Journées de méthodologie statistique de l'Insee 2022

Méthodes de carroyage du recensement de la population dans les communes de 10 000 habitants et plus

Martin Chevalier (*), Gabrielle Gallic (**), Clément Guillo (***), Gaël Guymarc (**), Céline Pilorge (*)

(*) Insee, Département de la Démographie
(**) Insee, Département de l'Action Régionale
(***) Insee, Département des Méthodes Statistiques

martin.chevalier@sante.gouv.fr ; gabrielle.gallic@insee.fr ; clement.guillo@insee.fr ; gael.guymarc@insee.fr ; celine.pilorge@insee.fr

Mots-clés : Carroyage, estimation sur petits domaines, imputation.

Domaines Statistique spatiale – Estimation sur petits domaines, carroyage ; Démographie – Recensement.

Résumé

Ce document présente les méthodes de carroyage du recensement de la population (RP) dans les communes de 10 000 habitants et plus (« grandes » communes). La population y vivant en ménage ordinaire est enquêtée tous les ans par sondage, sur un échantillon d'adresses représentant 8 % des logements. En cumulant cinq enquêtes, environ 40 % des logements des grandes communes sont recensés. Chaque année, la population légale des communes est estimée sur cette base et sur le nombre de logements issu d'un référentiel d'adresses d'habitation en grande commune, le répertoire d'immeubles localisés (RIL), avec pour référence l'année médiane du cycle. Pour être en mesure de diffuser les données du RP au carreau, un des principaux enjeux est de réaliser des estimations fiables dans les grandes communes, malgré le faible nombre de logements échantillonnés dans certains carreaux.

Trois méthodes de carroyage du RP dans les grandes communes sont présentées dans ce document. Il s'agit de méthodes d'imputation de données à l'adresse : elles permettent d'imputer des caractéristiques de logements et d'individus à l'ensemble des adresses du RIL à partir des données des adresses enquêtées et d'informations auxiliaires d'origine fiscale, préalablement appariées au RIL.

La méthode d'imputation par modélisation est une méthode d'estimation du type « petits domaines » de niveau individuel (ici, l'adresse). Elle consiste à estimer les variables d'intérêt pour

les adresses non enquêtées par le recensement à partir de modèles de régression linéaire. Chaque variable est estimée par un modèle spécifique, construit à partir des données auxiliaires.

La méthode d'imputation par *hot deck* consiste à imputer à chaque adresse des données issues des collectes du recensement au sein de la même zone infra-communale (Iris), sous contraintes. L'idée est de reconstruire un *simili* RP exhaustif à partir des seules adresses échantillonnées, en dupliquant chaque adresse enquêtée à hauteur de son poids et en allouant ces poids dupliqués aux adresses ayant des caractéristiques proches. Deux variantes de cette méthode sont mises en œuvre.

Ces méthodes de carroyage permettent d'estimer la population au carreau de manière plus fiable que ne le fait l'estimateur usuel du recensement, appliqué au carreau. Différents critères sont définis pour analyser les atouts et faiblesses de chacune des méthodes : des critères de cohérence, des critères de performance et des critères de mise en œuvre. Au final, l'Insee a retenu la méthode d'imputation par *hot deck*, privilégiant au strict critère de performance la cohérence des estimations à différents niveaux géographiques et la souplesse du cadre d'estimation.

Abstract

This paper presents the estimation methods conceived by INSEE to disseminate Census data on a 1 km² grid in municipalities with 10,000 inhabitants or more, where population living in ordinary households is enumerated through annual sample surveys. The main challenge is to provide reliable estimates, despite the low number of sampled dwellings in some grid cells. Two sets of estimation methods have been tested : "imputation by modelling", using linear regression models, and "hot deck imputation", reconstructing a mock exhaustive Census from the sampled residential buildings (with two variants). All three methods allow for a more reliable estimate of the population at the grid cell than the usual census estimate. In the end, INSEE retained the method of imputation by hot deck with dwellings target, favouring estimates' consistency at different geographical levels and the flexibility of the estimation framework over strict performance criterion.

Introduction

Ce document présente les méthodes de carroyage du recensement de la population (RP) dans les communes de 10 000 habitants et plus (« grandes » communes). La population y vivant en ménage ordinaire est enquêtée tous les ans par sondage, sur un échantillon d'adresses représentant 8 % des logements (encadré 1). En cumulant cinq enquêtes, environ 40 % des logements des grandes communes sont recensés. Chaque année, la population légale des communes est estimée sur cette base et sur le nombre de logements issu d'un référentiel d'adresses d'habitation en grande commune, le répertoire d'immeubles localisés (RIL), avec pour référence l'année médiane du cycle. Plusieurs indicateurs sont diffusés à des niveaux infra-communaux mais aucune donnée n'est encore produite à la maille des carreaux.

Le carroyage du recensement est impulsé par une demande européenne qui prévoit, dans le cadre du Recensement européen 2021 (Census 2021), la fourniture de différentes variables au niveau de carreaux de 1 km de côté pour la France métropolitaine. Cette demande de données carroyées s'inscrit dans une tendance de plus long terme : un projet de livraison annuelle à Eurostat à partir de 2025 ou de 2026 est en cours d'élaboration [1] et une demande de données carroyées à une maille infra-communale plus fine que les Iris s'exprime fortement au niveau national.

1. Règlement ESOP (European statistics on population).

Dans les grandes communes, l'enjeu est de pouvoir réaliser des estimations fiables sur des carreaux de 1 km², malgré le faible nombre de logements échantillonnés dans certains carreaux. Dans les communes de moins de 10 000 habitants (« petites communes ») en revanche, il n'y a pas de difficulté d'estimation, le recensement étant exhaustif (annexe A). C'est également le cas pour les personnes qui ne vivent pas en logements ordinaires en grande commune : résidents en communautés et dans une résidence touristique.

Les travaux méthodologiques menés conduisent à appréhender la problématique d'estimation comme un problème d'imputation de données à l'adresse [2]. Il s'agit d'imputer l'ensemble du RIL à partir des seules adresses enquêtées et d'informations auxiliaires. Pour réaliser ces imputations, le RIL a été préalablement enrichi par des données fiscales. Deux types de méthodes sont envisagées : une méthode d'imputation par modélisation, qui correspond à une méthode d'estimation du type « petits domaines » de niveau adresse, et une méthode d'imputation par *hot deck*, dans laquelle les données des adresses non enquêtées sont des réplications de celles d'adresses enquêtées, sous certaines contraintes. Deux variantes de cette dernière méthode ont été testées, selon la cible retenue, le nombre de logements à l'adresse ou la population fiscale à l'adresse.

Ces trois méthodes permettent d'estimer la population au carreau de manière plus fiable que ne le fait l'estimateur usuel du recensement, appliqué au carreau. Des critères de choix ont été définis pour trancher en faveur d'une méthode : performance en termes de biais et de variance, cohérence avec les données du recensement diffusées à d'autres mailles géographiques, facilité de mise en œuvre et caractère généralisable ou non de la méthode. La méthode retenue *in fine* est la méthode d'imputation par *hot deck* avec cible de logements, l'Insee ayant choisi de privilégier la cohérence au strict critère de performance, eu égard à l'utilisation qui sera faite de ces données (essentiellement de la cartographie).

La première partie de l'article présente les enjeux du carroyage de la population dans les grandes communes et les données disponibles pour y répondre. La deuxième partie décrit les méthodes de carroyage mises en œuvre : la méthode d'imputation par modélisation et la méthode d'imputation par *hot deck* (avec deux variantes). La troisième partie propose une comparaison des atouts et faiblesses des 3 méthodes d'estimation envisagées.

2. Une adresse fait référence ici à un bâtiment d'habitation individuelle ou collective, qui peut différer dans certains cas de l'adresse postale (par exemple, lorsqu'il existe plusieurs bâtiments à une même adresse postale, des adresses distinctes sont créées dans le recensement de la population).

Encadré 1 : Le recensement de la population dans les communes de 10 000 habitants et plus

En France, la méthode de recensement dépend de la taille de la commune et de la catégorie de population. Dans les communes de 10 000 habitants et plus (« grandes communes »), le recensement de la population vivant en logement ordinaire est réalisé chaque année par sondage, auprès d'un échantillon d'environ 8 % des logements de la commune. Au cours d'un cycle de cinq ans, environ 40 % des logements sont interrogés. Plus précisément, les adresses des grandes communes sont réparties en cinq groupes (un groupe par année du cycle de cinq ans), équilibrés notamment par rapport au nombre de logements. Chaque année, un échantillon d'adresses est tiré dans un groupe. Les adresses nouvelles, les grandes adresses et les structures touristiques sont toutes retenues dans l'échantillon (leur poids de sondage est alors égal à 1). Les autres adresses sont échantillonnées, de sorte qu'au total environ 40 % des logements du groupe d'adresses soient retenues (leur poids de sondage, qui correspond à l'inverse du taux de sondage, est généralement compris entre 2,5 et 5 et plus souvent près de 3 [3]). Tous les logements et individus d'une adresse échantillonnée sont enquêtés exhaustivement.
La méthode d'estimation utilisée est une estimation dite « par le ratio ». Cette méthode utilise les observations enquêtées pour produire une estimation de la population communale. Elle consiste à ajuster les poids de sondage des logements et individus enquêtés en utilisant l'information issue du répertoire d'immeubles localisés (RIL). Plus précisément, les poids finaux sont obtenus en multipliant les poids de sondage par le ratio du nombre de logements au RIL médian (au 1^{er} janvier de l'année médiane du cycle du recensement) sur le nombre de logements estimé sur les cinq années du cycle. En multipliant ce poids d'estimation par le nombre d'individus enquêtés, on obtient une estimation de la population. Toutes les variables du recensement de la population (population selon le sexe, l'âge, nombre de logements, etc.) sont estimées de cette façon.
Ces calculs sont réalisés à une échelle infra-communale [4], l'Iris (Îlots regroupés pour l'information statistique), puis ces estimations sont sommées au niveau de la commune. Les Iris ont été construits en concertation avec les communes, à partir de critères géographiques et statistiques. La cible de population initiale pour un Iris est de l'ordre de 2 000 habitants : les communes d'au moins 10 000 habitants et la majorité des communes de 5 000 à 10 000 habitants sont découpées en Iris. Dans ces communes, l'Iris est la maille de base de la diffusion des statistiques infra-communales. Dans les communes non irisées, les statistiques sont publiées à l'échelle communale.

Les personnes vivant en communauté [5] ou dans une résidence touristique [6] sont recensées exhaustivement au cours d'un cycle de cinq ans. Elles sont donc exclues du champ de l'étude méthodologique car leur estimation ne pose pas de problème spécifique. Elles seront toutefois bien comptabilisées *in fine* dans les données carroyées. En revanche, les sans-abris et les personnes résidant dans une habitation mobile ne seront pas comptabilisées, faute de géolocalisation précise. Elles seront affectées à un carreau virtuel « unallocated ».

2. Le poids moyen pour l'ensemble des logements et individus d'une commune est de 1/0,40 = 2,5. Sachant par ailleurs que le poids des logements et individus des adresses enquêtées exhaustivement est de 1, les poids des adresses échantillonnées sont généralement supérieurs à 2,5.

3. Sauf exceptions : dans les départements ultramarins, la maille utilisée est l'îlot et non l'Iris (regroupement d'îlots) ; par ailleurs, en métropole les Iris d'activité, contenant peu de population, sont par exemple regroupés au sein d'une même commune.

4. Par exemple : maison de retraite, foyer de travailleurs, gendarmerie, résidence étudiante, établissement pénitentiaire, etc.

5. Hôtel, camping ou résidence hôtelière.

1 Le carroyage du recensement de la population dans les grandes communes : enjeux et données

Dans le cadre du Census 2021, plusieurs variables seront diffusées sur des carreaux de 1 km de côté : la population totale, la population selon le sexe, l'âge (moins de 15 ans, 15-64 ans, 65 ans ou plus), le lieu de naissance (France, autre pays de l'Union européenne, pays en dehors de l'Union européenne), le lieu de résidence un an auparavant (inchangé, ailleurs en France, à l'étranger), et, dans la mesure du possible, la population en emploi. Des estimations provisoires sont attendues dès décembre 2022 pour la population totale, puis la livraison des estimations définitives sera effectuée en mars 2024 pour l'ensemble des variables.
Les analyses présentées dans ce rapport portent sur le recensement 2017, qui mobilise les données des enquêtes annuelles de recensement de 2015 à 2019. Les communes ayant connu un changement de géographie – fusion ou scission – ou ayant franchi le seuil des 10 000 habitants – à la hausse ou à la baisse – entre ces deux dates sont exclues du champ de l'analyse ; elles seront intégrées lors de la mise en production des données au carreau. L'étude porte au total sur 922 grandes communes de France métropolitaine n'ayant connu aucune modification de géographie ni de méthode d'estimation.

1.1 Le recensement par sondage : comment estimer la population dans un carreau où aucun logement n'a été enquêté ?

Le fait de réaliser une enquête par sondage dans les grandes communes conduit à ce qu'il y ait parfois peu ou pas de logements échantillonnés dans un carreau de 1 km de côté. La France métropolitaine compte près de 558 000 carreaux de 1 km de côté, dont environ 33 130 situés (au moins en partie) dans une grande commune. Parmi eux, 22 457 carreaux comptent au moins une adresse dans le RIL médian de 2017 (année de référence de l'étude) ; 10 673 carreaux sont donc vides [6]. Parmi les 22 457 carreaux qui comportent au moins une adresse au RIL médian en 2017, 27 % en comptent moins de 10 (tableau 1). Par ailleurs, environ 10 % des carreaux ne comptent aucune adresse échantillonnée au RP 2017 et près de 30 % en comptent entre 1 et 9 (tableau 2). Les chiffres sont semblables en termes de nombre de logements : 25 % des carreaux comptent moins de 10 logements au RIL médian au RP 2017 (tableau 3) et 10 % n'ont aucun logement échantillonné (tableau 4). En outre, près de la moitié des carreaux ont un taux de logements échantillonnés inférieur à 40 %, taux moyen pour chaque commune. Ce faible niveau d'information dans certains carreaux remet en cause la fiabilité de la méthode d'estimation usuelle (encadré 1) à un niveau aussi fin. L'enjeu de cette étude est de définir une méthodologie permettant de répondre à ce problème d'estimation au niveau de carreaux de 1 km de côté. L'objectif est de produire des données de population au carreau fiables tout en restant cohérentes avec les données publiées à d'autres niveaux géographiques, notamment la commune. Les méthodes d'estimation envisagées consistent à imputer des données du recensement aux adresses non échantillonnées à partir des adresses échantillonnées et d'une information auxiliaire.

6. Au sens où il n'y a aucun logement en grande commune dans ces communes. Il peut en revanche y avoir des logements en petite commune, s'ils ne sont pas totalement inclus dans des grandes communes.

Tableau 1 – Nombre de carreaux en fonction du nombre d'adresses dans le RIL médian de 2017 (année de référence de l'étude)

1-9	10-49	50-99	100-199	200 et plus	Total
6 015	4 656	2 018	2 118	7 650	22 457

Source : Recensement de la population 2017.
Champ : les 22 457 carreaux des grandes communes de France métropolitaine comptant au moins une adresse au RIL médian au RP 2017.
Lecture : au RP 2017, 22 457 carreaux comptent au moins une adresse au RIL médian. Parmi eux, 6 015 comptent moins de 10 adresses.

Tableau 2 – Nombre de carreaux en fonction du nombre d'adresses échantillonnées au recensement de la population de 2017

0	1-9	10-49	50-99	100-199	200 et plus	Total
2 208	6 491	4 739	2 419	2 890	3 710	22 457

Source : Recensement de la population 2017.
Champ : les 22 457 carreaux des grandes communes de France métropolitaine comptant au moins une adresse au RIL médian au RP 2017.
Lecture : au RP 2017, 22 457 carreaux comptent au moins une adresse au RIL médian. Parmi eux, 2 208 carreaux n'ont aucune adresse échantillonnée.

Tableau 3 – Nombre de carreaux en fonction du nombre de logements dans le RIL en 2017

1-9	10-49	50-199	200-499	500-999	1 000 et plus	Total
5 532	4 374	3 476	2 348	2 108	4 619	22 457

Source : Recensement de la population 2017.
Champ : les 22 457 carreaux des grandes communes de France métropolitaine comptant au moins une adresse au RIL médian au RP 2017.
Lecture : au RP 2017, 5 532 carreaux comptent entre 1 et 9 logements au RIL médian.

Tableau 4 – Nombre de carreaux en fonction du nombre de logements échantillonnés

0	1-9	10-49	50-199	200-499	500-999	1 000 et plus	Total
2 335	5 823	4 079	3 527	2 802	2 095	1 796	22 457

Source : Recensement de la population 2017.
Champ : les 22 457 carreaux des grandes communes de France métropolitaine comptant au moins une adresse au RIL médian au RP 2017.
Lecture : au RP 2017, 2 335 carreaux ne comptent aucun logement échantillonné.

1.2 Description des données mobilisées

1.2.1 Les adresses : coordonnées géographiques et nombre de logements

La mise en œuvre des méthodes de carroyage par imputation nécessite de disposer de la liste exhaustive des adresses des grandes communes. Cela est possible grâce à l'existence d'un répertoire d'immeubles localisés (RIL), identifiant toutes les adresses d'habitation dans ces communes. Ce répertoire est actualisé chaque année par un partenariat Insee-communes. Il est utilisé pour le tirage d'échantillon des enquêtes annuelles de recensement (la base de sondage est constituée à partir du RIL), le calcul des populations légales et l'édition des plans de collecte. Pour chaque adresse, le nombre de logements habitables en 2017 est connu (par la suite, on parlera du « nombre de logements au RIL médian », 2017 étant l'année médiane du cycle considéré). Ces adresses sont géolocalisées. Leurs coordonnées géographiques ont été converties dans le référentiel de projection du Census 2021, ETRS89 Lambert Azimuthal Equal-Area. L'identifiant de carreau officiel est également associé à chaque adresse.

1.2.2 Ajout d'une information auxiliaire pour chaque adresse au RIL médian à partir des données fiscales

Les adresses du RIL médian sont appariées avec les données fiscales de l'année 2017, à partir des libellés d'adresses, afin de récupérer des informations auxiliaires, indispensables pour procéder à l'imputation. Ces informations (de niveau adresse) sont de nature socio-démographique : population par sexe, âge et lieu de naissance ; mobilité en France [7] au cours de l'année ; revenus des ménages ; caractéristiques des logements. Toutes n'ont pas été mobilisées dans le cadre de cette étude, mais pourraient l'être lors de développements ultérieurs pour améliorer les performances. Environ 90 % des adresses du RIL médian sont appariées aux données fiscales. Pour les 10 % d'adresses non appariées, les valeurs des variables auxiliaires sont imputées : pour chaque adresse on considère, pour chacune des variables, le nombre moyen de personnes par logement (NMPL) dans les données fiscales (population totale, nombre moyen de femmes, nombre moyen d'hommes, nombre moyen de moins de 15 ans, etc.), à l'échelle de l'Iris, que l'on multiplie ensuite par le nombre de logements à l'adresse. Le NMPL dans l'Iris est calculé sur la base des données fiscales des adresses appariées [8]. Lorsque le taux d'appariement dans l'Iris est faible, on considère le NMPL à l'échelle de la commune [9]. Pour imputer le nombre de femmes par exemple, on retient le nombre moyen de femmes par logement dans les données fiscales, dans l'Iris (ou la commune selon les cas) de l'adresse non appariée, et on le multiplie par le nombre de logements à l'adresse. Afin d'assurer la cohérence entre la somme des modalités d'une variable et la population totale à l'adresse, on considère au final la structure de chaque variable après imputation et on la multiplie par la population totale imputée. Par exemple, pour le nombre de moins 15 ans, on applique la part de moins de 15 ans à l'adresse à la population totale imputée à cette adresse.

7. Mobilité en France (au sein de la même commune ou entre deux communes différentes) exclusivement. La mobilité d'un autre pays vers la France n'est pas renseignée.

8. Idéalement, le NMPL dans l'Iris pourrait être calculé sur la base des adresses fiscales non appariées. Toutefois, le volume d'adresses non appariées du RIL diffère parfois beaucoup du volume d'adresses non appariées de Fidéli pour un même Iris. Dans la mesure où le passage en production utilisera un nouvel appariement plus performant entre le RIL et les données fiscales, nous avons choisi de nous en tenir au NMPL des adresses appariées dans le cadre de cette étude.

9. Différents seuils ont été testés : lorsque le taux d'appariement dans l'Iris est inférieur à 60 %, le NMPL à l'échelle de la commune est retenu. Au final, une fois l'imputation réalisée, l'écart observé entre le NMPL dans les données du RP et le NMPL dans les données fiscales imputées est faible. Le NMPL dans les données fiscales est légèrement plus élevé : pour 80 % des adresses non appariées, l'écart de NMPL entre les données du RP et les données fiscales est compris entre -0,740 et -0,013, avec une médiane à -0,239. L'imputation réalisée n'est donc pas de nature à biaiser sensiblement les résultats.

1.2.3 Les données collectées par le recensement de la population

Pour les adresses échantillonnées et enquêtées au cours du cycle 2015-2019, nous disposons des données issues du recensement pour l'ensemble des variables demandées au carreau. Aucune estimation n'est nécessaire pour ces adresses.

1.2.4 Récapitulatif de la base de travail

En mobilisant les informations fournies par le RIL, les fichiers fiscaux et le recensement de la population, nous avons construit une base de données contenant, pour chaque adresse du champ, ses coordonnées géographiques, le nombre de logements au 1^{er} janvier 2017 et des caractéristiques socio-économiques de la population issues des sources fiscales (informations auxiliaires de la base imputée). Pour les adresses du champ enquêtées par le recensement de la population, cette base contient également les résultats de l'enquête ainsi que leur poids d'estimation respectif.

2 Méthodes de carroyage du recensement de la population

Trois méthodes d'estimation au carreau sont envisagées. Leur principe général est d'imputer l'ensemble des adresses du RIL médian à partir des adresses échantillonnées et d'une information auxiliaire. Ces méthodes diffèrent néanmoins dans la manière de réaliser cette imputation. L'output est un fichier de données individuelles, contenant une estimation des variables d'intérêt pour chaque adresse du RIL médian, qu'elles aient été enquêtées ou non par une collecte du recensement. Les estimations à chaque adresse peuvent ensuite être sommées pour estimer des indicateurs sur les carreaux. Pour toutes les méthodes, les fichiers de données individuelles sont construits indépendamment du zonage d'intérêt – les carreaux ici. Ils pourraient donc être mobilisés pour d'autres zonages, comme les quartiers prioritaires de la politique de la ville (QPV) ou des zonages à façon.

2.1 Méthode d'imputation par modélisation

2.1.1 Principe général de la méthode

La méthode d'imputation par modélisation consiste à estimer les variables d'intérêt Y pour les adresses du RIL médian non enquêtées par le recensement *via* des modèles de régression linéaire [10]. Chaque variable est estimée par un modèle spécifique, construit à partir des données auxiliaires. Seules les adresses non-enquêtées par le recensement sont imputées car nous disposons déjà de valeurs collectées pour les adresses enquêtées. Les modèles sont estimés au niveau de chaque commune.
Les données auxiliaires sont disponibles grâce à un appariement entre les données du recensement et les données fiscales (paragraphe 1.2.2).
Les variables de revenus et sur les caractéristiques des logements (par exemple, la distinction entre un logement social et un logement du parc privé) n'ont pas été mobilisées à ce stade ; elles pourraient toutefois l'être lors de travaux ultérieurs, notamment en vue d'une diffusion au carreau d'autres variables que celles étudiées dans le cadre du recensement européen, comme le diplôme ou la catégorie sociale.

10. La linéarité de la relation entre population du recensement et population fiscale a été vérifiée au préalable. Elle se vérifie mieux dans les très grandes communes que dans les plus petites.

2.1.2 Choix du modèle : sélection des variables explicatives et domaine d'estimation

Le choix du modèle est une étape centrale de ce processus d'imputation. L'objectif est de construire un modèle pour chaque variable d'intérêt Y que l'on souhaite diffuser au carreau.
La variété des variables explicatives disponibles conditionne largement les possibilités d'imputation des variables d'intérêt. On anticipe ainsi des performances inégales entre les estimations des variables d'intérêt fortement corrélées aux variables explicatives (population totale, population par sexe, âge et lieu de naissance) et celles des variables d'intérêt moins corrélées (comme la population par type de diplôme, variable également testée dans cette étude).

Les modèles sont estimés commune par commune. Cela permet de tenir compte d'éventuelles spécificités territoriales en termes de liens entre les variables explicatives et les variables à expliquer, contrairement à un modèle unique pour toutes les adresses en grande commune métropolitaine. Le domaine d'estimation communal permet par ailleurs d'avoir suffisamment de données pour construire un modèle fiable (contrairement aux Iris). Il s'agit enfin de la maille à laquelle le recensement est organisé et diffusé de manière privilégiée.

Pour une variable d'intérêt Y donnée, la définition des modèles (au sens de la liste des variables explicatives les composant) est la même pour toutes les grandes communes métropolitaines. Ce parti pris permet d'avoir une méthode d'estimation identique pour toutes les communes de 10 000 habitants ou plus.

Pour les variables d'intérêt Y ayant leur homologue dans les données fiscales (population, population par âge et par sexe), la sélection des modèles est faite par validation croisée (encadré 2). Les modèles retenus sont les modèles « minimaux », c'est-à-dire que le modèle n'est composé que de la variable homologue dans les sources fiscales.
Ainsi, pour prédire les variables Y_k de population, population par sexe et population par tranche d'âge, les modèles suivants sont utilisés :

$$Y_{k,i} = \alpha_k X_{k,i} + \epsilon_k \tag{1}$$

avec $X_k \in$ {population, population par sexe, population par âge},
i = (1, …, n) les adresses du RIL médian non enquêtées,
α_k les paramètres estimés par les modèles de régression (coefficients de régression) et
ϵ_k des variables aléatoires non observées (erreurs).

La modélisation précédente ne fait pas intervenir de constante : on suppose donc un lien multiplicatif entre une variable du RP donnée et son homologue dans les variables auxiliaires. Le fait de forcer la droite de régression à rejoindre l'origine permet de neutraliser l'effet des adresses proches de l'origine (population faible), pour lesquelles un fort déséquilibre est observé entre la donnée RP et la donnée fiscale.
Pour prédire les autres variables d'intérêt Y (par exemple, la population résidant en France un an auparavant), les modèles sont une combinaison linéaire des variables fiscales :

$$Y_{k,i} = \alpha + \alpha_1 X_{1,i} + \alpha_2 X_{2,i} + \alpha_3 X_{3,i} + \epsilon_k \tag{2}$$

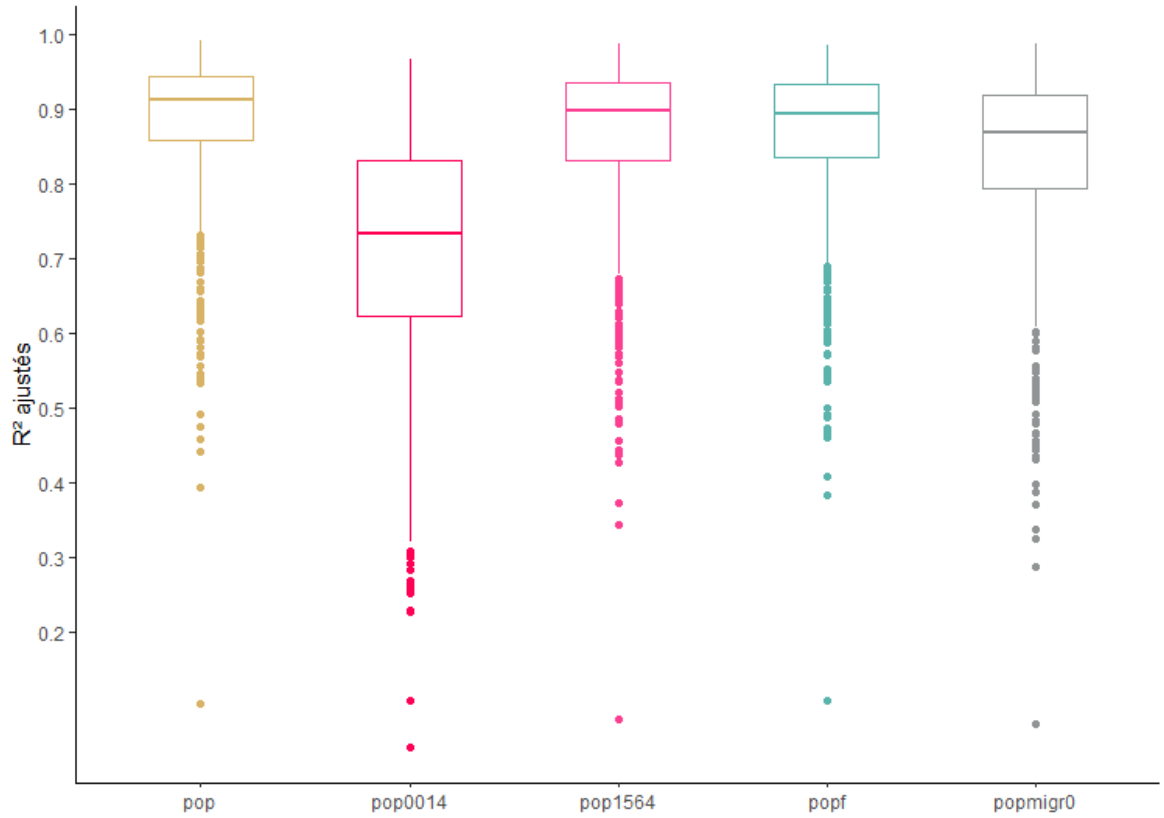
avec $Y_k \in$ {population par lieu de naissance, population par lieu de résidence il y a un an, etc.},
X_1 = population, X_2 = population par âge, X_3 = population par sexe,
α_1, α_2, α_3 les coefficients de régression,
i = (1, …, n) les adresses du RIL médian non enquêtées,
ϵ_k des variables aléatoires non observées (erreurs).

Les variables auxiliaires « population par pays de naissance » et « population ayant déménagé en France au cours de l'année » n'apportent pas de gain significatif pour prédire respectivement la

population par pays de naissance et par lieu de résidence un an auparavant [11]. Compte-tenu de ce faible apport et de la qualité incertaine de l'imputation des valeurs manquantes, ces variables auxiliaires n'ont pas été conservées dans les modèles.

Ce sont globalement les mêmes communes (une centaine) qui sont concernées par un ajustement linéaire faible pour les différentes variables Y à expliquer. Les plus petites communes sont surreprésentées : 70 % d'entre elles ont entre 10 000 et 20 000 habitants, alors que ce type de communes représente la moitié des 922 communes du champ.

GRAPHIQUE 1 – Distribution des R^2 ajustés associés aux modèles retenus pour l'ensemble des GC



Source : Recensement de la population 2017 et Fidéli 2017 et 2018.
Champ : adresses des grandes communes de France métropolitaine, hors changement de géographie et franchissement de seuil des 10 000 habitants (grande commune devenant petite ou petite commune devenant grande), ayant été enquêtées sur le cycle du RP 2017.
Note : « pop » désigne la population à l'adresse ; « pop0014 », le nombre de personnes de moins de 15 ans ; « pop1564 », le nombre de personnes entre 15 et 64 ans ; « popf », le nombre de femmes et « popmigr0 », le nombre de personnes qui résidaient déjà à cette adresse un an auparavant.

11. Pour mesurer l'apport de la variable auxiliaire "population née en France" dans un modèle où la variable à expliquer est la population née en France, on compare, par validation croisée sur l'ensemble des grandes communes du champ, un modèle comprenant une combinaison des variables auxiliaires de population, population par sexe et population par âge avec un modèle comprenant cette même combinaison ainsi que la population née en France. Le critère de comparaison est l'erreur moyenne absolue (MAE). On remarque que pour chaque commune l'ajout de la population née en France n'améliore que très légèrement la MAE.

Le graphique 1 illustre les distributions des coefficients de détermination R^2 associés aux modèles retenus pour plusieurs variables d'intérêt Y. Pour le modèle associé à la population par exemple, l'ajustement linéaire est de très bonne qualité dans la majorité des cas (526 communes ont un R^2 supérieur à 0,9), mais pour quelques communes, la qualité de l'ajustement est nettement plus faible : pour 40 communes le R^2 est inférieur à 0,7. Ce même constat se retrouve globalement pour les autres variables représentées (population entre 15 et 64 ans, population de femmes, population résidant en France un an auparavant). Pour les moins de 15 ans, la qualité de l'ajustement est légèrement moindre : cela peut s'expliquer par le fait que cette population est moins bien comptabilisée dans les données fiscales sur les millésimes utilisés [12], les individus ne déclarant pas de revenus étant moins bien identifiés.

2.1.3 Mise en cohérence des estimations

2.1.3.1 Au niveau des adresses

À ce stade de la procédure d'imputation par modélisation, chaque variable d'intérêt a été estimée indépendamment. Dès lors, rien ne garantit une cohérence entre ces variables au niveau de l'adresse : rien ne garantit par exemple que la somme du nombre d'hommes et du nombre de femmes d'une adresse imputée soit égale à la population estimée de cette adresse.

Une étape de mise en cohérence est donc opérée. Pour les variables de population par sexe par exemple, cette étape consiste à :

1. calculer la structure par sexe induite par les estimations du nombre d'hommes et de femmes au niveau de l'adresse (part d'hommes et part de femmes à l'adresse) ;
2. puis appliquer cette structure à l'estimation de la population à cette adresse pour obtenir des estimations du nombre d'hommes et du nombre de femmes cohérentes.

Procéder ainsi permet une gestion relativement simple des estimations. Une autre piste envisagée, qui consiste à ajuster le nombre d'hommes par soustraction entre les estimations de population et du nombre de femmes, se révèle en effet plus compliquée en cas de partition en 3 ou plus de la variable de référence. Elle permet aussi une gestion plus simple des estimations égales à 0 (aucun homme à une adresse par exemple). Cette étape de mise en cohérence s'applique de la même manière à toutes les variables qui sont une partition d'une autre variable également estimée (population par sexe donc et population par tranche d'âge, par lieu de naissance et lieu de résidence un an auparavant). Concrètement, cette mise en cohérence implique donc de décrire en amont des estimations toutes les relations d'ensemble entre les variables à estimer, comme par exemple que la somme du nombre de 0-14ans, de 15-64 ans et de 65 ans ou plus soit égale à la population totale. Par conséquent, cette méthode est assez « rigide » aux modifications d'un modèle pour une variable à expliquer à cause des implications sur les autres variables à expliquer que cela entraîne pour garantir une cohérence globale des données. Par exemple, modifier le modèle pour estimer le nombre d'hommes implique de revoir également les estimations du nombre de femmes pour garantir que la somme des deux corresponde à la population totale.

2.1.3.2 Au niveau communal (ou des Iris)

La dernière étape de l'imputation par modélisation vise à garantir une cohérence au niveau d'une maille de diffusion plus agrégée, en particulier pour deux variables principales du recensement : la population et le nombre de logements. Cette maille doit être *a minima* la commune (des tests sont également réalisés pour garantir une cohérence au niveau des Iris). Ainsi, en sommant les

12. La moindre comptabilisation des personnes de moins de 15 ans est surtout observée sur le millésime 2017 et, dans une moindre mesure, sur le millésime 2018. A partir du millésime 2020, les personnes de moins de 15 ans deviennent mieux comptabilisées avec la mise en place de la déclaration tacite

données de toutes les adresses d'une commune (qu'elles aient été imputées ou qu'elles soient directement issues de la collecte), les totaux doivent être égaux à ceux estimés directement à partir des données collectées pondérées (méthode d'estimation « habituelle » du recensement). Un calage sur marges est effectué pour atteindre cette cohérence (encadré 3).

Encadré 2 : Sélection des modèles par validation croisée

Pour chaque commune, les modèles optimaux associés aux variables d'intérêt de population, population par âge et par sexe sont déterminés par validation croisée à 7 blocs.
Le principe est le suivant : pour une commune donnée et pour une variable d'intérêt, par exemple la population, l'ensemble des combinaisons des variables explicatives est explicité. Autrement dit, pour chaque commune, tous les modèles possibles pour expliquer la population sont définis à partir des différentes combinaisons des variables explicatives. Ensuite, pour chaque commune, les adresses enquêtées sont réparties aléatoirement dans 7 groupes (le nombre de groupes est un arbitrage entre la taille du jeu d'adresses d'apprentissage et la taille du jeu d'adresses de validation, usuellement compris entre 5 et 10). Les adresses de 6 groupes sont utilisées pour définir les coefficients des modèles explicités (phase d'apprentissage). Les adresses du 7^e groupe sont utilisées pour mesurer les erreurs moyennes absolues (MAE) associées à chaque modèle (phase de validation).
L'erreur moyenne absolue (Mean Absolute Error) :

$$MAE = \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{n} \tag{3}$$

On répète l'opération avec 6 autres groupes d'adresses et ainsi de suite jusqu'à obtenir, pour chaque modèle testé, 7 MAE associées. On en déduit une MAE moyenne par modèle. Finalement, le modèle retenu pour prédire la population des adresses qui n'ont pas été enquêtées (pour une commune donnée) est celui qui a la plus petite MAE moyenne. En cas d'égalité entre deux modèles, le modèle avec le moins de variables explicatives est conservé.
On sélectionne ainsi un modèle par variable et par commune, pour les 922 grandes communes du champ.
Les modèles retenus *in fine* sont ceux qui sont comptabilisés le plus de fois parmi l'ensemble des 922 grandes communes.

Variable d'intérêt à expliquer	Modèle (variables explicatives issues des sources fiscales)	Nombre de fois que le modèle minimise la MAE (sur 922 communes)
Population	Population	623
Nombre d'hommes	Nombre d'hommes	567
Nombre de femmes	Nombre de femmes	616
Population des 0-24 ans	Population des 0-24 ans	442
Population des 25-64 ans	Population des 25-64 ans	522
Population des 65 ans et plus	Population des 65 ans et plus	726

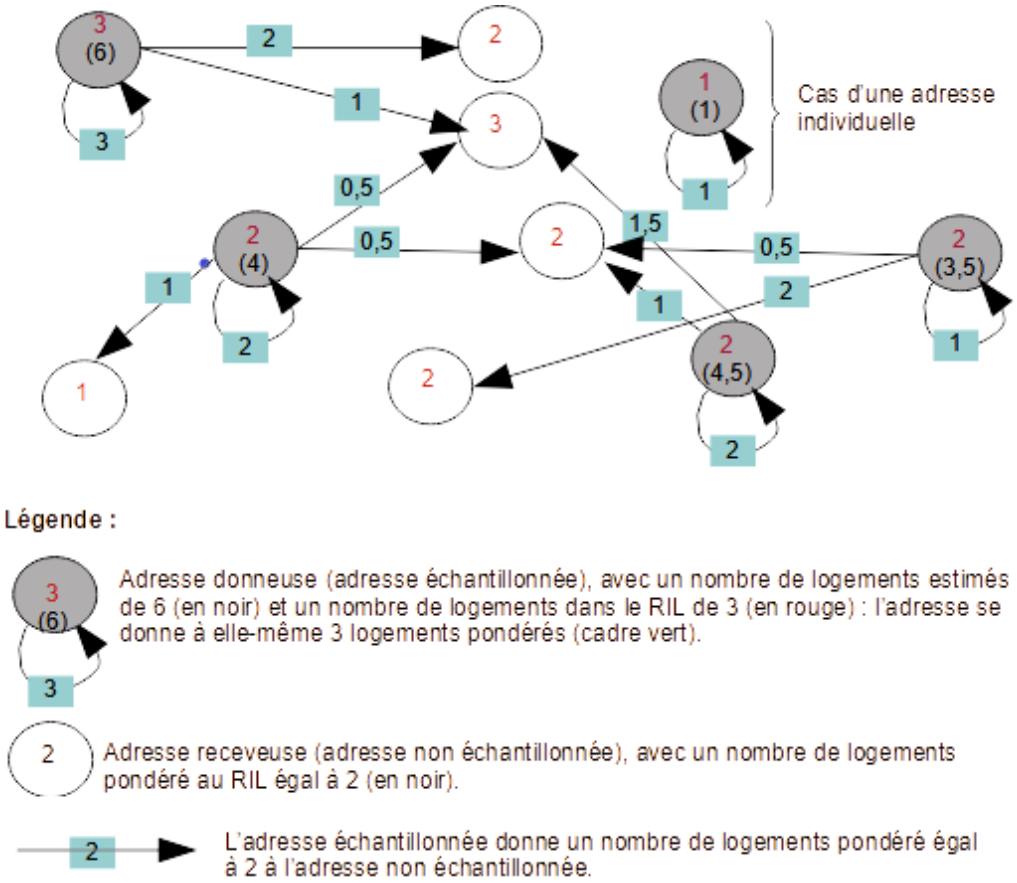
2.2 Méthode d'imputation par *hot deck*

2.2.1 Description de la méthode d'imputation par *hot deck*

2.2.1.1 Idée générale de la méthode

La méthode d'imputation par *hot deck* consiste à imputer à chaque adresse du RIL médian des données issues des collectes du recensement au sein de la même zone infra-communale (Iris), sous certaines contraintes.
Les adresses recensées lors des enquêtes annuelles ont un poids d'estimation, calculé de telle sorte que le nombre estimé de logements à l'Iris soit égal au nombre de logements du RIL médian (encadré 1). L'idée est d'imputer des fractions d'adresses recensées pondérées aux adresses non-échantillonnées du RIL médian de telle sorte que toutes les adresses du RIL aient un nombre de logements imputés égal à son nombre de logements dans le RIL, comme si le recensement était exhaustif. Les adresses recensées, appelées également adresses donneuses, se donnent à elles-mêmes (les donneurs sont « stables »), puis donnent aux adresses non enquêtées la partie restante de leur poids d'estimation. Ce poids restant peut être réparti entre plusieurs adresses receveuses. Par ailleurs, les adresses receveuses peuvent recevoir des fractions de poids issues de différentes adresses recensées (graphique 2). Elles se voient alors attribuer une fraction des caractéristiques de plusieurs adresses enquêtées. Avec cette méthode d'imputation, toutes les variables du recensement sont renseignées pour l'ensemble des adresses du RIL, à partir des données collectées « dupliquées ».
Cette répartition des poids entre les adresses est réalisée sur le principe d'une minimisation d'une distance entre des adresses donneuses (les adresses enquêtées) et des adresses receveuses (l'ensemble des adresses des grandes communes). Cette distance est exprimée en termes de variables auxiliaires (information issue des données fiscales). En cas d'égalité entre adresses, la distance géographique est également considérée (voir *infra*). Elle est pondérée par la fraction que chaque adresse échantillonnée donne à chaque adresse receveuse. Ce programme de minimisation répond à des contraintes, propres aux variantes de la méthode d'imputation par *hot deck* considérées : une variante dans laquelle on impose d'imputer à chaque adresse son nombre de logements du RIL, méthode présentée ci-dessus pour comprendre l'esprit de la méthode (paragraphe 2.2.2.1), une variante dans laquelle on impose d'imputer à chaque adresse la population auxiliaire, issue des données fiscales (paragraphe 2.2.2.2). Ainsi, selon la variante retenue, le programme d'optimisation peut conduire à une répartition différente des poids d'estimation entre les adresses.

Graphique 2 – Représentation de la méthode d'imputation par *hot deck* dans le cas où l'on impute aux adresses non échantillonnées son nombre de logements du RIL



Lecture : ce schéma illustre la répartition des logements estimés pour les adresses échantillonnées (adresses donneuses) vers les adresses non échantillonnées (adresses receveuses), pour le cas général où les donneurs donnent exactement leur poids d'estimation.

2.2.1.2 Programme d'optimisation

La méthode d'imputation par *hot deck* repose sur un problème d'optimisation linéaire. La minimisation de la distance est définie à un niveau infra-communal, l'Iris, qui est la strate d'estimation usuelle du recensement de la population en grandes communes (encadré 1).

$$\begin{aligned} \min_{\lambda_{d,r}} \quad & \sum_{d,r} \lambda_{d,r} dist_{d,r} \\ \text{s.c.} \quad & \forall d \in L^{(D)} \sum_{r \in L^{(R)}} \lambda_{d,r} \leq w_d * coef \\ & \forall r \in L^{(R)} \sum_{d \in L^{(D)}} qte\_cible_d \lambda_{d,r} = qte\_cible_r \end{aligned}$$

$L^{(D)}$ est l'ensemble des adresses donneuses, $L^{R)}$ l'ensemble des adresses receveuses.

w_d est le poids d'estimation de l'adresse dans le recensement de la population.

$\lambda_{d,r}$ est l'indicateur d'appariement entre adresses donneuses et receveuses.

$coef$ est un coefficient de pondération permettant aux donneurs de donner plus que leur poids

d'estimation ($coef$ est supérieur ou égal à 1).

Par défaut, *coef* est égal à 1, mais différentes variantes sont testées afin d'augmenter les degrés de liberté, et éventuellement la précision : un donneur peut donner jusqu'à 1,2 fois, 1,5, fois, 2 fois ou 3 fois son poids d'estimation. Dans ce cas, certaines adresses donnent plus que son poids d'estimation et d'autre moins. Par ailleurs, le modèle est construit de telle sorte que les adresses donneuses se donnent à elles-mêmes (les donneurs sont « stables »).

qte_cible définit la quantité cible que nous souhaitons atteindre en « distribuant » les poids des adresses donneuses. Deux variantes sont envisagées : l'une avec une cible de nombre de logements du RIL et l'autre avec une cible de population fiscale.

La distance $dist_{d,r}, r$ correspond à la distance entre deux adresses aux sens d'une ressemblance en termes de caractéristiques socio-démographiques, issues des informations fiscales : elle est calculée à partir du nombre moyen de personnes par logement et de parts pour les variables de structures (population par sexe et par classe d'âge [13]) [14], préalablement centrées et réduites. Il s'agit d'une distance euclidienne multidimensionnelle.

Sur cette base, il peut y avoir un nombre important de valeurs égales dans les matrices de distance entre adresses donneuses et receveuses, impliquant dès lors un temps d'exécution du programme d'optimisation assez long. Afin de pallier ce problème, une distance hybride est considérée : les valeurs *ex æquo* au sens de la distance en termes d'information auxiliaire sont ordonnées plus finement en fonction de leur distance géographique. Par ailleurs, le nombre de paires adresses donneuses / adresses receveuses est réduit en se limitant aux paires d'adresses donneuses les plus proches d'une adresse receveuse, et inversement, sachant que les donneurs se donnent à eux-mêmes.

Les contraintes du modèle sont propres à la cible retenue : le nombre de logements ou la population auxiliaire.

2.2.2 Deux variantes de la méthode d'imputation par *hot deck* : une avec une cible de logements et une avec une cible de population

2.2.2.1 Méthode d'imputation par hot deck avec cible de logements

Avec cette méthode d'imputation, la cible à atteindre (variable qte_cible_r dans l'équation) est le nombre de logements au RIL médian de chaque adresse. Les contraintes du modèle sont les suivantes :

- d'une part, chaque adresse échantillonnée donne au plus son poids d'estimation, donc au plus son nombre de logements estimé dans le recensement [15] ($coef = 1$) ;
- d'autre part, chaque adresse reçoit son nombre de logements au RIL médian. Les donneurs étant « stables », ils se donnent et reçoivent leur nombre de logements au RIL médian.

13. Les classes d'âge utilisées pour le calcul de la distance entre adresses donneuses et receveuses sont les suivantes : moins de 15 ans, 15-24 ans, 25-39 ans, 40-64 ans, 65-74 ans et 75 ans ou plus.

14. Une variante dans laquelle la distance entre adresses donneuses et adresses receveuses est calculée à partir du nombre moyen de personnes par logement à l'adresse ainsi que de la structure par sexe, par âge, par lieu de naissance et par lieu de résidence antérieur a également été testée. Les résultats étaient moins bons pour les petits carreaux (carreaux les moins peuplés) et identiques sinon. La faible dispersion des variables de lieu de naissance et de lieu de résidence antérieur (la part de personnes nées en France à une adresse est souvent égale à 1 ; c'est également le cas pour la part d'individus n'ayant pas changé de lieu de résidence au cours de l'année) ne permet pas d'améliorer la précision du calcul de la distance dans les carreaux peuplés, et peut, à l'inverse, rendre le calcul plus complexe dans les carreaux avec peu d'individus. Le choix de ne retenir que la structure par sexe et par âge dans la matrice de distances permet par ailleurs de proposer une méthode de carroyage la moins adhérente possible aux variables diffusées.

15. Le nombre de logements estimé dans le recensement est égal au nombre de logements collectés, multiplié par le poids d'estimation.

Par construction, le nombre de logements estimés est égal au nombre de logements du RIL médian. Ainsi, chaque adresse donneuse donne exactement son nombre de logements estimés (car elles ne peuvent donner plus que leur poids d'estimation [16]), et donc son poids d'estimation. La première contrainte du programme d'imputation est donc une égalité avec $coef = 1$: chaque adresse échantillonnée donne exactement son poids d'estimation. Cette variante présente donc la très bonne propriété de préserver toutes les estimations au niveau Iris.
Avec cette méthode, les adresses échantillonnées donnent une fraction de leur poids à environ 4 adresses en moyenne. Les adresses receveuses se voient attribuer une fraction de poids par 1,5 adresse donneuse en moyenne.

2.2.2.2 Méthode d'imputation par hot deck avec cible de population

Avec cette méthode d'imputation, la cible à atteindre est la population auxiliaire de chaque adresse.

- D'une part, chaque adresse échantillonnée peut donner soit jusqu'à son poids d'estimation, soit jusqu'à 1,2 fois, 1,5 fois, 2 fois ou 3 fois plus que son poids d'estimation. Dans les faits, pour que le modèle converge systématiquement [17], la déformation maximale du poids des donneurs doit être strictement supérieure à 1, mais une déformation maximale supérieure à 1,2 n'améliore pas la précision. *In fine*, la variante retenue est la suivante : chaque adresse échantillonnée peut donner jusqu'à 1,2 fois son poids d'estimation.
- D'autre part, chaque adresse reçoit sa population fiscale. Là aussi, le principe de donneurs « stables » est appliqué (chaque donneur reçoit de lui-même sa population auxiliaire).

L'objectif de cette méthode est d'améliorer les performances du modèle en termes de précision :

- d'une part, la population issue des données fiscales est a priori plus corrélée à la population du RP que ne l'est le nombre de logements au RIL médian ;
- d'autre part, le fait que les adresses échantillonnées puissent donner plus que leur poids augmente les degrés de liberté pour l'optimisation.

Avec cette méthode, les adresses échantillonnées donnent une fraction de leur poids à 3 adresses en moyenne. Les adresses receveuses se voient attribuer une fraction de poids par 1,1 adresse donneuse en moyenne.
Contrairement à la méthode d'imputation avec une cible de logements, la méthode avec une cible de population ne garantit pas la cohérence avec les données du recensement diffusées à l'échelle de l'Iris ou de la commune (les adresses donneuses ne donnant pas exactement leur poids). Afin d'assurer une cohérence minimale, un calage sur marges est effectué pour que la population carroyée soit égale à la population légale de la commune et pour que le nombre de logements soit égal au nombre de logements du RIL au niveau communal (paragraphe 3). En revanche, la cohérence avec les autres résultats du recensement (population par sexe, par âge, etc.) n'est pas assurée. Par ailleurs, la cohérence avec la population et le nombre de logements au RIL n'est pas respectée à l'échelle infra-communale de l'Iris.

16. Différents scénarios ont été testés, avec une déformation maximale du poids des donneurs égale à 1,2, 1,5, 2 ou 3. Le modèle le plus précis est celui pour lequel chaque donneur donne exactement son poids d'estimation ($coef = 1$). Par ailleurs, les autres scénarios ne permettent pas une cohérence parfaite avec les données diffusées au niveau Iris.

17. Contrairement au modèle avec cible de logements, nous ne sommes pas dans un cas favorable ici. Rien ne garantit que la somme pondérée de la population fiscale (sur les adresses échantillonnées) soit égale à la population fiscale (contrairement au nombre de logements du RIL). Il est donc nécessaire d'être moins strict pour que le modèle converge.

3 Comparaison des méthodes de carroyage

3.1 Critères de sélection des méthodes

Les trois méthodes de carroyage envisagées (imputation par modélisation, imputation par *hot deck* avec cible de logements, imputation par *hot deck* avec cible de population) sont évaluées à la lumière de plusieurs critères :

- Critères de cohérence, interne et externe :
  - La cohérence interne consiste à ce que la somme des différentes modalités d'une variable soit égale à la population totale : par exemple, le nombre d'hommes estimé et le nombre de femmes estimé doit correspondre à la population totale estimée.
  - La cohérence externe concerne les données diffusées sur les autres niveaux géographiques – notamment les communes, les Iris et les quartiers prioritaires de la politique de la ville [18] (QPV).
- Critères de performance : celle-ci est analysée à travers des indicateurs de précision des estimations, en termes de biais et de variance. Ces indicateurs sont observés pour trois types de variables : la population totale, qui est une variable centrale, les variables directement corrélées à l'information auxiliaire issue des données fiscales utilisée dans nos méthodes de carroyage (sexe, âge), ainsi que les variables moins corrélées à l'information auxiliaire (lieu de naissance, lieu de résidence un an auparavant par exemple).
- Critères de mise en œuvre : au-delà des critères de cohérence et de performance, il est également important de ne pas retenir une méthode trop complexe pour envisager un passage en production et assurer sa pérennité. Les délais de traitement sont également pris en considération.

3.2 Comparaison des 3 méthodes en termes de cohérence interne et externe

3.2.1 Cohérence interne

La cohérence interne (entre les variables) est respectée dans les trois méthodes d'imputation envisagées. Dans les méthodes par *hot deck*, l'imputation concerne l'ensemble des variables d'une adresse recensée, ce qui permet d'avoir naturellement la cohérence entre variables. Dans la méthode par modélisation, c'est le processus de mise en cohérence des estimations qui permet de l'obtenir (partie 2.1.3.1).

3.2.2 Cohérence externe

Deux contraintes de cohérence ont été imposées :

1. la somme des populations carroyées au sein d'une commune1 doit être égale à la population légale de celle-ci ;
2. la somme du nombre de logements carroyés doit être égal au nombre de logements au RIL de la commune.

La cohérence des estimations au carreau avec les diffusions aux mailles infra-communales (Iris et QPV) n'est pas requise (elle imposerait trop de contraintes aux modèles d'imputation). Les trois méthodes satisfont les deux contraintes ci-dessus. La méthode d'imputation par *hot deck* avec

18. Les quartiers prioritaires de la politique de la ville sont des territoires d'intervention du Ministère de la cohésion des territoires et des relations avec les collectivités territoriales, définis par la loi de programmation pour la ville et la cohésion urbaine du 21 février 2014. En métropole, ils ont été identifiés selon un critère unique, celui du revenu par habitant.

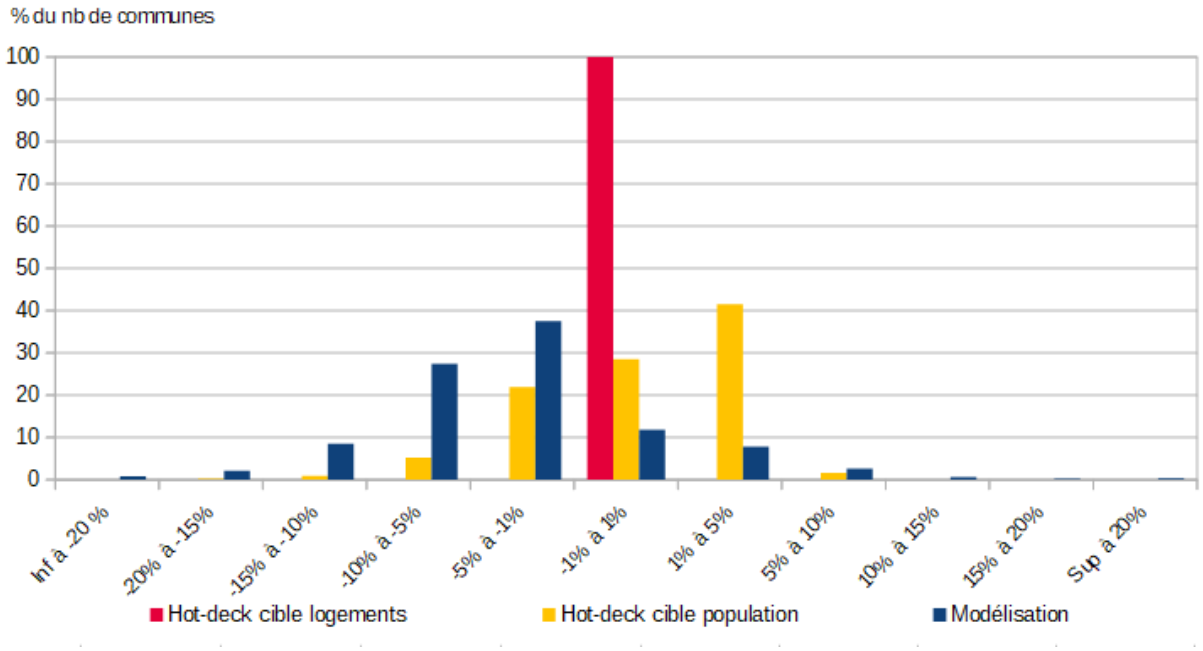
cible de logements y parvient par construction alors que les deux autres méthodes doivent faire appel à un processus de calage sur marges (encadré 3) pour atteindre les objectifs de cohérence au niveau communal.
La méthode d'imputation par *hot deck* avec cible de logements va plus loin car elle assure une cohérence parfaite pour l'ensemble des variables, au niveau Iris et aux niveaux supérieurs. Les deux autres méthodes n'assurent quant à elles qu'une cohérence à la maille communale et uniquement pour les variables de population et de nombre de logements au RIL. Ainsi, avec ces deux méthodes, il peut y avoir des écarts pour les autres variables qui ne participent pas au calage (population par sexe, par âge, par lieu de naissance et lieu de résidence 1 an avant) entre les données carroyées et les données estimées par la méthode « usuelle » au niveau communal. Les graphiques 2 et 3 illustrent ces écarts.

Encadré 3 : Le calage sur marges des méthodes d'imputation par modélisation et d'imputation par *hot deck* avec cible de population

Ces deux méthodes ne permettent pas d'assurer la cohérence des estimations au carreau avec les données de population et de nombre de logements diffusées à une maille supérieure, communale ou infra-communale (Iris). Afin de garantir cette cohérence, un calage sur marges est réalisé. Deux méthodes de calage ont été testées : une méthode de calage de type « raking ratio » et une de type logit bornée. La méthode de type logit bornée permet de borner la déformation des poids, en pratique on essaie de réaliser le calage le plus contraint au départ (10 % de déformation des poids maximum), puis en cas d'échecs on élargit les bornes de déformation jusqu'à ce que le calage réussisse (on revient au raking ratio en cas d'échecs sur tous les jeux de bornes essayés). Cette méthode est retenue car, en limitant la déformation des poids, elle améliore sensiblement la performance des estimations par rapport au calage de type « raking ratio ». Le choix est fait de l'appliquer à l'échelle de la commune, le calage à un niveau infra-communal (Iris) réduisant plus sensiblement la performance des méthodes d'imputation. Par ailleurs, la cohérence avec la population légale de chaque commune est plus importante pour la communication que la cohérence avec des données infra-communales.

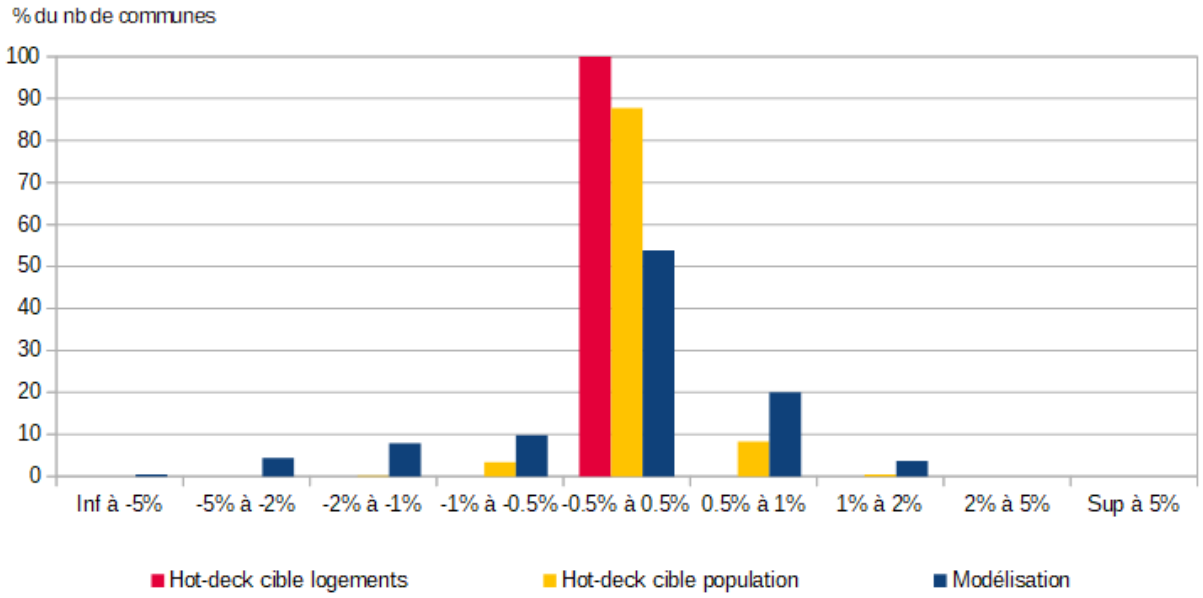
Pour la population des moins de 15 ans, au niveau communal, la méthode par modélisation produit des estimations plus éloignées des estimations usuelles du recensement (encadré 1) que la méthode d'imputation par *hot deck* avec cible de population (graphique 3). Les écarts sont moindres pour la population n'ayant pas changé de lieu résidence au cours de l'année, mais peuvent néanmoins atteindre entre -2 % et -5 % pour 5 % des communes (graphique 4). Par ailleurs, ces résultats montrent bien la parfaite cohérence des estimations de la méthode d'imputation *hot deck* avec cible de logements avec celles du recensement au niveau communal.

Graphique 3 – Écarts entre les estimations des trois modèles d'imputation et les estimations « usuelles » du recensement dans l'ensemble des grandes communes étudiées, à l'échelle des communes, pour la population des moins de 15 ans



Champ : adresses des grandes communes de France métropolitaine, hors changement de géographie.
Source : Recensement de la population 2017 et Fidéli 2017 et 2018.
Lecture : Pour 100 % des communes, il n'y a aucun écart entre les estimations usuelles du recensement de la population et les estimations par *hot deck* avec cible de logements. Pour 29 % des communes, l'écart entre les estimations usuelles du recensement et les estimations par *hot deck* avec cible de population est compris entre -1 % et +1 %. Cette proportion est de 12 % pour la méthode d'imputation par modélisation.

Graphique 4 – Écarts entre les estimations des trois modèles d'imputation et les estimations usuelles du recensement dans l'ensemble des grandes communes étudiées, à l'échelle des communes, pour la population n'ayant pas changé de lieu de résidence au cours de l'année



Champ : adresses des grandes communes de France métropolitaine, hors changement de géographie.
Source : Recensement de la population 2017 et Fidéli 2017 et 2018.
Lecture : Pour 100 % des communes, il n'y a aucun écart entre les estimations usuelles du recensement de la population et les estimations par *hot deck* avec cible de logements. Pour 88 % des communes, l'écart entre les estimations usuelles du recensement et les estimations par *hot deck* avec cible de population est compris entre -1 % et +1 %. Cette proportion est de 54 % pour la méthode d'imputation par modélisation.

Les constats dressés au niveau communal se vérifient également à l'échelle infra-communale des Iris (annexes B et C).

3.3 Comparaison de la performance des trois méthodes en termes de précision

3.3.1 Mise en œuvre de l'évaluation sur des données exhaustives

Les performances respectives d'estimation des 3 méthodes sont évaluées sur 5 communes particulières [19]. Il s'agit de communes ayant récemment franchi le seuil des 10 000 habitants, si bien que l'on dispose à la fois d'une collecte exhaustive récente et d'un RIL. En effet, ces communes étaient considérées jusqu'à très récemment comme des « petites communes », pour lesquelles la collecte du recensement est exhaustive. Ainsi, la vraie valeur pour chaque carreau est connue

19. Les cinq communes retenues pour l'évaluation des méthodes de carroyage sont : Biganos (33051), Juvignac (34123), Haillan (33200), Montévrain (77307), Villebon-sur-Yvette (91661). Ces communes sont les seules à partir desquelles il est possible de mener l'évaluation à l'heure actuelle : au-delà des informations nécessaires au tirage des échantillons, elles peuvent être appariées aux données fiscales et permettent donc de disposer de l'information auxiliaire nécessaire aux méthodes d'imputation.

(au sens où il n'y a pas d'aléa de sondage). La présence nouvelle d'un RIL, constitué lors du franchissement de seuil, permet par ailleurs la géolocalisation des données et la construction des groupes de rotation. Grâce aux informations sur chaque adresse que contient le RIL, nous sommes en outre en mesure de pouvoir simuler différents tirages d'échantillons d'enquêtes annuelles de recensement. Au total, 2 000 échantillons pour 5 EAR consécutives sont simulés dans chaque commune, selon la même méthode que le tirage d'échantillon réalisé dans les grandes communes pour l'enquête annuelle par sondage (encadré 4). On dispose ainsi de 2 000 RP simulés.

Encadré 4 : Le tirage des échantillons dans les communes servant pour l'évaluation des méthodes de carroyage

Les adresses des communes retenues pour l'évaluation des méthodes de carroyage sont réparties en cinq groupes, comme c'est le cas des adresses des communes de 10 000 habitants et plus (encadré 1). Dans chaque groupe, un échantillon d'adresses est tiré, représentant 8 % des logements de la commune, afin de simuler les échantillons de 5 EAR successives. Au sein de chaque groupe de rotation, les adresses de grande taille (60 logements ou plus) constituent une strate exhaustive et sont donc tirées d'office. Les autres adresses sont tirées dans chaque groupe, avec comme variables d'équilibrage le nombre de logements, le nombre de logements en adresse collective et le nombre de logements par Iris (par construction, il n'y a pas d'adresses nouvelles, habituellement enquêtées exhaustivement, dans les 5 communes ayant servi aux simulations). Pour chaque groupe, un vecteur de poids de sondage est obtenu (ajusté par le ratio du nombre de logements par Iris sur la somme des poids des logements échantillonnés). Pour chaque commune, une matrice de poids est ainsi disponible.

3.3.2 Résultats de l'évaluation

À partir des trois méthodes de carroyage envisagées, la population au carreau est estimée pour chaque commune sur les 2 000 RP simulés (population totale, population selon le sexe, la classe d'âge, le lieu de naissance et la population n'ayant pas changé de lieu de résidence au cours de l'année). La moyenne des estimations sur ces 2 000 RP simulés est ensuite mise en regard des données exhaustives du recensement.

Les résultats des différentes méthodes de carroyage sont comparés sur la base de trois indicateurs. L'objectif est d'estimer les performances de ces méthodes comparativement à l'estimateur le plus simple que l'on puisse calculer, celui qui correspond à l'estimateur usuellement utilisé pour la diffusion à l'Iris ou à la commune. Il est appelé dans la suite « estimateur du recensement » et il est calculé à partir des poids de sondage des adresses, calés sur le nombre de logements au RIL, au niveau de l'Iris (encadré 1). Les trois indicateurs sont :

- la racine de l'erreur quadratique moyenne relative (relative root mean squared error, RRMSE), qui rapporte la racine carrée de l'erreur quadratique moyenne [20] de l'estimateur d'une méthode de carroyage à la racine carrée de l'erreur quadratique moyenne de l'estimateur du recensement, représentant ainsi le gain en termes de précision par rapport à l'estimateur du recensement ;
- le biais relatif, qui représente l'écart entre l'estimateur d'une méthode de carroyage et la vraie valeur de cet estimateur, rapportée à cette vraie valeur ;
- l'écart-type relatif, qui mesure la dispersion de l'estimateur d'une méthode de carroyage, rapporté à la vraie valeur de cet estimateur.

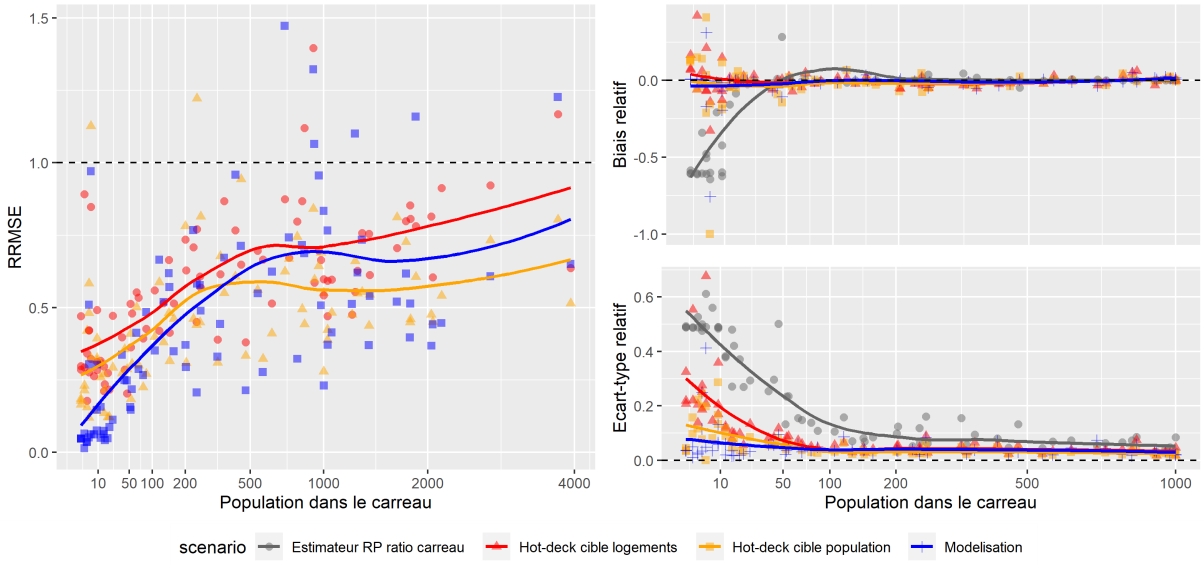
Les résultats de la performance des méthodes de carroyage sont présentés, pour chaque variable, à travers trois graphiques : le graphique de gauche illustre la RRMSE et ceux de droite repré-

20. L'erreur quadratique moyenne (mean squared error, MSE) est égale à la somme du biais de l'estimateur au carré et de sa variance.

sentent le biais relatif (en haut) et l'écart-type relatif (en bas), pour chacune des trois méthodes envisagées. Ces trois indicateurs sont représentés selon la taille des carreaux, en termes de nombre d'habitants [21].

Il ressort que pour chacune des méthodes de carroyage, l'estimation de la population au carreau est plus précise que celle fournie par l'estimateur usuel du recensement. Pour les carreaux les plus « petits » (jusqu'à 400 habitants par carreau environ), pour lesquels l'enjeu du carroyage de la population est le plus fort en termes de fiabilité des résultats, la méthode d'imputation par modélisation est la plus précise : c'est pour cette méthode en effet que la RRMSE est la plus faible (graphique 5). La méthode d'imputation par *hot deck* avec cible de population l'est un peu moins, mais elle est plus performante que la méthode avec cible de logements. La représentation du biais montre que les trois méthodes de carroyage sont moins biaisées que l'estimateur usuel du recensement sur les petits carreaux, et sans biais sur les carreaux plus importants. Par ailleurs, la représentation de l'écart-type montre une plus faible dispersion des estimations des méthodes de carroyage que l'estimateur usuel du recensement : cette dispersion est la plus faible pour les estimations de la méthode d'imputation par modélisation. Au final, les gains de précision de la méthode d'imputation par modélisation sont d'un facteur 8 pour les plus petits carreaux. Ils sont plus faibles pour la méthode d'imputation par *hot deck*, d'un facteur 6 pour la méthode avec cible de population et d'un facteur 4 pour la méthode avec cible de logements.

Graphique 5 – RRMSE, biais relatif et écart-type relatif pour les 3 méthodes de carroyage, pour la variable de population, en fonction de la taille du carreau

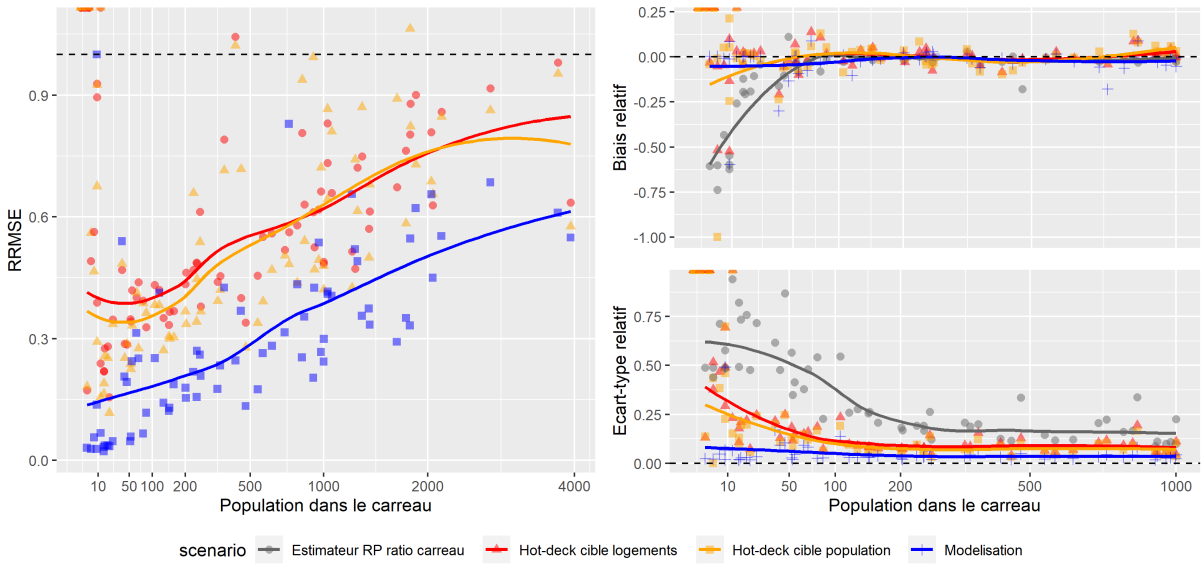


Champ : les carreaux des 5 communes sur lesquelles porte l'évaluation.
Source : Recensement de la population et données fiscales Fidéli.
Lecture : en abscisse, figure la taille du carreau en termes de population au dernier recensement exhaustif. Une RRMSE inférieure à 1 signifie un gain par rapport à la méthode usuelle du recensement appliquée au carreau. Plus la RRMSE est faible, meilleure est la prédiction par rapport à la méthode usuelle du recensement. Plus le biais et l'écart-type sont faibles, meilleure est la prédiction par rapport à la vraie valeur.

21. Pour les cinq communes sur lesquelles porte l'analyse, 25 % des carreaux comptent jusqu'à 20 habitants, la moitié ont entre 20 et 1 000 habitants et 25 % comptent au moins 1 000 habitants.

Pour la population de moins de 15 ans, représentée ici en taux [22], la méthode d'imputation par modélisation est également la plus précise des trois méthodes, quelle que soit la taille des carreaux. Les deux méthodes par *hot deck* apportent le même niveau de précision (graphique 6). Les estimations des trois méthodes de carroyage présentent un biais très faible et sont faiblement dispersées par rapport aux données du recensement.

Graphique 6 – RRMSE, biais relatif et écart-type relatif pour les 3 méthodes de carroyage, pour la part de la population de moins de 15 ans, en fonction de la taille du carreau



Champ : les carreaux des 5 communes sur lesquelles porte l'évaluation.
Source : Recensement de la population et données fiscales Fidéli.
Lecture : en abscisse, figure la taille du carreau en termes de population au dernier recensement exhaustif. Une RRMSE inférieure à 1 signifie un gain par rapport à la méthode usuelle du recensement appliquée au carreau. Plus la RRMSE est faible, meilleure est la prédiction par rapport à la méthode usuelle du recensement. Plus le biais et l'écart-type sont faibles, meilleure est la prédiction par rapport à la vraie valeur.

Les résultats diffèrent en revanche pour des variables moins corrélées avec l'information auxiliaire utilisée pour les estimations et qui ont donc moins de chances d'être bien estimées. Pour l'estimation de la part de la population n'ayant pas changé de lieu de résidence au cours de l'année, les trois méthodes apportent la même précision pour les petits carreaux, précision plus grande que l'estimateur usuel du recensement (graphique 7). Pour les carreaux de taille moyenne ou de grande taille, la méthode d'imputation par modélisation devient en revanche moins précise que l'estimateur du recensement (RRMSE supérieure à 1). Les deux autres méthodes d'imputation par *hot deck* apportent le même niveau de précision. Là aussi, le biais est très faible pour chaque estimateur, de même que l'écart-type. Les résultats sont semblables pour l'estimation de la part de la population née en France [23] : la précision est identique pour les trois méthodes sur les

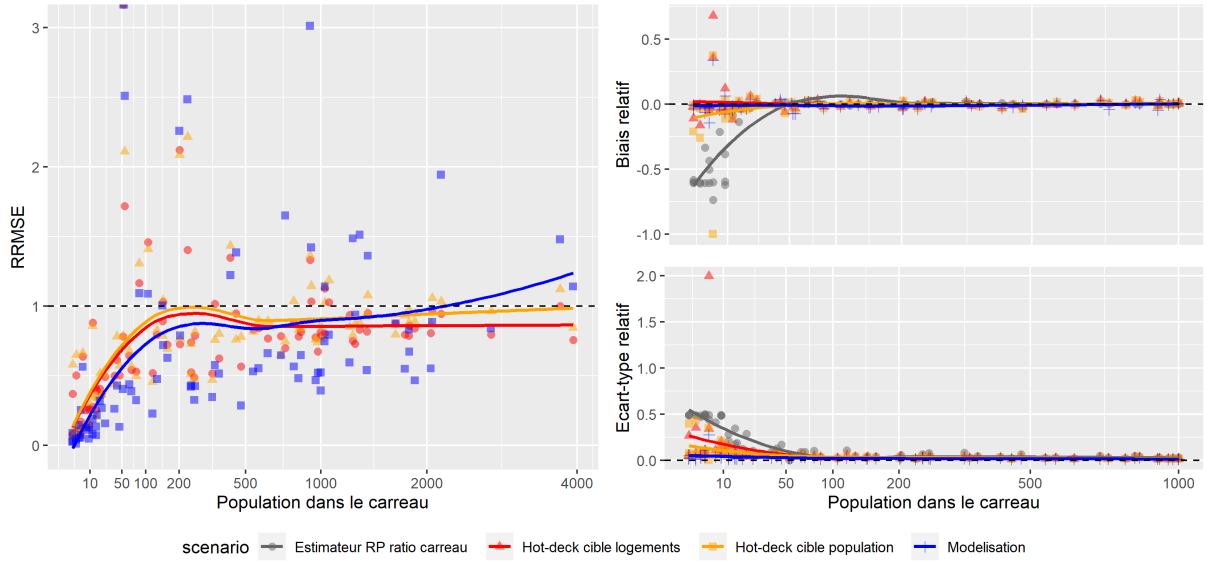
22. La performance sur les variables autres que la population totale est estimée sur les parts de manière à séparer les différences de performance sur les niveaux (mesurée sur la population) et sur les structures.

23. La méthode d'imputation par modélisation estime la population née en France à partir d'une combinaison linéaire des variables de population, population par sexe et population par âge, la variable auxiliaire homologue de la population par pays de naissance n'apportant pas de gain significatif pour la

petits carreaux, tandis que pour les carreaux de taille moyenne ou de grande taille, la méthode d'imputation par modélisation perd nettement en précision (annexe D, graphique D-1). La performance des méthodes a également été étudiée pour la variable de diplôme (non demandée dans le cadre du recensement européen), pour laquelle il n'existe pas d'information homologue dans les données fiscales. Pour l'estimation de la part de personnes diplômées du baccalauréat [24], la méthode d'imputation offre la meilleure précision (annexe D, graphique D-2).

Au global, les résultats montrent que la méthode d'imputation par modélisation offre la meilleure précision des estimations au carreau, sauf pour des variables dont les modalités sont peu dispersées (lieu de résidence un an auparavant par exemple).

Graphique 7 – RRMSE, biais relatif et écart-type relatif pour les 3 méthodes de carroyage, pour la part de la population n'ayant pas changé de lieu de résidence au cours de l'année, en fonction de la taille du carreau



Champ : les carreaux des 5 communes sur lesquelles porte l'évaluation.
Source : Recensement de la population et données fiscales Fidéli.
Lecture : en abscisse, figure la taille du carreau en termes de population au dernier recensement exhaustif. Une RRMSE inférieure à 1 signifie un gain par rapport à la méthode usuelle du recensement appliquée au carreau. Plus la RRMSE est faible, meilleure est la prédiction par rapport à la méthode usuelle du recensement. Plus le biais et l'écart-type sont faibles, meilleure est la prédiction par rapport à la vraie valeur.

3.4 Comparaison de la performance des trois méthodes en termes de facilité de mise en œuvre

La méthode d'imputation par modélisation repose sur une méthode assez simple, de régressions linéaires et de calage sur marges. Sa « difficulté » réside dans le fait de devoir relancer l'ensemble de la modélisation pour chaque ajout de nouvelle variable à carroyer (définition de la séquence

prédiction (cf. paragraphe 2.1.2.).

24. Le baccalauréat correspond à un des diplômes marquant la fin des études secondaires. Ce niveau de diplôme est intégré dans la modalité « ISCED 3 » de la nomenclature européenne du niveau éducatif.

des estimations en amont). La méthode d'imputation par *hot deck* est quant à elle un peu plus complexe à mettre en œuvre car elle repose sur un problème d'optimisation linéaire. La variante avec cible de population implique par ailleurs un calage sur marges. Avec cette méthode, l'ensemble des variables du recensement sont disponibles. Les délais de traitement sont également un critère de sélection : le temps d'exécution est légèrement plus important pour les méthodes par *hot deck*, mais reste acceptable (quelques heures).

3.5 Synthèse des résultats

Dans cette partie, nous présentons une synthèse comparative des trois méthodes d'imputation (tableau). Les résultats montrent que deux types de critères s'opposent : les critères de précision des estimations d'une part et, d'autre part, les critères de cohérence avec les données diffusées par ailleurs et la possibilité de carroyer facilement d'autres variables que celles définies en amont. L'arbitrage entre les différentes méthodes dépend donc du degré d'importance accordée à la précision des estimations ou à la cohérence avec les données publiées à d'autres niveaux géographiques, ainsi que de la volonté de diffuser des données nationales au carreau sur plus de variables que celles prévues dans le cadre du Census 2021.

Tableau 5 – Synthèse comparative des trois modèles d'imputation, selon les critères de sélection définis

Critères de sélection	Méthode d'imputation par modélisation	Méthode d'imputation par *hot deck* avec cible de logements issue du RIL	Méthode d'imputation par *hot deck* avec cible de population issue des données fiscales
Critères de cohérence :			
Cohérence entre variables	Assurée	Parfaite	Parfaite
Cohérence avec les données diffusées du recensement de la population	Uniquement sur la population et le nombre de logements à la commune	Totale au niveau IRIS. Respect du nombre de logements du RIL à chaque adresse.	Uniquement sur la population et le nombre de logements à la commune
Critères de précision :			
Performance pour la variable de population	+++++	+++	++++
Performance pour les autres variables corrélées à l'information auxiliaire	+++	++	++
Performance pour les variables non corrélées à l'information auxiliaire	+	+	+
Critères de mise en oeuvre de la méthode :			
Complexité de la méthode	Moyenne	Moyenne	Moyenne
Champ des variables carroyées	Seulement les variables prédéfinies. L'ajout de variables supplémentaires peut être complexe.	Toutes les variables du RP	Toutes les variables du RP
Temps d'exécution	Quelques minutes par variable	Quelques heures pour toutes les variables	Quelques heures pour toutes les variables

Conclusion

Les méthodes d'imputation envisagées permettent toutes d'estimer la population au carreau de manière plus fiable que ne le fait l'estimateur usuel du recensement de la population. Chaque méthode présente des atouts et faiblesses au regard des critères de sélection.

- Les trois méthodes respectent la cohérence des estimations avec les données du recensement à l'échelle communale, après un calage sur marges pour deux d'entre elles. En revanche, seule la méthode d'imputation par *hot deck* avec cible de logements assure la cohérence avec les données publiées à l'échelle infra-communale des Iris. Pour les deux autres méthodes, il est possible d'imposer des contraintes de cohérence au niveau Iris, mais cela se fait au détriment de la précision.
- La méthode d'imputation par modélisation permet d'avoir les estimations les plus précises pour la variable de population et pour les variables corrélées à l'information auxiliaire (sexe, âge). Pour les variables non corrélées à l'information auxiliaire (lieu de naissance, lieu de résidence antérieure), la méthode d'imputation par modélisation perd un peu en précision.
- Par ailleurs, la méthode d'imputation par modélisation est moins flexible quant à l'ajout de nouvelles variables. Ce point entre en considération dans la perspective du futur règlement européen ESOP, qui pourrait demander des estimations au carreau pour un spectre plus large de variables que le Census 2021.

Au regard de ces différents critères, l'Insee a choisi de privilégier la méthode d'imputation par *hot deck* avec cible de logements. Dans la mesure où l'utilisation première de données carroyées est la diffusion de cartographies, la cohérence des estimations à différents niveaux géographiques paraît plus importante que la précision des estimations. Par ailleurs, en vue d'une diffusion nationale des données et de la réponse au futur règlement européen ESOP, il paraît important de pouvoir facilement généraliser la méthode à d'autres variables. La rigidité des modèles d'estimation dans la méthode d'imputation par modélisation constituait un frein à cet égard. Enfin, la moindre adhérence des estimations aux données auxiliaires dans le cas de la méthode par *hot deck* avec cible logement (où ces dernières n'interviennent que dans le calcul de la distance) permet de moins distordre la structure des données du recensement.

Bibliographie

[1] EUROSTAT (2019), "EU legislation on the 2021 population and housing censuses, explanatory notes". Theme Population and social conditions, Collection Manuals and guidelines. Février 2019.

[2] Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km² grid. OJ L 296, 22.11.2018, p. 19–27.

[3] Ardilly P., « Panorama des principales méthodes d'estimation sur les petits domaines », Documents de travail Insee n° M0602, septembre 2006.

[4] Gallic G., Pagès J. (2022), « La géolocalisation du recensement de la population dans les communes métropolitaines de moins de 10 000 habitants », communication aux Journées de méthodologie statistique de l'Insee de 2022.

Annexes

Annexe A : L'estimation de données carroyées dans les communes de moins de 10 000 habitants

Les communes de moins de 10 000 habitants sont réparties en cinq groupes de rotation et l'enquête annuelle de recensement y est réalisée exhaustivement de façon tournante : chaque année, les communes d'un des groupes de rotation sont recensées exhaustivement, si bien qu'en 5 ans, toutes les communes de moins de 10 000 habitants ont été recensées intégralement. Malgré cette enquête tournante, les résultats sont produits chaque année pour toutes les communes. Différentes méthodes d'estimation sont mobilisées : extrapolation et interpolation. À la fin de l'année N, on calcule les populations municipales en date du 1^{er} janvier N-2. Si la commune a été recensée en N-2, la population est obtenue directement en sommant le nombre de bulletins de recensement. Si la commune a été recensée en N-3 ou N-4, on extrapole le résultat de la collecte à partir de l'évolution du nombre de logements constatée dans les données fiscales combinées à une estimation de la variation de la taille moyenne des ménages. Si la commune a été recensée en N-1 ou en N, on effectue une interpolation linéaire entre la dernière population légale (recensement N-1) et la nouvelle collecte.

Le recensement étant exhaustif dans ces communes, le carroyage de la population ne pose pas de problème spécifique dès lors que les données sont géolocalisés (*cf.* Gallic et Pagès, JMS 2022). En effet, chaque année, les données de chaque commune sont disponibles au niveau adresse à l'aide des méthodes exposées ci-dessus. Les données carroyées s'obtiennent alors par simple sommation. L'hypothèse sous-jacente est que la variation de population estimée au niveau communal est identique sur tous les carreaux. Il s'agit là d'un proxy acceptable, dans la mesure où, d'après les données fiscales, l'évolution de la population dans les carreaux d'une commune donnée est globalement proche de l'évolution de la population dans l'ensemble de la commune. En effet, la moitié des carreaux ont une évolution de population distante de moins de 5 points en valeur absolue par rapport à l'évolution de population au sein de leur commune [25] (tableau A-1). Cette proportion varie selon la taille des carreaux : elle est plus forte dans les petits carreaux que dans les grands. Pour les carreaux comprenant entre 5 et 9 habitants, la moitié des carreaux ont une évolution de population distante de 10 points ; elle de 8,5 points pour la moitié des carreaux comptant entre 10 et 19 habitants. Pour les très petits carreaux en revanche (moins de 5 habitants), l'évolution de population est distante de 3 points pour la moitié d'entre eux. Pour les carreaux les plus grands, de plus 200 habitants, l'écart médian d'évolution de la population au carreau et à la commune est de 2 points. Ces points de pourcentage peuvent paraître élevés, mais les écarts d'effectifs sont en réalité très faibles (tableau A-2). Ainsi, en appliquant le taux d'évolution de la population communale aux carreaux, l'écart en nombre d'individus par rapport à la donnée fiscale observée est de 1,4 habitant pour la moitié des carreaux. Cet écart médian est très faible pour les petits carreaux (inférieur à 1 habitant) et s'élève à 9 habitants environ pour les plus grands carreaux.

25. On se restreint ici aux carreaux porteurs de population qui ne sont pas à cheval sur plusieurs communes (soit environ 80 % des carreaux porteurs de population en petite commune).

Tableau A-1 – Quantiles des écarts en valeur absolue de l'évolution 2018-2019 de la population issue des sources fiscales au carreau et à la commune, dans les communes de moins de 10 000 habitants, selon la taille des carreaux

Taille des carreaux*	Nombre de carreaux	Quantiles des écarts en valeur absolue de l'évolution de la population entre 2018 et 2019 entre le carreau et la commune (en points de pourcentage)							
		0 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %
Moins de 5 habitants	55 211	0,0	0,2	0,4	1,3	3,1	24,6	54,2	100,3
5-9 habitants	44 310	0,0	0,3	0,7	2,0	10,0	21,5	41,2	57,6
10-19 habitants	49 311	0,0	0,5	1,0	3,2	8,5	17,1	28,7	37,8
20-49 habitants	56 596	0,0	0,5	1,1	3,0	6,6	12,2	19,5	25,4
50-99 habitants	31 267	0,0	0,3	0,8	2,0	4,6	8,2	12,9	16,9
100-199 habitants	23 022	0,0	0,2	0,5	1,4	3,2	5,9	9,4	12,6
200 habitants et plus	30 226	0,0	0,2	0,3	0,9	2,0	3,9	6,5	9,1
Ensemble des carreaux	289 943	0,0	0,3	0,6	1,8	4,8	12,4	27,9	45,2

* Taille des carreaux en 2018
Champ : carreaux des communes de moins de 10 000 habitants porteurs de population en 2018 et en 2019 n'étant pas à cheval sur plusieurs communes.
Source : Fidéli 2018 et 2019.
Lecture : pour la moitié des carreaux, l'écart entre l'évolution de la population au carreau et l'évolution de la population au niveau de la commune entière est inférieur à 4,8 points en valeur absolue.

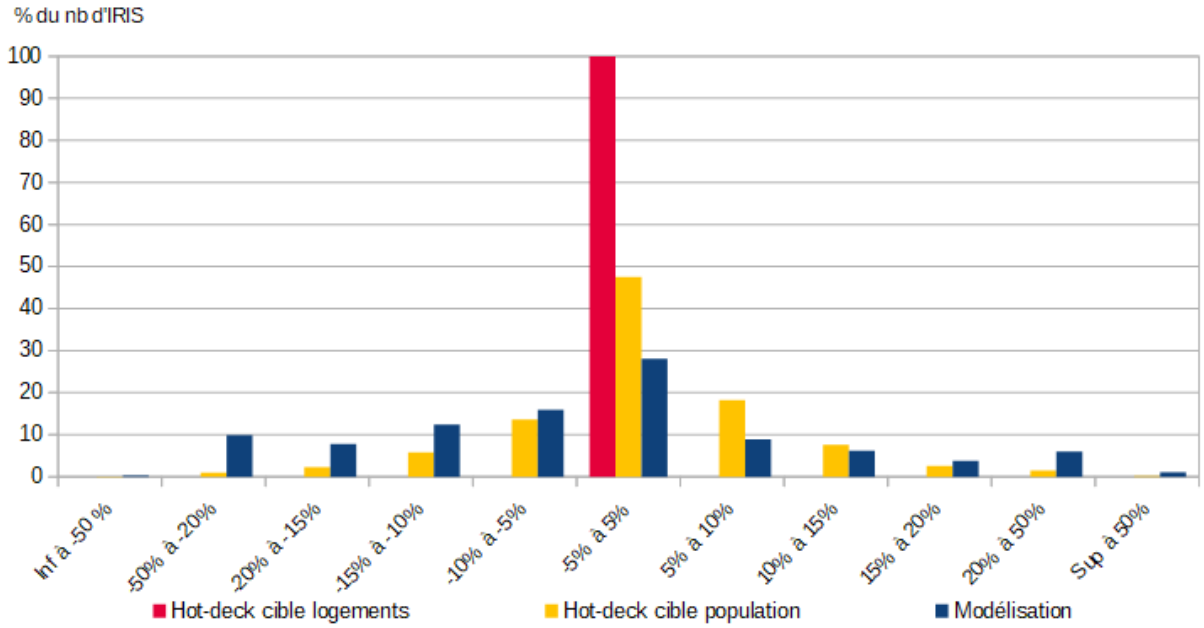
Tableau A-2 – Quantiles des écarts en valeur absolue de l'évolution 2018-2019 de la population issue des sources fiscales au carreau et à la commune, dans les communes de moins de 10 000 habitants, selon la taille des carreaux

Taille des carreaux*	Nombre de carreaux	Quantiles des écarts en valeur absolue au niveau carreau entre le nombre de personnes observées dans les données fiscales en 2019 et le nombre de personnes estimées en appliquant à chaque carreau l'évolution 2018-2019 de la population communale (en points de pourcentage)							
		0 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %
Moins de 5 habitants	55 211	0,0	0,0	0,0	0,0	0,1	0,9	1,5	2,1
5-9 habitants	44 310	0,0	0,0	0,0	0,1	0,7	1,4	2,9	3,9
10-19 habitants	49 311	0,0	0,1	0,1	0,4	1,1	2,3	4,0	5,1
20-49 habitants	56 596	0,0	0,2	0,3	0,9	2,0	3,8	6,0	7,7
50-99 habitants	31 267	0,0	0,2	0,5	1,4	3,1	5,7	9,0	11,8
100-199 habitants	23 022	0,0	0,3	0,7	2,0	4,4	8,2	13,1	17,5
200 habitants et plus	30 226	0,0	0,7	1,4	3,8	8,9	17,8	33,7	52,3
Ensemble des carreaux	289 943	0,0	0,0	0,0	0,3	1,4	3,8	8,2	13,6

* Taille des carreaux en 2018
Champ : carreaux des communes de moins de 10 000 habitants porteurs de population en 2018 et en 2019 n'étant pas à cheval sur plusieurs communes.
Source : Fidéli 2018 et 2019.
Lecture : pour la moitié des carreaux, l'écart en nombre d'individus en appliquant à chaque carreau l'évolution 2018-2019 de la population communale par rapport à la donnée fiscale observée est de 1,4 habitants.

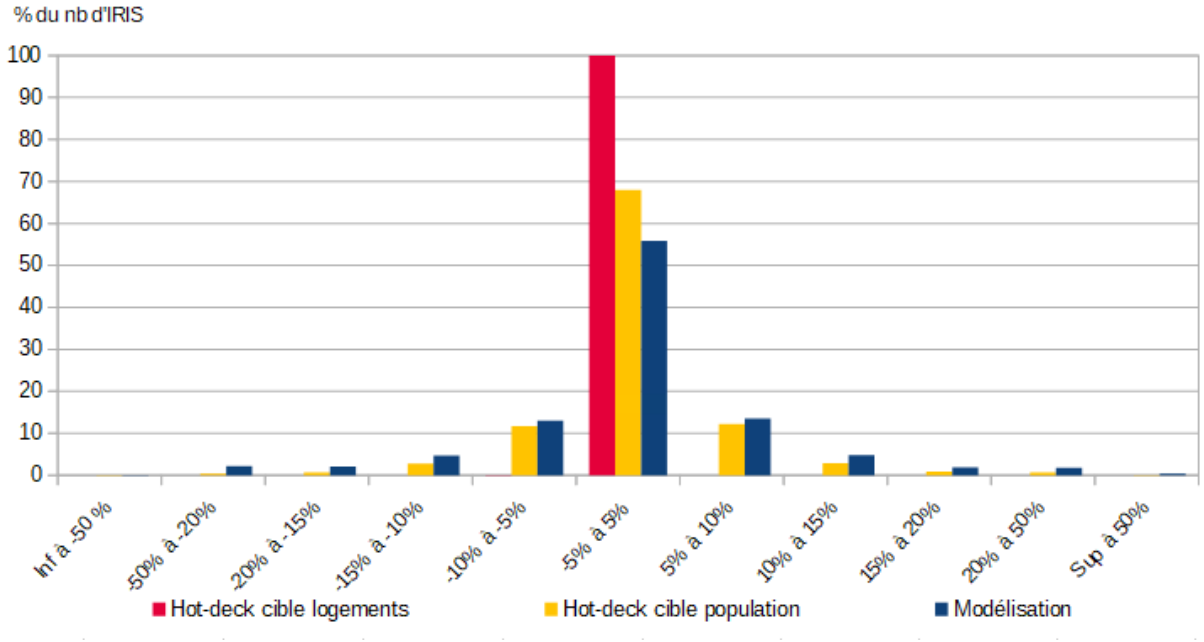
Annexe B : Écarts entre les estimations des trois modèles d'imputation et les estimations usuelles du recensement dans l'ensemble des grandes communes étudiées, à l'échelle des Iris

Graphique B-1 – Écarts entre les estimations des trois modèles d'imputation et les estimations usuelles du recensement dans l'ensemble des grandes communes étudiées, à l'échelle des Iris, pour la population des moins de 15 ans



Champ : les adresses des grandes communes de France métropolitaine, hors changement de géographie.
Source : Recensement de la population 2017 et Fidéli 2017 et 2018.
Lecture : Pour 100 % des Iris, il n'y a aucun écart entre les estimations usuelles du recensement de la population et les estimations par *hot deck* avec cible de logements. Pour 47 % des Iris, l'écart entre les estimations usuelles du recensement et les estimations par *hot deck* avec cible de population est compris entre -5 % et +5 %. Cette proportion est de 28 % pour la méthode d'imputation par modélisation.

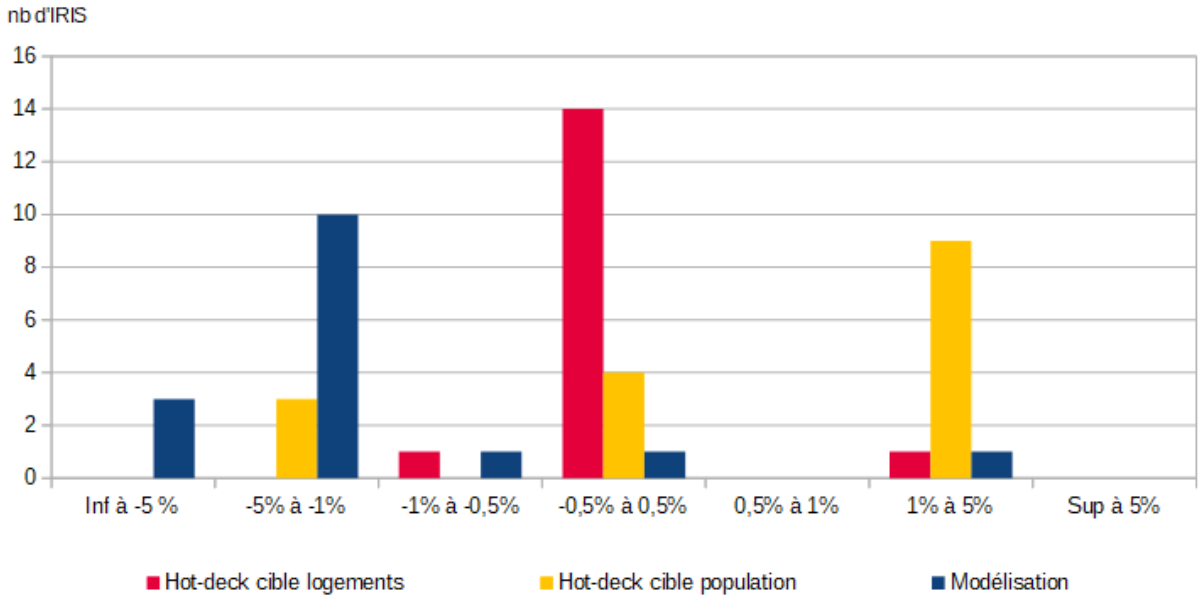
Graphique B-2 – Écarts entre les estimations des trois modèles d'imputation et les estimations « usuelles » du recensement dans l'ensemble des grandes communes étudiées, à l'échelle des Iris, pour la population n'ayant pas changé de lieu de résidence au cours de l'année



Champ : les adresses des grandes communes de France métropolitaine, hors changement de géographie.
Source : Recensement de la population 2017 et Fidéli 2017 et 2018.
Lecture : Pour 100 % des Iris, il n'y a aucun écart entre les estimations usuelles du recensement de la population et les estimations par *hot deck* avec cible de logements. Pour 68 % des Iris, l'écart entre les estimations usuelles du recensement et les estimations par *hot deck* avec cible de population est compris entre -5 % et +5 %. Cette proportion est de 56 % pour la méthode d'imputation par modélisation.

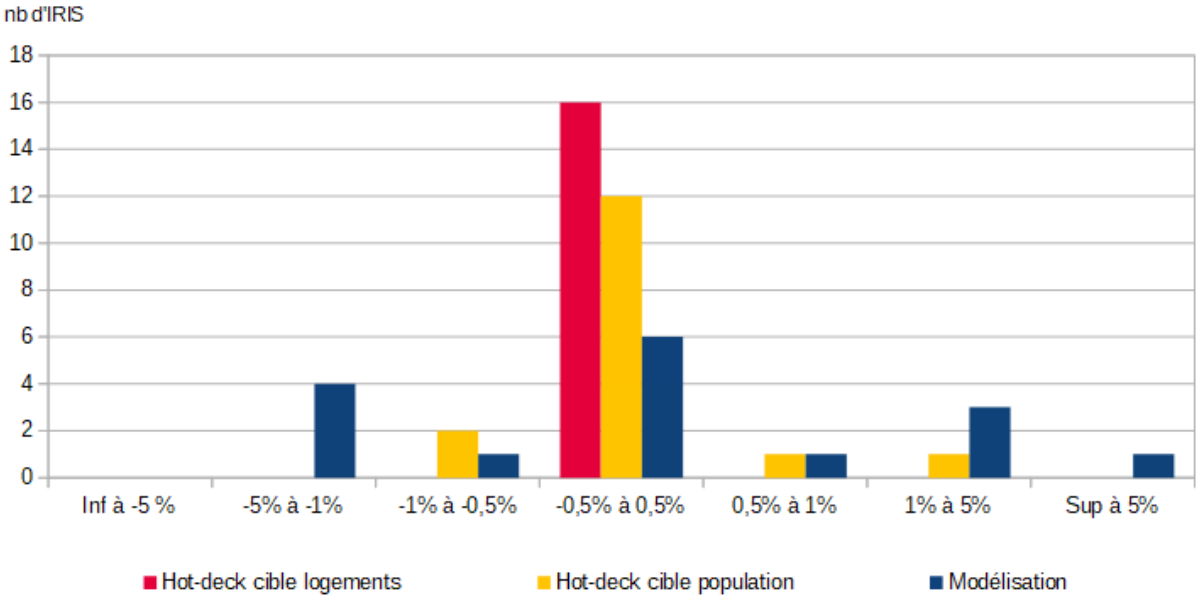
Annexe C : Écarts entre les estimations des trois modèles d'imputation et les données exhaustives du recensement pour les cinq communes servant à l'évaluation des méthodes, à l'échelle des Iris

Graphique C-1 – Écarts entre les estimations des trois modèles d'imputation et les données exhaustives du recensement pour les cinq communes retenues pour l'évaluation, à l'échelle des Iris, pour la population des moins de 15 ans



Champ : les 5 communes retenues pour l'évaluation.
Source : Recensement de la population et Fidéli.
Lecture : Pour 14 des 16 Iris des 5 communes servant pour l'évaluation des méthodes de carroyage, il n'y a aucun écart entre la population collectée exhaustivement et les estimations par *hot deck* avec cible de logements (cet écart s'explique par le fait que l'on compare ici par rapport à la vraie valeur, et non par rapport à l'estimation habituelle du recensement). Pour 4 Iris, l'écart entre la population collectée exhaustivement et les estimations par *hot deck* avec cible de population est compris entre -0,5 % et 0,5 % de la population. C'est le cas pour 1 Iris pour la méthode d'imputation par modélisation.

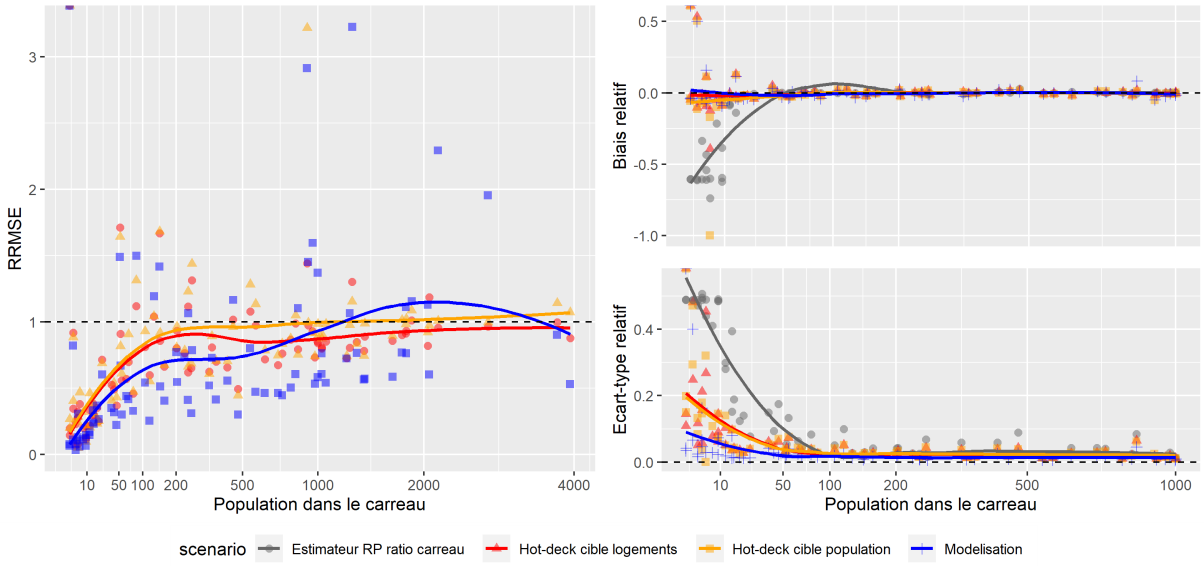
Graphique C-2 – Écarts entre les estimations des trois modèles d'imputation et les données exhaustives du recensement pour les cinq communes retenues pour l'évaluation, à l'échelle des Iris, pour la population n'ayant pas changé de lieu de résidence au cours de l'année



Champ : les 5 communes retenues pour l'évaluation.
Source : Recensement de la population et Fidéli.
Lecture : Pour la totalité des Iris des 5 communes servant pour l'évaluation des méthodes de carroyage, il n'y a aucun écart entre la population collectée exhaustivement et les estimations par *hot deck* avec cible de logements. Pour 12 Iris, l'écart entre la population collectée exhaustivement et les estimations par *hot deck* avec cible de population est compris entre -0,5 % et 0,5 % de la population. C'est le cas pour 6 Iris pour la méthode d'imputation par modélisation.

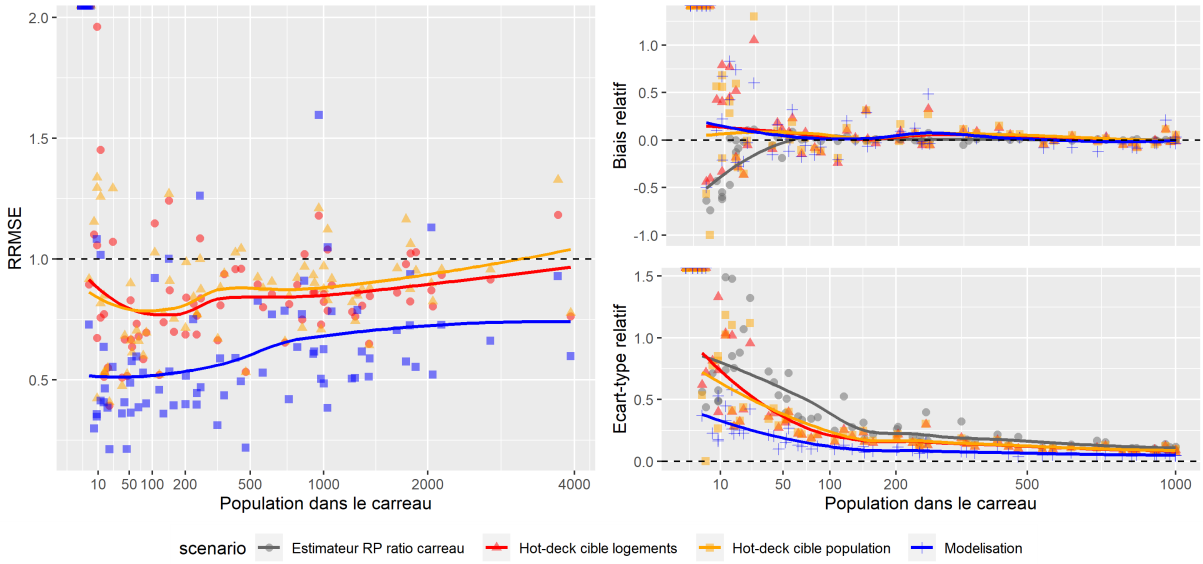
Annexe D : Analyse de la précision des trois méthodes de carroyage, pour la part des personnes nées en France et pour la part des personnes diplômées du baccalauréat

Graphique D-1 – RRMSE, biais relatif et écart-type relatif pour les 3 méthodes de carroyage, pour la part des personnes nées en France, en fonction de la taille du carreau



Champ : les carreaux des 5 communes sur lesquelles porte l'évaluation.
Source : Recensement de la population et données fiscales Fidéli.
Lecture : en abscisse, figure la taille du carreau en termes de population au dernier recensement exhaustif. Une RRMSE inférieure à 1 signifie un gain par rapport à la méthode usuelle du recensement appliquée au carreau. Plus la RRMSE est faible, meilleure est la prédiction par rapport à la méthode usuelle du recensement. Plus le biais et l'écart-type sont faibles, meilleure est la prédiction par rapport à la vraie valeur.

Graphique D-2 – RRMSE, biais relatif et écart-type relatif pour les 3 méthodes de carroyage, pour la part des personnes diplômées du baccalauréat, en fonction de la taille du carreau



Champ : les carreaux des 5 communes sur lesquelles porte l'évaluation.
Source : Recensement de la population et données fiscales Fidéli.
Lecture : en abscisse, figure la taille du carreau en termes de population au dernier recensement exhaustif. Une RRMSE inférieure à 1 signifie un gain par rapport à la méthode usuelle du recensement appliquée au carreau. Plus la RRMSE est faible, meilleure est la prédiction par rapport à la méthode usuelle du recensement. Plus le biais et l'écart-type sont faibles, meilleure est la prédiction par rapport à la vraie valeur.