AMÉLIORATION DU REDRESSEMENT DE LA NON-RÉPONSE DES COMMUNAUTÉS DANS LE RECENSEMENT

Lise REYNAERT[1]*,* Division « Méthodes et Traitements des Recensements », Direction des Statistiques Démographiques et Sociales, INSEE

Résumé

En France, environ 9 millions de personnes sont recensées chaque année, toutes catégories de population confondues. Ces catégories distinguent : la population des ménages, celle des habitations mobiles ou sans abri et celle des individus en communautés à laquelle s'intéresse cette étude.
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, visibles à un niveau communal : la première particularité est le lien entre le comportement de réponse et l'ordre du lot de saisie, entraînant des redressements massifs de non-répondants successifs par un unique donneur ; la seconde est l'hétérogénéité des populations en communauté, provoquant 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 : 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 envisagerons des méthodes de redressement alternatives, qui pourraient être mises en place en cas de refonte à moyen terme des applications du recensement. Pour cela, nous testerons et comparerons le hot-deck par classe, le hot-deck métrique à partir d'une mesure de similarité basée sur le V de Cramer et la technique de repondération.

Abstract

The population census distinguishes several components of population : people residing in ordinary residences, those who lives in mobile residences and homeless people, and those living in communities like boarding schools. For the latter, the nonresponse is adjusted by a sequential hot-deck, 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 doesn'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 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, redressement de la non-réponse

[1] lise.reynaert@insee.fr

Table des matières

Introduction

Le recensement de la population distingue différentes catégories de population[2] : celle des ménages (97,7 %), celle des habitations mobiles ou sans abri (0,2 %) et celle des individus en communautés (2,1 %) à laquelle s'intéresse cette étude. 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 % 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 %) ;
- les cités universitaires (5,5 %) ;
- les établissements pénitentiaires (3,9 %) ;
- les établissements militaires (3,3 %) ;
- les communautés religieuses (2 %) ;
- 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[3].

Les opérations post-collecte s'effectuent simultanément pour les individus en 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. La liste des critères pour les variables de cette étude figure en annexe 3.

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 actualisés en fin décembre de chaque année (cf. encadré). La méthode se révèle 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 en communautés. 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, provoquant 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.

[2] Le décret n°2003-485 du 5 juin 2003 fixe les catégories de population et leur composition.

[3] Un répertoire des communautés permet à l'Insee de gérer la collecte. Ce répertoire recense toutes les structures répondant à la définition de communauté et maintient à jour les informations qui y sont associées. Il a été constitué à partir des informations du recensement 1999 et est mis à jour grâce à des sources administratives et des retours terrain.

- 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 limités de redressements aberrants visibles à un niveau communal.

L'objectif de cette étude 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 de hot-deck séquentiel, en restant dans la logique des traitements standards du RP. 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, qui pourraient être mises en place en cas de refonte à moyen terme des applications du recensement.

Encadré : les principaux enjeux 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.

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 s'assurer de la compatibilité entre les changements envisagés et le fait que les résultats du recensement se basent sur les cinq enquêtes annuelles les plus récentes.

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

- La collecte s'effectue en janvier-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 ;

- Les résultats statistiques sont élaborés de janvier à avril de l'année N+1 et diffusés en juin N+1.

1 Éléments préalables à l'étude des redressements

Dans cette étape préalable, nous mettrons en évidence les particularités de la non-réponse dans les communautés. Nous ferons également le point sur les variables auxiliaires permettant modéliser le comportement de non-réponse et/ou les variables d'intérêt. Enfin, nous définirons les critères permettant d'apprécier la qualité des différentes méthodes de redressement testées.

1.1 Champ d'étude et particularité de la non-réponse dans les communautés

Nous utilisons les réponses des 1 604 055 individus en communauté collectées entre 2009 et 2013. Cette analyse à partir des réponses cumulées durant cinq années de collecte successives est menée spécifiquement pour cette étude, et les résultats obtenus diffèrent des redressements actuels menés en production car ces derniers sont effectués indépendamment par année de collecte. Nous limitons cette étude aux dix variables[4] 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[5] : 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[5] : la situation principale et l'indicatrice du fait de travailler actuellement.

L'ensemble des variables du questionnaire et leurs modalités figurent en annexe 1.

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 1). Cela s'explique par les consignes de collecte : 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 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 très nettement plus élevée dans les communautés que dans les ménages ; par exemple, la situation principale[6] n'est pas renseignée pour 28,9 % des individus en communauté, contre seulement 4,9 % des individus en ménage.

Tableau 1 : taux de non-réponse pour les variables d'intérêts en distinguant les individus en communauté des individus en ménage

	Taux de non réponse (en %)	
Variable	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%
Etat 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

[4] Ces variables servent à l'exploitation dite « principale » du recensement. L'exploitation principale traite toutes les informations pouvant être codifiées aisément après la saisie des questionnaires.

[5] Ces informations ne sont pas collectées dans les centres de détention.

[6] Les modalités de la variable « situation principale » sont ; emploi, apprentissage, études, chômage, retraite, homme ou femme au foyer ou autre situation.

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 2).

Tableau 2 : probabilité empirique de non-réponse sachant que l'individu précédent est non-répondant pour la variable concernée

Variable	Probabilité empirique de non-réponse	
	Communautés	Ménages***
Sexe	0,27	0,14
Année de naissance	0,83	0,47
Nationalité	0,82	0,30
Inscription dans un établisement d'étude	0,92	0,39
Indicateur de résidence antérieure	0,87	0,26
Vie en couple**	0,87	0,08
Etat 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

1.2 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 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.

Les modalités de ces variables auxiliaires se trouvent en annexe 2.

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 Khi-deux. 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 3 pour la variable « nationalité »). La non-réponse est notamment très présente dans les foyers de travailleurs, les cités universitaires et les établissements sociaux de court séjour.

Tableau 3 : dépendance entre la sous-catégorie de communauté et le fait de répondre à la question « nationalité »

Effectif de population = 1,6 million d'individus
Khi-deux = 121 625
P-value < 0.0001

		Ecart (%) des effectifs observés par rapport à la situation hypothétique d'indépendance	
Sous-catégorie de communauté	Effectifs de population	Réponse	Non-réponse
Maison de retraite, hospice	513 994	2%	-10%
Foyer sonacotra	55 901	-35%	168%
Autre foyer de travailleurs	71 554	-18%	84%
Service 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%
Etablissement pénitentiaire	61 952	1%	-7%
Etablissement social de court séjour	7 762	-23%	112%
Autre communauté	763	-11%	51%

A noter qu'à partir de la collecte 2015, la sous-catégorie « services de moyen ou de long séjour » a été ventilé en cinq postes plus homogènes[7], permettant ainsi d'accroître le pouvoir explicatif de la variable « sous-catégorie de communauté ».

1.3 Définition des critères de qualité

Afin d'appréhender la qualité des différentes méthodes testées, nous définissons quatre critères :

Critère n°1 : la méthode testée améliore-t-elle les cas de redressement aberrants de la méthode actuelle ? L'effectif de ces cas est limité, mais ces derniers restent visibles pendant au moins cinq ans. Ci-dessous 4 cas problématiques, parmi la dizaine de cas majeurs identifiés dans les enquêtes 2009 à 2013 :

- Cas des cadres nonagénaires dans les Hauts-de-Seine : lors de la collecte 2009, une vingtaine d'individus en maison de retraite qui avaient uniquement renseigné leur date de naissance et leur sexe ont été redressés par un cadre habitant en foyer de travailleur de même sexe (masculin) et de même tranche d'âge (65 ans et plus), créant ainsi des individus cadres de plus de 90 ans.

- Cas des déplacements domicile-travail 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 des Yvelines : lors de l'enquête 2010, une centaine d'individus en foyer de demandeur 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 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

[7] Les services de long et moyen séjour ont été découpés en :
- structures pour personnes nécessitant des soins médicaux (enfants, adultes) ;
- structures pour les enfants handicapés ;
- structures pour les adultes handicapés ;
- structures d'aide sociale à l'enfance et de protection judiciaire pour enfants et jeunes majeurs ;
- structures adultes et familles nécessitant un accompagnement social et psychologique.

redressés à partir d'un individu en maison de retraite né en 1920, déformant ainsi la pyramide des âges de la commune.

Nous analyserons ce premier critère en nous assurant que les redressements donnent des caractéristiques imputées plus adaptées aux non-répondants.

Critère n°2 : quelle est la qualité prédictive de la méthode de redressement en cas de non-réponse diffuse ? 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 suivant un tirage aléatoire indépendant des caractéristiques des répondants, puis le redressement de la non-réponse est réalisé. Enfin, les variables d'intérêt avant et après l'imputation sont comparées. Cette opération est réalisée trente fois pour calculer un taux moyen d'individus bien classés.

Critère n°3 : la méthode de redressement induit-elle une distorsion dans la distribution des variables en cas de non-réponse massive et concentrée ? pour tester les limites de la méthode de redressement, nous générerons trente fois une non-réponse massive dans 50 % des communautés de 18 communes tests, choisies -arbitrairement- pour leur hétérogénéité du point de vue de la population en communauté.

Pour quantifier la distorsion provoquée par le redressement, 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 tests ;
- *J* le nombre de modalités de la variable ;
- $Ftj\_répondants$ la proportion de la modalité *j* parmi les répondants de la commune *t* ;
- $Ftj\_imputation$ 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 mesure de distorsion *D* comme l'écart moyen par rapport à la variable avant imputation :

$$D = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{J}\sum_{j=1}^{J}.\left|Ftj\_imputation - Ftj\_répondants\right|$$

Critère n°4 : la méthode de redressement limite-t-elle le nombre maximal de don par individu répondant ? Nous souhaitons éviter le redressement massif de communautés entières non-répondantes par un unique donneur. Pour ce critère aucune simulation de données n'est réalisée, 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 quantiles.

2 Amélioration de la méthode actuelle par hot-deck séquentiel

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

2.1 Les principes de l'imputation par hot-deck séquentiel

La méthode par hot-deck séquentiel « classique » consiste à déterminer, parmi les variables renseignées par les répondants et les non-répondants, les plus corrélées à la variable à imputer à partir des observations correspondant aux seuls répondants. Le fichier est ensuite trié selon ces variables, et pour chaque valeur manquante, la modalité de la variable prise par le répondant précédent est imputée. Une valeur initiale est définie, au cas où la première observation du fichier trié est manquante.
Dans le cas du recensement, l'ordre de tri est fixé une fois pour toutes pour l'ensemble des variables : le fichier est trié selon l'identifiant du recensement (département, commune, rang d'adresse, rang d'individus), permettant ainsi de limiter les temps de traitement. Pour pallier à ce tri fixe, des contraintes d'imputation sont ajoutées ; par exemple, on cherche le dernier donneur en communauté de même tranche d'âge ou de même sexe. Par ailleurs, le redressement s'effectue variable par variable, dans un ordre logique prédéfini (cf. tableau 4). 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. Cette méthode a l'avantage de prendre en compte les éventuelles relations de dépendance entre les différentes variables à imputer.

Tableau 4 : ordre des variables pour le redressement

1	Sexe
2	Age
3	Etat 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

Nous pouvons améliorer le hot-deck séquentiel actuel à deux égards :

- en trouvant des donneurs plus proches des non-répondants via le regroupement de l'ensemble des individus en communauté dans un unique lot de saisie et la redéfinition les contraintes d'imputation ;
- en élargissant le champ des donneurs potentiels avec l'ajout d'un aléa d'imputation.

2.2 Première proposition : ajout de nouvelles contraintes d'imputation

Une première amélioration de la méthode actuelle consiste à redéfinir les contraintes d'imputation, à défaut de pouvoir modifier l'ordre de tri du fichier.

2.2.1 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 déjà imputées). On utilise pour ce faire les observations correspondant aux répondants et un modèle polytomique non ordonné.

La probabilité d'observer, pour l'individu *i*, la modalité *j* de la variable à expliquer s'écrit :

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

où les x_i sont les caractéristiques de l'individu *i* prises en compte dans le modèle et les β_j et β_h les paramètres estimés.
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 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. Nous ajoutons une contrainte qui consiste à garder le département comme variable explicative (compte tenu du tri géographique qui joue de toutes façons dans la recherche du donneur, nous retenons systématiquement comme variable explicative le code du département, afin de déterminer si les autres variables explicatives apportent de l'information supplémentaire par rapport à l'information départementale).
En procédant pas à pas, on construit, pour chaque variable du questionnaire, un modèle qui inclut obligatoirement la variable département (cf. tableau 5), et qui s'écrit par exemple pour le diplôme, en prenant la modalité 3 « Pas de diplôme, mais scolarité au-delà du collège » :

$$\text{Ln}\left(\frac{P(\text{dipl} = 1/\text{xi})}{P(\text{dipl} = 3/\text{xi})}\right) = \beta 01 + \beta 11.\text{dep\_code}i + \beta 21.\text{id\_scat}i + \beta 31.\text{tranche\_age}i + \beta 41.\text{matr}i + \beta 51.\text{inscr}i$$

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 ».

Tableau 5 : résultats de la sélection des variables à utiliser comme contrainte d'imputation

	Contraintes d'imputation	
Variable à imputer	**Hot-deck séquentiel avec redéfinition des contraintes d'imputation**	**Hot-deck séquentiel actuel**
Sexe	Sous catégorie de communauté	
	Effectif de la communauté	
Age	Sous catégorie de communauté	
	Effectif de la communauté	
	Sexe	
Etat matrimonial légal	Sous catégorie de communauté	
	Effectif de la communauté	
	Sexe	
	Tranche d'âge	
Vie en couple	Sous catégorie de communauté	Tranche d'âge
	Effectif de la communauté	Etat matrimonial
	Tranche d'âge	Sexe
	Etat matrimonial légal	
Indicatrice de nationalité	Sous catégorie de communauté	
	Effectif de la communauté	
	Tranche d'âge	
Inscription dans un établissement d'enseignement	Sous catégorie de communauté	Tranche d'âge
	Tranche d'âge	
Situation principale	Sous catégorie de communauté	Indicatrice "travaille actuellement"
	Inscription dans un établissement d'enseignement	Sexe
	Tranche d'âge	Tranche d'âge
Diplômes obtenus	Sous catégorie de communauté	Tranche d'âge
	Tranche d'âge	Indicatrice de nationalité

Les modalités de ces différentes variables se trouvent en annexe 2.

2.2.2 Résultats

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 diffuse de non-réponse) : le taux de bien classés de la nouvelle méthode n'est pas significativement meilleur que celui de la méthode actuelle (cf. tableau 6). 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 6 : le taux moyen de bien classés du hot-deck séquentiel avec redéfinition des contraintes est similaire au hot-deck actuel

Variable à imputer	Taux de bien classés (en %)	
	Hot-deck séquentiel avec redéfinition des contraintes d'imputation	Méthode actuelle simulée
Sexe	70,9%	70,8%
Age	76,3%	75,9%
Etat matrimonial légal	76,6%	76,3%
Indicatrice "Vie en couple"	95,9%	95,9%
Indicatrice de nationalité	89,3%	89,7%
Indicatrice travaille actuellement	91,2%	91,00%
Inscription dans un établissement d'enseignement	96,8%	96,2%
Situation principale	91,5%	89,5%
Indicateur de résidence antérieure	67,2%	67,2%
Diplômes obtenus	49,7%	50,0%

Critère n°3 (simulation d'une non-réponse massive et concentrée dans les communes test) : le hot-deck avec redéfinition des contraintes diminue significativement la distorsion de la distribution des variables, par rapport au hot-deck actuel (cf. tableau 7).

Tableau 7 : mesure de distorsion des variables en cas de non-réponse massive et concentrée

	Méthode actuelle simulée	Hot-deck avec redéfinition des contraintes
Tranche d'âge	14,0	5,0
Situation principale	7,5	4,0
Situation matrimoniale	10,1	6,0
Diplôme	7,4	6,2

Critère n°4 (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 10). 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 2900 individus de 8 cités universitaires se situant à la suite dans le lot de saisie.

2.3 Deuxième proposition : ajout d'un aléa d'imputation permettant de réduire le nombre de dons

Afin de limiter le nombre de dons par individu, nous proposons d'ajouter un aléa dans le processus d'imputation. 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 3 répondants qui précèdent dans le fichier et qui correspondent aux contraintes d'imputation.
Les résultats de l'imputation pour les deux premiers critères sont similaires à ceux du hot-deck séquentiel avec redéfinition des contraintes. Le taux de bien classés en présence de non-réponse diffuse (critère n°2) est globalement similaire à celui du hot-deck séquentiel actuel et à celui avec redéfinition des contraintes (cf. tableau 8).
L'ajout de l'aléa ne réduit pas de manière significative 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°3, cf. tableau 9). Il permet en revanche de limiter légèrement le nombre maximum de dons par individus, même si ce dernier reste tout de même très élevé (cf. tableau 10) : 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 8 : pour une non-réponse diffuse, le taux moyen de bien classés du hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation est dans l'ensemble similaire au hot-deck actuel

	Taux de bien classés (en %)	
Variable à imputer	**Hot-deck séquentiel avec redéfinition des contraintes d'imputation et aléa d'imputation**	**Méthode actuelle simulée**
Sexe	70,0%	70,8%
Age	74,8%	75,9%
Etat matrimonial légal	74,8%	76,3%
Indicatrice "Vie en couple"	95,0%	95,9%
Indicatrice de nationalité	88,9%	89,7%
Indicatrice "travaille actuellement"	90,9%	91,0%
Inscription dans un établissement d'enseignement	96,5%	96,2%
Situation principale	90,7%	89,5%
Indicateur de résidence antérieure	67,3%	67,2%
Diplômes obtenus	33,0%	50,0%

Tableau 9 : mesure de distorsion des variables en cas de non-réponse massive et concentrée sur les différents hot-decks séquentiels testés

	Méthode actuelle simulée	**Hot-deck avec redéfinition des contraintes**	**Hot-deck avec redéfinition des contraintes et aléa d'imputation**
Tranche d'âge	14,0	5,0	4,5
Situation principale	7,5	4,0	4,4
Situation matrimoniale	10,1	6,0	5,8
Diplôme	7,4	6,2	5,1

Tableau 10 : distribution du nombre de dons par individu pour les hot-decks séquentiels

			Nombre de dons		
	Effectif des non-répondants	**Distribution**	**Hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation**	**Hot-deck séquentiel avec redéfinition des contraintes**	**Hot-deck séquentiel actuel**
Sexe	6106	Maximum	26	56	56
		99 ème quantile	5	6	6
		90ème quantile	3	3	3
Age	32259	Maximum	436	872	502
		99 ème quantile	58	103	120
		90ème quantile	18	21	20
Etat matrimonial légal	429077	Maximum	1 596	3 265	2 706
		99 ème quantile	61	115	146
		90ème quantile	21	31	29
Couple	422649	Maximum	2 543	4 640	3 839
		99 ème quantile	71	113	188
		90ème quantile	25	34	47
Indicatrice de nationalité	276904	Maximum	1 546	2 975	2 703
		99 ème quantile	55	81	112
		90ème quantile	18	19	14
Inscription dans un établissement d'enseignement	507355	Maximum	1 589	2 959	2 370
		99 ème quantile	124	187	194
		90ème quantile	47	57	62
Situation principale	454578	Maximum	1 964	3 683	1 985
		99 ème quantile	102	153	138
		90ème quantile	37	45	43
Indicateur de résidence antérieure	466565	Maximum	2 160	3 184	2 370
		99 ème quantile	60	90	128
		90ème quantile	19	22	27
Diplômes obtenus	830023	Maximum	1 550	2 911	2 375
		99 ème quantile	104	170	187
		90ème quantile	39	25	23

2.4 Troisième proposition : regrouper les individus en communauté 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[8], 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 à la suite dans un lot de saisie sont susceptibles d'être très différents, car très éloignés géographiquement ; 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 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 travailleur dans le lot de saisie.

2.5 Bilan : les améliorations possibles du hot-deck actuel

La méthode actuelle peut-être améliorée sensiblement du point de vue de nos critères de qualité grâce à trois propositions :

Proposition n°1 : 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°2 : ajouter un aléa d'imputation permettant de limiter les redressements massifs de communautés entières par un unique individu (critères n°3 et 4).

[8] C'est-à-dire que les individus d'une commune se trouvent dans le même lot de saisie

Proposition n°3 : 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 (critères n°1 et 2).

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.

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

Dans cette troisième partie, nous mettrons en évidence que le hot-deck par classe et le hot-deck métrique pourraient être envisagés 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.

3.1 Le hot-deck par classe

3.1.1 Les principes de l'imputation par hot-deck par classe

La méthode du hot-deck par classe consiste à remplacer une valeur manquante par la valeur observée pour un individu répondant choisi au hasard à l'intérieur de la classe à laquelle appartient le receveur. Les classes sont définies de manière à être homogènes par rapport à la probabilité de réponse et par rapport à la réponse donnée à la variable d'intérêt. 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 de l'imputation.

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. Cette méthode a l'avantage de prendre en compte les éventuelles relations de dépendance entre les différentes variables à imputer.

3.1.2 Création de classes homogènes de repondération

Nous utiliserons la méthode du score pour former nos classes d'imputation. Cette méthode nous conduira ainsi, 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 de la robustesse au modèle si un des scores est mal spécifié.

Pour chaque variable du questionnaire, nous procédons ainsi :

- nous estimons les probabilités de réponse p_i grâce à un modèle logistique ;
- nous estimons les réponses à la question r_i grâce à un modèle polytomique non-ordonné ;
- nous utilisons un algorithme de classification de type « k-means »[9] afin d'obtenir des classes homogène par rapport aux scores p_i et r_i.

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 est 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 11)

[9] Cette méthode a l'avantage d'être efficace et très rapide. 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.

Tableau 11 : coefficient de corrélation résultant d'une analyse de la variance entre le score p_i 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

3.1.3 Résultats

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 diffuse de non-réponse) : le taux de bien classés du hot-deck par classe est sensiblement inférieur à celui de la méthode actuelle pour l'ensemble des variables du questionnaire (cf. tableau 12). 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.

Tableau 12 : le taux moyen de bien classé du hot-deck par classe par rapport à celui de la méthode actuelle

	Taux de bien classés (en %)	
Variable à imputer	Hot-deck par classe	Méthode actuelle simulée
Sexe	59,0%	70,8%
Age	66,0%	75,9%
Etat matrimonial légal	67,3%	76,3%
Indicatrice "Vie en couple"	62,5%	95,9%
Indicatrice de nationalité	82,7%	89,7%
Indicatrice "travaille actuellement"	85,5%	91,2%
d'enseignement	89,7%	96,2%
Situation principale	70,0%	89,5%
Indicateur de résidence antérieure	51,0%	67,2%
Diplômes obtenus	35,0%	50,0%

Critère n°3 (simulation d'une non-réponse massive et concentrée dans des communes test) : le hot-deck par classe réduit fortement la distorsion par rapport à la méthode actuelle (cf. tableau 13). Il améliore plus modérément la distorsion par rapport au hot-deck séquentiel avec redéfinition des contraintes.

Tableau 13 : comparaison de la mesure de distorsion des variables en cas de non-réponse massive et concentrée entre le hot-deck par classe et la méthode actuelle

	Méthode actuelle simulée	Hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation	Hot-deck par classe
Tranche d'âge	14,0	4,5	2,5
Situation principale	7,5	4,4	3,8
Situation matrimoniale	10,1	5,8	4,4
Diplôme	7,4	5,1	4,1

Critère n°4 (nombre maximal de dons) : le hot-deck par classe permet d'éviter les donneurs multiples, de part le tirage aléatoire au sein de classes relativement vastes.

3.3 Bilan sur le hot-deck par classe

Le hot-deck par classe permet de résoudre le cas des redressements aberrants (critère n°1) et de limiter le nombre de dons multiples (critère n°4). Toutefois, le taux de bien classés en cas de non-réponse diffuse est inférieure à celle du hot-deck séquentiel (critère n°2) ; cela est dû au fait que certaines sous-catégories de communautés contiennent des individus aux caractéristiques très hétérogènes. La décision de redécouper la sous-catégorie « services de long et moyen séjour » à partir de la collecte 2015 est susceptible d'améliorer les résultats de ce hot-deck.
Le hot-deck par classe 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°3). Il améliore plus modérément la distorsion par rapport au hot-deck séquentiel avec redéfinition des contraintes.

3.2 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.

3.2.1 Les principes de l'imputation par 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 variables auxiliaires et renseignées disponibles ; elle doit être définie de façon à respecter la corrélation entre ces variables et la variable d'intérêt, 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 variables précédemment imputées de manière à conserver le lien entre les différentes variables du questionnaire. Ainsi, pour chaque variable, le champ des donneurs potentiel 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.

3.2.2 Une mesure de similarité basée sur le V de Cramer

3.2.2.1 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[10] 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.

[10] Le V de Cramer est une mesure de la liaison entre deux variables qualitatives dérivée du khi-deux. Contrairement à ce dernier, le V de Cramer présente l'avantage de ne pas dépendre du nombre d'observations ni du nombre de modalités.
Le V de Cramer entre deux variables qualitatives X et Z est défini ainsi :

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

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

La mesure de similarité entre le receveur et le donneur potentiel 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 ;
- ω_j le poids de la variable auxiliaire j au sens du V de Cramer ;
- $\delta_{j,rd}$ vaut 0 si la variable auxiliaire 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 14 : pondérations affectées aux 6 variables auxiliaires pour chacune des variables du questionnaire

V de Cramer	Sous-catégorie de communauté	ZEAT	Effectif de la communauté	tranche d'aire urbaine 2010	catégorie d'aire urbaine 2010	Indicatrice "première collecte"
sexe	0,46	0,09	0,23	0,09	0,09	0,04
Année de naissance	0,46	0,08	0,18	0,10	0,10	0,08
Etat matrimonial légal	0,48	0,07	0,17	0,08	0,10	0,09
Indicatrice "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 dans un établissement d'enseignement	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
Indicateur de résidence antérieure	0,39	0,11	0,17	0,10	0,10	0,14
diplômes obtenus	0,34	0,09	0,22	0,12	0,13	0,11

3.2.2.2 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 ;
- ω_j le poids de la variable auxiliaire j au sens du V de Cramer ;
- ω_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 15 : les poids associés aux variables auxiliaires et variables du questionnaire pour la variable « situation principale » :

variable	pondération
Sous-catégorie de communauté	0,11
ZEAT	0,01
Effectif de la communauté	0,04
tranche d'aire urbaine 2010	0,02
catégorie d'aire urbaine 2010	0,02
Indicatrice "première collecte"	0,03
sexe	0,07
Tranche d'âge	0,12
Nationalité	0,02
Inscription dans un établissement d'enseignement	0,18
Indicateur de résidence antérieure	0,02
Indicatrice "Vie en couple"	0,03
Etat matrimonial légal	0,09
diplômes obtenus	0,07
situation principale	0,00
Indicatrice "travaille actuellement"	0,17

3.2.2 Inconvénient pratique majeur de la méthode et alternative

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 choisir celui qui la minimise. 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).

Un procédé alternatif 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.

3.2.3 Résultats

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 redressement aberrants, grâce à l'importance de la sous-catégorie de communauté dans la mesure de similarité.

Critère n°2 (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 16).

Tableau 16 : le taux moyen de bien classé du hot-deck métrique en cas de non-réponse diffuse par rapport à celui de la méthode actuelle

	Taux de bien classés (en %)	
Variable à imputer	**Hot-deck métrique**	**Méthode actuelle simulée**
Sexe	69,2%	70,8%
Age	72,0%	75,9%
Etat matrimonial légal	76,6%	76,3%
Indicatrice "Vie en couple"	94,4%	95,9%
Indicatrice de nationalité	87,9%	89,7%
Indicatrice "travaille actuellement"	86,1%	91,0%
Inscription dans un établissement d'enseignement	96,8%	96,2%
Situation principale	96,8%	89,5%
Indicateur de résidence antérieure	61,0%	67,2%
Diplômes obtenus	40,0%	50,0%

Critère n°3 (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 17). Il n'améliore que pour certaines variables la distortion par rapport au hot-deck séquentiel avec redéfinition des contraintes.

Tableau 17 : comparaison de la mesure de distorsion en cas de non-réponse massive et concentrée entre le hot-deck métrique et la méthode actuelle

	Méthode actuelle simulée	**Hot-deck séquentiel avec redéfinition des contraintes et aléa d'imputation**	**Hot-deck métrique**
Tranche d'âge	14,0	4,5	1,2
Situation principale	7,5	4,4	6,4
Situation matrimoniale	10,1	5,8	3,3
Diplôme	7,4	5,1	3,7

Critère n°4 (nombre maximal de dons) : le hot-deck métrique permet de réduire le nombre de dons maximum par rapport à la méthode actuelle (cf. tableau 18).

Tableau 18 : distribution du nombre de dons par individu pour le hot-deck métrique et le hot-deck séquentiel actuel

		Nombre de dons	
	Distribution	**Hot-deck métrique**	**Hot-deck séquentiel actuel**
Sexe	Maximum	31	56
	99 ème quantile	5	6
	90ème quantile	4	3
Age	Maximum	61	502
	99 ème quantile	6	120
	90ème quantile	3	20
Etat matrimonial légal	Maximum	745	2 706
	99 ème quantile	18	146
	90ème quantile	8	29
Couple	Maximum	591	3 839
	99 ème quantile	19	188
	90ème quantile	8	47
Indicatrice de nationalité	Maximum	2 125	2 703
	99 ème quantile	22	112
	90ème quantile	6	14
Inscription dans un établissement d'enseignement	Maximum	1 307	2 370
	99 ème quantile	91	194
	90ème quantile	23	62
Situation principale	Maximum	1 246	1 985
	99 ème quantile	185	138
	90ème quantile	58	43
Indicateur de résidence antérieure	Maximum	1 462	2 370
	99 ème quantile	75	128
	90ème quantile	21	27
Diplômes obtenus	Maximum	1 516	2 375
	99 ème quantile	106	187
	90ème quantile	29	23

3.2.4 Bilan sur le hot-deck métrique

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°4). 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°2). 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 (critère n°3). Il n'améliore que pour certaines variables la distortion par rapport au hot-deck séquentiel avec redéfinition des contraintes.

3.3 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 il ne coïncide pas avec les

groupes de réponse homogènes (GRH). En effet, il n'est pas envisageable de faire coïncider les communes avec les groupes de réponse homogènes, 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 on a 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° obs	Commune	poids initial	Variables 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 adapté 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, pour 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 conduit ainsi ici à une perte d'information au niveau communal.

4 Conclusion

La méthode actuelle par hot-deck séquentiel se révèle globalement satisfaisante, malgré quelques cas limités mais très visibles de redressements aberrants dans les communautés. Ces cas problématiques 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 hot-deck séquentiel a l'avantage principal d'être robuste, simple et efficace ; il 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 de février 2002 fixant la publication des populations légales en fin décembre de chaque année et également de par le nombre important d'applications informatiques en jeu.

Telles qu'elles ont été mises en œuvre et au vu de nos critères de qualité, les méthodes de redressement plus sophistiquées telles que le hot-deck par classe ou le hot-deck métrique n'améliorent qu'à la marge la méthode actuelle.

A moindre coût en matière de production courante et d'évolution des processus, une amélioration sensible de la méthode actuelle pourrait être obtenue :

- en regroupant l'ensemble des individus en communauté dans un unique lot de saisie. Cette proposition permettrait également de limiter les différences entre des individus consécutifs dans le lot de saisie.
- 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);

Annexes

Annexe 1A : première page du bulletin individuel communauté - hors détenus

Recensement de la population - 2014
Bulletin individuel – Communauté

Exemple : DUPAS, épouse MAURIN
Nom :
Prénom :
Identifiant de la communauté
Cadre à remplir par l'enquêteur
Imprimé n° 7

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

3 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 :

4 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

5 Où habitiez-vous le 1er janvier 2013 ?
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

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

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

8 Quel est votre état matrimonial légal ?
• Célibataire (jamais légalement marié(e)) ☐ 1
• Marié(e) (ou séparé(e) mais non divorcé(e)) ☐ 2
• Veuf, veuve ☐ 3
• Divorcé(e) ☐ 4

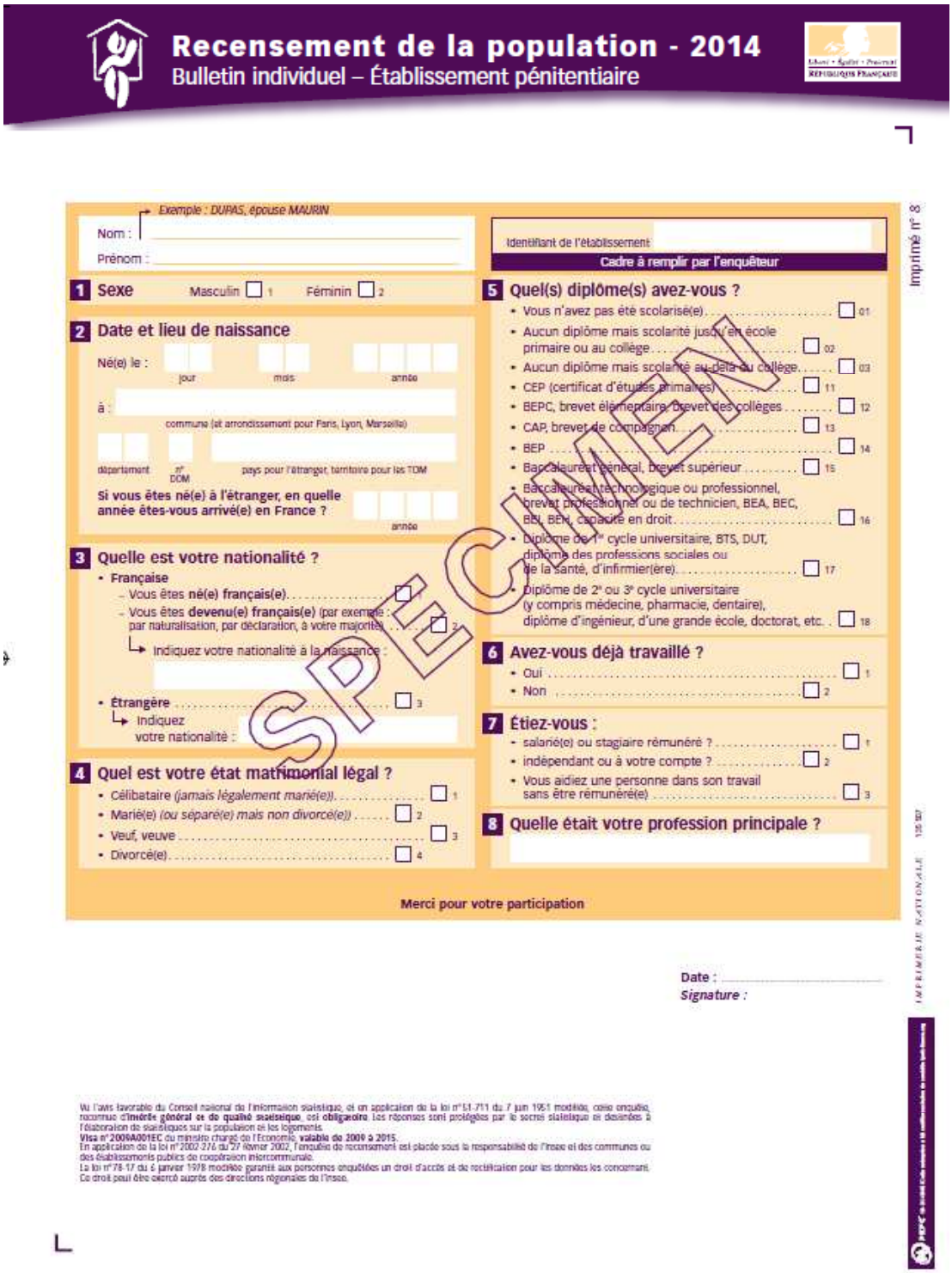
9 Quel(s) diplôme(s) avez-vous ?
• Vous n'avez pas été scolarisé(e) ☐ 01
• Aucun diplôme mais scolarité jusqu'en école primaire ou au collège ☐ 02
• Aucun diplôme mais scolarité au-delà du collège ☐ 03
• CEP (certificat d'études primaires) ☐ 11
• BEPC, brevet élémentaire, brevet des collèges ☐ 12
• CAP, brevet de compagnon ☐ 13
• BEP ☐ 14
• Baccalauréat général, brevet supérieur ☐ 15
• Baccalauréat technologique ou professionnel, brevet professionnel ou de technicien, BEA, BEC, BEI, BEH, capacité en droit ☐ 16
• Diplôme de 1er cycle universitaire, BTS, DUT, diplôme des professions sociales ou de la santé, d'infirmier(ère) ☐ 17
• Diplôme de 2e ou 3e cycle universitaire (y compris médecine, pharmacie, dentaire), diplôme d'ingénieur, d'une grande école, doctorat, etc. ☐ 18

10 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 17 ☐ 1
• Apprentissage sous contrat ou stage rémunéré ⇨ cochez puis passez en 17 ☐ 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

11 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 17 ☐ 1
• Non ⇨ cochez puis passez en 12 ☐ 2

Continuez page suivante et n'oubliez pas de signer →

Annexe 1B : bulletin individuel pour les détenus



Recensement de la population - 2014
Bulletin individuel – Établissement pénitentiaire

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Imprimé n° 8

Exemple : DURAS, 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

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

3 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 :

4 Quel est votre état matrimonial légal ?

- Célibataire *(jamais légalement marié(e))* ☐ 1
- Marié(e) *(ou séparé(e) mais non divorcé(e))* ☐ 2
- Veuf, veuve ☐ 3
- Divorcé(e) ☐ 4

5 Quel(s) diplôme(s) avez-vous ?

- Vous n'avez pas été scolarisé(e) ☐ 01
- Aucun diplôme mais scolarité jusqu'en école primaire ou au collège ☐ 02
- Aucun diplôme mais scolarité au-delà du collège ☐ 03
- CEP (certificat d'études primaires) ☐ 11
- BEPC, brevet élémentaire, brevet des collèges ☐ 12
- CAP, brevet de compagnon ☐ 13
- BEP ☐ 14
- Baccalauréat général, brevet supérieur ☐ 15
- Baccalauréat technologique ou professionnel, brevet professionnel ou de technicien, BEA, BEC, BEI, BEH, capacité en droit ☐ 16
- Diplôme de 1er cycle universitaire, BTS, DUT, diplôme des professions sociales ou de la santé, d'infirmier(ère) ☐ 17
- Diplôme de 2e ou 3e cycle universitaire (y compris médecine, pharmacie, dentaire), diplôme d'ingénieur, d'une grande école, doctorat, etc. ☐ 18

6 Avez-vous déjà travaillé ?

- Oui ☐ 1
- Non ☐ 2

7 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

8 Quelle était votre profession principale ?

Merci pour votre participation

Date :

Signature :

IMPRIMERIE NATIONALE 135 307

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° 2009A001EC du ministre chargé de l'Économie, **valable de 2009 à 2015.**
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.

Annexe 2 : signification des modalités des variables

Sous-catégorie de la communauté :
11 : Maison de retraite, hospice
12 : Foyer sonacotra
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 : Etablissement pénitentiaire
61 : Etablissement social de court séjour
71 : Autre communauté

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 : Etablissement pénitentiaire
6 : Etablissement social de court séjour
7 : autre catégorie de 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 individus 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 :
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

Effectif de communauté :
0 : entre 0 et 19 résidents
1 : entre 20 et 74 résidents
2 : entre 75 et 199 résidents
3 : entre 200 et 399 résidents
4 : entre 400 et 879 résidents
5 : plus de 880 résidents

Annexe 3 : les contraintes d'imputation pour les variables étudiées

Variable à imputer	Contraintes d'imputation
sexe	
âge	
Etat matrimonial légal	
Vie en couple	Tranche d'âge Etat matrimonial Sexe
Indicatrice de nationalité	
Inscription dans un établissement d'enseignement	Tranche d'âge
Situation principale	Indicatrice "travaille actuellement" Sexe Tranche d'âge
diplômes obtenus	Tranche d'âge Indicatrice de nationalité

Bibliographie

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

http://insee.fr/fr/themes/document.asp?reg_id=0&id=1540

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

http://www.insee.fr/fr/publications-et-services/sommaire.asp?codesage=imeths01

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

support de cours Master de statistique publique

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

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

[6] Documentation Insee sur les redressements dans le recensement

http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/doc/traitement_donnees_rp.htm