccalauréat professionnel, brevet professionnel, de technicien ou d'enseignement, diplôme équivalent ☐ 15
- BTS, DUT, Deug, Deust, diplôme de la santé ou du social de niveau bac+2, diplôme équivalent ☐ 16
- Licence, licence pro, maîtrise, diplôme équivalent de niveau bac+3 ou bac+4 ☐ 17
- Master, DEA, DESS, diplôme grande école niveau bac+5, doctorat de santé ☐ 18
- Doctorat de recherche (hors santé) ☐ 19

7 Avez-vous déjà travaillé ?

- Oui ☐ 1
- Non ☐ 2

8 Étiez-vous :

- salarié(e) ou stagiaire rémunéré ? ☐ 1
- indépendant ou à votre compte ? ☐ 2
- Vous aidiez une personne dans son travail sans être rémunéré(e) ☐ 3

9 Quelle était votre profession principale ?

Merci pour votre participation

SPECIMEN

Date : .....

Signature :

Vu l'avis favorable du Conseil national de l'information statistique, et en application de la loi n°51-711 du 7 juin 1951 modifiée, cette enquête, reconnue d'**intérêt général et de qualité statistique**, est **obligatoire**. Les réponses sont protégées par le secret statistique et destinées à l'élaboration de statistiques sur la population et les logements.
Visa n° 2015A001EC du Ministre des finances et des comptes publics et du Ministre de l'économie, du redressement productif et du numérique, **valable** pour les années **2015 à 2019**.
En application de la loi n° 2002-276 du 27 février 2002, l'enquête de recensement est placée sous la responsabilité de l'Insee et des communes ou des établissements publics de coopération intercommunale.
La loi n° 78-17 du 6 janvier 1978 modifiée garantit aux personnes enquêtées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des directions régionales de l'Insee.

IMPRIMERIE NATIONALE 145 142

PEFC 10-31-1544 Cette entreprise a fait certifier sa chaîne de contrôle : pefc-france.org

A5 - Signification des modalités des variables auxiliaires

Catégorie de la communauté :

- 1 : Service de moyen ou de long séjour d'un établissement public ou privé de santé, établissement social de moyen ou de long séjour, maison de retraite, foyer ou résidence sociale ou assimilé
- 2 : Communauté religieuse
- 3 : Gendarmerie
- 4 : Établissement hébergeant des élèves ou des étudiants, y compris établissement militaire d'enseignement
- 5 : Établissement pénitentiaire
- 6 : Établissement social de court séjour
- 7 : Autre catégorie de communauté

Sous-catégorie de la communauté :

- 11 : Maison de retraite, hospice
- 12 : Foyer Adoma
- 13 : Autre foyer de travailleurs
- 14 : Service de moyen ou de long séjour
- 21 : Communauté religieuse
- 31 : Gendarmerie
- 32 : Quartier, base ou camp militaire
- 41 : Cité universitaire
- 42 : Autre internat
- 51 : Établissement pénitentiaire
- 61 : Établissement social de court séjour
- 71 : Autre communauté

Tranche d'aire urbaine 2010 : Ce code indique la tranche de taille de l'aire urbaine à laquelle appartient la commune au recensement de la population 2008.

- 01 : Commune hors aire urbaine ou appartenant à une aire urbaine de moins de 15 000 habitants
- 02 : Commune appartenant à une aire urbaine de 15 000 à 19 999 habitants
- 03 : Commune appartenant à une aire urbaine de 20 000 à 24 999 habitants
- 04 : Commune appartenant à une aire urbaine de 25 000 à 34 999 habitants
- 05 : Commune appartenant à une aire urbaine de 35 000 à 49 999 habitants
- 06 : Commune appartenant à une aire urbaine de 50 000 à 99 999 habitants
- 07 : Commune appartenant à une aire urbaine de 100 000 à 199 999 habitants
- 08 : Commune appartenant à une aire urbaine de 200 000 à 499 999 habitants
- 09 : Commune appartenant à une aire urbaine de 500 000 à 9 999 999 habitants
- 10 : Commune appartenant à l'aire urbaine de Paris

Catégorie de la commune dans le zonage en aires urbaines 2010

- 111 : Commune appartenant à un grand pôle (10 000 emplois ou plus)
- 112 : Commune appartenant à la couronne d'un grand pôle
- 120 : Commune multipolarisée des grandes aires urbaines

- 211 : Commune appartenant à un moyen pôle (5 000 à moins de 10 000 emplois)
- 212 : Commune appartenant à la couronne d'un moyen pôle
- 221 : Commune appartenant à un petit pôle (de 1 500 à moins de 5 000 emplois)
- 222 : Autres communes

Effectif collecté dans la communauté

- 0 : moins de 20 personnes ont été enquêtées
- 1 : entre 20 et 75 personnes ont été enquêtées
- 2 : entre 75 et 205 personnes ont été enquêtées
- 3 : entre 205 et 385 personnes ont été enquêtées
- 4 : entre 385 et 880 personnes ont été enquêtées
- 5 : plus de 880 personnes ont été enquêtées

Tranche d'âge de l'individu recensé :

- 0 : entre 0 et 17 ans
- 1 : entre 18 et 30 ans
- 2 : entre 30 et 50 ans
- 3 : entre 50 et 74 ans
- 4 : 75 ans et plus

A6 - Les communes « test » pour le critère de qualité n° 2

Code	Commune	Population municipale RP2012	Population en communauté RP2012
04070	Digne-les-Bains	16 844	1 080
04088	Forcalquier	4 775	328
04112	Manosque	22 099	801
05023	Briançon	12 301	1018
05046	Embrun	6 143	533
05061	Gap	40 761	1840
06004	Antibes	75 568	1 307
06027	Cagnes-sur-Mer	46 686	464
06029	Cannes	73 603	1 794
06030	Le Cannet	43 115	923
06048	Contes	7187	506
06069	Grasse	51 021	2 252
06083	Menton	29 073	603
06084	Mouans-Sartoux	10 214	304
06085	Mougins	17 884	500
06088	Nice	343 629	9 235
06152	Valbonne	12 619	623
13001	Aix-en-Provence	141 148	8 031
13201	Marseille 1er Arrondissement	38 972	571
13202	Marseille 2e Arrondissement	24 274	375
13203	Marseille 3e Arrondissement	44 666	1 553
13204	Marseille 4e Arrondissement	47 869	534
13205	Marseille 5e Arrondissement	46 460	1 074
13206	Marseille 6e Arrondissement	41 838	493
13208	Marseille 8e Arrondissement	78 065	1 870
13209	Marseille 9e Arrondissement	74 130	4 901
13210	Marseille 10e Arrondissement	54 279	1 363
13211	Marseille 11e Arrondissement	56 835	1 868
13212	Marseille 12e Arrondissement	60 528	1 970
13213	Marseille 13e Arrondissement	90 122	2 813
13214	Marseille 14e Arrondissement	60 949	1 245
13215	Marseille 15e Arrondissement	80 808	1 642
13216	Marseille 16e Arrondissement	17 101	304
21038	Auxonne	7 771	625
31555	Toulouse	453 317	11 419
33063	Bordeaux	241 287	5 453
33140	Créon	4 333	106
33243	Libourne	23 736	726
34172	Montpellier	268 456	9 448
35238	Rennes	209 860	11 209
59043	Bailleul	14 564	672

Code	Commune	Population municipale RP2012	Population en communauté RP2012
67482	Strasbourg	274 394	10 805
69046	Charly	4 470	103
70550	Vesoul	15 637	1 098
75104	Paris 4e Arrondissement	27 769	444
75105	Paris 5e Arrondissement	60 179	2 670
75106	Paris 6e Arrondissement	43 224	2 592
75107	Paris 7e Arrondissement	57 092	1 193
75108	Paris 8e Arrondissement	38 749	676
75109	Paris 9e Arrondissement	59 474	409
75110	Paris 10e Arrondissement	94 474	993
75111	Paris 11e Arrondissement	155 006	1 569
75112	Paris 12e Arrondissement	144 925	3 882
75113	Paris 13e Arrondissement	182 386	6 187
75114	Paris 14e Arrondissement	141 102	9 209
75115	Paris 15e Arrondissement	238 190	4 327
75116	Paris 16e Arrondissement	167 613	2 386
75117	Paris 17e Arrondissement	170 156	1 340
75118	Paris 18e Arrondissement	201 374	2 259
75119	Paris 19e Arrondissement	186 116	2 769
75120	Paris 20e Arrondissement	197 311	3 485
76114	Bolbec	11 753	424
76165	Caudebec-lès-Elbeuf	10 030	340
76217	Dieppe	30 632	894
76259	Fécamp	19 262	577
76322	Le Grand-Quevilly	24 563	385
76331	Grugny	915	493
76351	Le Havre	173 142	4 245
76451	Mont-Saint-Aignan	19 798	1 718
76484	Oissel	11 445	692
76498	Le Petit-Quevilly	22 089	631
76540	Rouen	111 557	3 456
76563	Saint-Aubin-Routot	1 902	681
76575	Saint-Étienne-du-Rouvray	28 616	831
76681	Sotteville-lès-Rouen	28 622	776
76758	Yvetot	11 644	483
77459	Sourdun	1 436	322
91286	Grigny	27 713	87
92002	Antony	61 624	1 481

Code	Commune	Population municipale RP2012	Population en communauté RP2012
95018	Argenteuil	104 962	1 838
95051	Beauchamp	8 753	313
95063	Bezons	28 172	313
95127	Cergy	60 528	873
95176	Cormeilles-en-Parisis	23 369	581
95203	Eaubonne	24 714	518
95219	Ermont	27 352	388
95277	Gonesse	26 343	745
95424	Montigny-lès-Cormeilles	20 018	399
95428	Montmorency	20 842	727
95476	Osny	16 366	1 271
95487	Persan	11 233	336
95500	Pontoise	30 164	1 406
95572	Saint-Ouen-l'Aumône	23 470	522
95582	Sannois	26 559	564
95585	Sarcelles	57 499	576
95607	Taverny	26 094	415
95680	Villiers-le-Bel	27 496	516