Chroniques

Conseil national de l'information statistique

Vers un carroyage des données du recensement

La demande d'accès à des données finement localisées s'exprime régulièrement lors de commissions et événements organisés par le Conseil national de l'information statistique (Cnis), en phase avec l'avis général n° 5 du Moyen terme 2019-2023. Si la commission Territoires est celle où ce besoin est plus souvent évoqué, comme dans l'avis de la séance du 25 novembre 2021, qui « encourage tout particulièrement les efforts de géolocalisation, pour répondre au besoin croissant de données finement localisées dans de nombreux domaines de l'action publique » (compte rendu de la réunion, p. 21), elle est loin d'être le seul lieu d'expression de ce besoin. La commission Environnement et développement durable, ainsi que les commissions sociales confirment l'importance de ces données dans des domaines aussi différents que la culture, l'emploi, le développement durable… Lors du séminaire du Cnis sur le questionnaire et la diffusion des résultats du recensement de la population, qui s'est tenu le 7 octobre 2020, des chercheurs et des représentants des agences d'urbanisme ainsi que des collectivités locales ont souligné leur souhait d'avoir accès à des données du recensement carroyées, voire géolocalisées, en indiquant que les données à l'IRIS (îlots regroupés pour l'information statistique : un découpage de la majorité des communes de plus de 5 000 habitants en zones de taille homogène d'environ 2 000 habitants) sont utiles, mais insuffisantes eu égard à leurs besoins.

Des données finement localisées indispensables aux politiques locales

Dès le 14 octobre 2003, une réunion sur les statistiques locales et régionales[1] illustrait la diversité des utilisateurs et des données finement localisées. De même, l'avis n° 5 de la Commission nationale d'évaluation du recensement de la population (Cnerp) du Moyen terme 2019-2023 met en exergue le besoin d'adapter le recensement aux exigences et demandes externes, y compris celles concernant la production de données finement localisées. Ce besoin de données localisées s'exprime aussi indirectement au Cnis par les demandes croissantes d'accès à des données administratives au titre de l'article 7bis de la loi de 1951 sur l'obligation, la coordination et le secret en matière de statistiques qui permettent souvent la production d'information localisée, notamment en recourant à des appariements de fichiers. Le sujet des appariements et des réflexions que ces opérations entraînent, a d'ailleurs fait l'objet de la rencontre du Cnis qui a eu lieu le 28 janvier 2022[2].

Malgré cette demande importante et croissante, peu de sources statistiques (par exemple le recensement de la population et certains dispositifs fondés sur des données administratives) permettent de diffuser des informations à un niveau géographique fin, comme la commune ou l'EPCI (établissement public de coopération intercommunale). La demande portée par le séminaire de 2020 sur le recensement de la population vise principalement le niveau infra-communal, voire des données individuelles géolocalisées.

Les zonages infra-communaux (îlot, jusqu'en 1999, date du dernier recensement exhaustif, mais surtout IRIS et QPV[3]) permettent déjà aujourd'hui de diffuser régulièrement des données à partir du recensement de la population. Le passage à un recensement tournant a réduit légèrement la

1. https://www.cnis.fr/wp-content/uploads/2018/02/CR_2003_FORM_statistiques_locales.pdf
2. https://www.cnis.fr/evenements/appariements-de-donnees-individuelles-entre-richesse-de-linformation-statistique-et-respect-de-la-vie-privee/?category=1067
3. Quartier prioritaire de la politique de la ville.

Chroniques n° 28
Mai 2022
Directrice de la publication : Françoise Maurel
Rédacteur en chef : François Guillaumat-Tailliet
Maquette : Brigitte Rols
Publication diffusée gratuitement,
ne peut être vendue
ISBN 978 -2-11-162339-2

finesse de la diffusion, du fait de la suppression de la diffusion à l'îlot. Les IRIS sont de surface variable selon le niveau de dispersion de la population sur le territoire concerné pour atteindre la taille moyenne de 2 000 habitants visée à leur construction. Le zonage de diffusion infra-communale repose sur un arbitrage entre d'une part la finesse de la zone d'intérêt et d'autre part la robustesse des données et le respect de la confidentialité.

Pour compléter ce dispositif de diffusion infra-communale à partir du recensement de la population, l'Insee produit également des indicateurs sur les logements et les individus sur des zonages à façon, à la demande d'organismes ayant une mission de service public. Ces territoires doivent être situés au sein d'une ou plusieurs communes de plus de 10 000 habitants et comprendre au moins 1 000 logements dans chaque composante communale. Ce service de « Diffusion infra-communale à façon » (Diaf-RP) a récemment évolué : il permet désormais la diffusion d'indicateurs sur des zonages à façon à cheval sur plusieurs communes, il est disponible dans les DOM (hors Mayotte) et permet des analyses en évolution sur un pas de 5 ans. L'Insee poursuit les travaux d'amélioration de ce service pour l'étendre à l'ensemble des communes.

La diffusion de données carroyées : une réponse aux demandes de données infra-communales

A côté de ces découpages aux contours *ad hoc*, la diffusion de données sur une grille géométrique fixe, dites aussi données carroyées, s'impose de plus en plus. Ainsi, l'Insee a diffusé à trois reprises des données démographiques et socio-économiques à partir des sources administratives et fiscales : en 2013 en 2019 et très récemment en 2022[4] . Des informations telles que le nombre d'individus présents dans ces fichiers, le nombre de logements sociaux, le nombre de ménages propriétaires ou encore le nombre de ménages pauvres ont été diffusées sur des carreaux de 200 mètres de côté.

Le carroyage présente de multiples avantages. Le premier est la stabilité du découpage dans le temps. Alors que les zonages administratifs (commune, arrondissements, etc.) ou statistiques (IRIS, îlots, etc.) peuvent évoluer au fil des modifications géographiques afin de s'adapter aux évolutions du terrain ou des besoins changeants, la grille géométrique peut être fixée une fois pour toutes. Cela facilite grandement les analyses en évolution. Les carreaux assurent également une meilleure comparabilité spatiale, car chaque carreau dispose des mêmes caractéristiques géographiques : même forme, même surface notamment. Une critique régulièrement exprimée est que la construction de la grille géographique ne se fonde pas sur des éléments physiques présents sur le territoire, comme des limites naturelles (fleuve, forêt) ou artificielles (route, voie ferrée, etc.). Toutefois, cela en constitue également l'une des forces dès lors que ces frontières réelles ne sont pas pertinentes pour ce que l'on cherche à analyser (analyse d'une zone autour d'une infrastructure, comme une gare, population exposée à un bruit, etc.). Il faut voir les carreaux comme la brique de base (petite maille géographique permettant de diffuser des données de qualité tout en respectant la confidentialité) permettant d'approcher le zonage d'intérêt [Darriau, 2020]. Plus les carreaux sont petits, plus l'approximation d'un zonage par agrégation de carreaux est de bonne qualité.

Eurostat s'est récemment inscrit dans cette optique en demandant aux États membres la fourniture de données carroyées[5] du recensement . Ceux-ci devront ainsi mettre à disposition d'Eurostat en décembre 2022 des populations provisoires relatives à l'année 2021 sur une grille de carreaux de 1 km de côté. En mars 2024, devront être transmises une dizaine de variables de population relatives à l'année 2021 sur cette même grille : population par sexe, par âge (moins de 15 ans, 15-64 ans et 65 ans ou plus), par lieu de naissance (France, autre pays de l'Union européenne, autre), par lieu de résidence habituelle un an auparavant (même résidence, déplacement à l'intérieur de la France, déplacement depuis l'extérieur de la France) et, *dans la mesure du possible*, population en emploi. Au-delà des règlements encadrant le recensement européen de 2021, des discussions sont en cours autour d'un projet de nouveau règlement européen sur les statistiques de population et de logements (ESOPH – European statistics on population and housing) fusionnant plusieurs règlements actuellement en vigueur. Ce projet devrait prévoir en outre la fourniture de données carroyées tous les ans par les États membres. La demande européenne de disposer de données carroyées s'inscrit donc dans la durée.

Précision et respect de la confidentialité : deux préalables à la diffusion de données finement localisées

La production et l'utilisation de données carroyées se heurtent cependant à deux limites importantes. La première est la qualité des données produites. Si le carroyage de sources administratives exhaustives ne pose pas de problème de qualité, au-delà de la qualité de la source elle-même, le carroyage de fichiers non exhaustifs – comme le recensement dans les communes de plus de 10 000 habitants, effectué par sondage – repose sur un modèle d'estimation nécessairement imparfait, basé parfois sur peu de personnes enquêtées ou faisant intervenir de l'information auxiliaire hors du domaine d'estimation – en l'occurrence le carreau. À cela s'ajoute la nécessité de disposer d'une géolocalisation très précise et stable dans le temps des unités d'intérêt. Un faible écart entre la géolocalisation dans la source et la localisation sur le terrain peut conduire à ce que l'estimation d'une grandeur sur un carreau s'éloigne de la réalité : cas rencontrés parfois lors de carroyage de sources fiscales par exemple, où la géolocalisation peut être obtenue via l'étiquette de la parcelle cadastrale et non le bâtiment ou l'adresse d'habitation. De même, un changement de géolocalisation d'un grand bâtiment d'une année sur l'autre peut conduire à une forte variation des estimations pour un carreau donné. Cette variation ne devra pas être interprétée comme une évolution réelle, mais comme une modification de la qualité de la source

4. https://www.insee.fr/fr/statistiques/6215217

5. Plus précisément, la Commission européenne a adopté un règlement d'exécution relatif à l'établissement d'une « action statistique directe temporaire » pour la diffusion de thèmes sélectionnés du recensement de la population et des logements de 2021 géocodés selon une grille de 1 km (règlement n° 2018/1799 du 21 novembre 2018).

perceptible du fait de la finesse des données diffusées. L'utilisation de données carroyées fines requiert donc de la prudence, tant dans l'analyse en coupe qu'en évolution.
La deuxième limite est celle de la confidentialité. Diffuser des données sur des espaces de taille très réduite conduit à travailler sur des effectifs de population peu nombreux et risque d'induire une rupture du secret statistique. Il est ainsi impératif d'assurer la confidentialité des données, via un traitement spécifique, afin en particulier d'éviter l'inférence d'une caractéristique sur l'ensemble des unités d'intérêt du carreau et de ne pas pouvoir identifier a posteriori une unité. Plusieurs méthodes de brouillage des données existent, dont certaines sont recommandées au niveau européen (*swapping, cell key method*) et celle utilisée dans le cadre de la diffusion des données carroyées de Filosofi[6] (grilles superposées). Dans le cadre du recensement de la population, l'Insee a décidé de se doter de règles pour garantir la confidentialité des données carroyées produites (cf. encadré 1).

Encadré 1

Règles de gestion de la confidentialité pour la diffusion des données du recensement au carreau

Afin de respecter le secret statistique et la confidentialité des données diffusées, l'Insee souhaite se doter de règles de diffusion des données carroyées issues du recensement de la population. Ces règles s'inspirent de la politique de diffusion des données carroyées issues de Filosofi établie en 2019[7] . Au stade actuel de la réflexion, ces règles pourraient être :
- Les nombres d'habitants, de logements et de ménages peuvent être diffusés au carreau sans restriction.
- Pour les autres variables à l'exclusion des variables à diffusion restreinte (nationalité, pays de naissance, pays de résidence antérieure, date d'arrivée en France) :
  * sans nécessité de retraitement, les données agrégées sur des carreaux d'au moins 11 ménages estimés peuvent être diffusées ;
  * avec retraitement préalable en appliquant des méthodes de contrôle du secret statistique, les données agrégées sur des carreaux de moins de 11 ménages estimés peuvent être diffusées.
- Les variables à diffusion restreinte ne peuvent pas être diffusées au niveau carreau.

La méthode de retraitement des données pour les carreaux de moins de 11 ménages estimés est en cours de détermination.

Le carroyage du recensement de la population se heurte à deux obstacles méthodologiques en passe d'être résolus

Par rapport à une source administrative exhaustive et géolocalisée, le recensement de la population fait face à deux défis en vue de son carroyage : la géolocalisation des logements dans les communes de moins de 10 000 habitants et les estimations d'indicateurs au carreau à partir de l'échantillon de données collectées dans les communes de 10 000 habitants ou plus. Chaque commune de moins de 10 000 habitants est recensée exhaustivement une année sur cinq. Cependant, les logements recensés ne font pas l'objet d'une géolocalisation au moment de la collecte. Il est nécessaire de procéder à une géolocalisation des logements a posteriori à partir des informations disponibles : adresse, district de collecte, caractéristiques individuelles des habitants. L'Insee a ainsi instruit une stratégie de géolocalisation qui consiste à tirer profit de deux méthodes en les combinant au mieux, pour déterminer les coordonnées géographiques aussi précisément que possible (cf. encadré 2). Dans les communes de métropole de moins de 10 000 habitants, on considère au final que les coordonnées géographiques estimées sont de bonne qualité pour plus de 93 % des personnes [Gallic et Pagès, 2022]. Dans les communes de plus de 10 000 habitants, ce problème de géolocalisation ne se pose pas du fait de la disponibilité d'un répertoire de bâtiments d'habitation déjà géolocalisé.
Le deuxième obstacle pour la production des données carroyées à partir du recensement de la population est la fiabilité des estimations dans les communes de plus de 10 000 habitants. En effet, sur ces territoires, le recensement de la population est effectué par sondage : sur un cycle de 5 ans, 40 % des logements sont recensés. Si ce taux de sondage est élevé pour une enquête statistique, certains carreaux contiennent très peu de logements échantillonnés, ce qui nuit à la qualité des estimations à cette échelle fine. Ainsi, 36 % des carreaux de 1 km de côté et localisés dans une commune de plus de 10 000 habitants contiennent moins de 10 logements échantillonnés en 5 ans. Des méthodes spécifiques d'estimation doivent dès lors être mises en place pour améliorer la précision des indicateurs calculés sur les carreaux. L'Insee a testé différentes méthodes ayant pour caractéristique commune la mobilisation d'information auxiliaire, en l'occurrence les données fiscales. La méthode retenue est une méthode d'imputation par *hot deck*, qui crée un simili-RP exhaustif, permettant ainsi de produire des estimations sur n'importe quel zonage infra-communal par simple sommation [Chevalier et alii, 2022]. Les propriétés de la méthode garantissent par ailleurs que ce simili-RP exhaustif est parfaitement cohérent avec les données déjà diffusées sur les autres mailles géographiques (IRIS, communes, etc.).

6. Filosofi est un rapprochement des données fiscales et des données sur les prestations sociales. Ces données fournissent de l'information sur le revenu déclaré des ménages fiscaux et permettent de reconstituer leur revenu disponible aux niveaux infra-communaux, communaux et supra-communaux.
7. Cf. https://www.insee.fr/fr/statistiques/6215647?sommaire=6215217

Encadré 2

Méthodes de géolocalisation des logements

Pour la géolocalisation des logements, on peut avoir trois cas de figure dans le recensement de la population.

Dans *les communes de 10 000 habitants ou plus*, l'Insee et les communes mettent à jour en continu un répertoire qui liste tous les bâtiments d'habitation (immeubles ou maisons) de la commune. Ce Répertoire des immeubles localisés (RIL) sert à tirer l'échantillon des logements recensés. Il contient notamment les coordonnées géographiques de chacun de ces bâtiments.

Dans *les communes de métropole de moins de 10 000 habitants*, on ne dispose pas d'un tel répertoire. On tente alors de rapprocher les données du recensement d'un référentiel géolocalisé constitué à partir de fichiers fiscaux (du cadastre notamment). Deux méthodes sont mises en œuvre :

– la première s'appuie sur les éléments d'adressage (numéro, voie, complément d'adresse, lieu-dit, commune). Lorsque ceux-ci sont assez précis et se retrouvent dans le référentiel, on peut associer à chaque bâtiment enquêté dans le recensement des coordonnées géographiques. Mais lorsque les éléments d'adressage ne sont pas assez précis, par exemple s'ils se limitent au nom d'un lieu-dit (comme le hameau des coquelicots), il n'est pas possible d'obtenir la coordonnée précise d'un bâtiment.

– on recourt alors à la deuxième méthode. Celle-ci repose sur les caractéristiques des individus résidant dans chaque bâtiment. En effet, les données du recensement comportent, comme les fichiers fiscaux, des informations non-nominatives sur les personnes (date de naissance, sexe, commune de naissance). On considère alors qu'un bâtiment enquêté par le recensement correspond à un bâtiment du référentiel dès lors que les deux bâtiments partagent un certain nombre d'individus avec des caractéristiques très proches (même sexe, date de naissance et lieu de naissance, au sein d'une même commune de résidence).

Ces deux méthodes sont combinées pour garder pour chaque logement les coordonnées géographiques les plus fiables. Pour les logements pour lesquels aucune des deux méthodes ne fournit des coordonnées géographiques de qualité, on détermine les coordonnées par interpolation linéaire entre deux adresses encadrantes géolocalisées.

Dans *les communes de moins de 10 000 habitants des départements d'Outre-mer*, les enquêteurs de l'Insee réalisent avant chaque recensement une enquête cartographique des territoires à recenser. Ils disposent pour ce faire d'une tablette équipée d'un GPS qui permet de géolocaliser directement les bâtiments d'habitation. Les bâtiments d'habitation ajoutés pendant l'enquête annuelle de recensement sont géolocalisés par interpolation ou à défaut sont positionnés aléatoirement sur la commune.

Une première diffusion de données carroyées à partir du recensement pour 2024

L'Insee est donc en train de se mettre en ordre de marche pour produire des données carroyées à partir du recensement, et enrichir l'offre de données déjà disponibles. La première diffusion est prévue au premier semestre 2024 de manière concomitante avec la mise à disposition des données à Eurostat. Au niveau national, la France prévoit de diffuser plus de variables que ce qui est demandé par Eurostat. D'ici 2024, sera également étudiée l'opportunité de diffuser sur des carreaux plus petits que ceux demandés par Eurostat et sur un périmètre géographique plus large (opportunité de diffuser sur certains DOM par exemple). Les investissements méthodologiques vont donc se poursuivre au sein de l'Insee pour y parvenir.

Une articulation avec les données carroyées diffusées à partir de Filosofi est à construire afin de consolider l'offre de l'Insee sur les territoires infra-communaux. ■■■

Cristina D'Alessandro et Gwennaël Solard

Pour aller plus loin

Chevalier M., **Gallic G.**, **Guillo C.**, **Guymarc G.**, **Pilorge C.**, « Méthodes de carroyage du recensement de la population dans les communes de 10 000 habitants et plus », Insee, article pour les journées de méthodologie statistique de 2022, mars 2022 : http://www.jms-insee.fr/2022/S12_2_ACTE_PILORGE_JMS%202022.pdf

Darriau V., « Les données carroyées, des outils et méthodes innovants », Insee, Courrier des statistiques n° 5, décembre 2020 : https://www.insee.fr/fr/statistiques/fichier/5008701/courstat-5-5.pdf

De Bellefond M.-P., **Pages J.**, « Combien de Français habitent à plus de 10 minutes en voiture d'une boulangerie », Insee, article de blog, octobre 2021 : https://blog.insee.fr/combien-de-francais-habitent-a-plus-de-10-minutes-en-voiture-dune-boulangerie-mieux-mesurer-les-temps-de-trajet-pour-mieux-comprendre-le-fonctionnement-des-territoires/

Gallic G., **Pages J.**, « La géolocalisation du recensement de la population dans les communes métropolitaines de moins de 10 000 habitants », Insee, article pour les journées de méthodologie statistique de 2022, mars 2022 : http://www.jms-insee.fr/2022/S12_1_ACTE_PAGES-GALLIC_JMS2022.pdf