Journées de méthodologie statistique de l'Insee 2022

LA GÉOLOCALISATION DU RECENSEMENT DE LA POPULATION DANS LES COMMUNES MÉTROPOLITAINES DE MOINS DE 10 000 HABITANTS

Gabrielle GALLIC (*), Jeanne PAGÈS (**)

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

gabrielle.gallic@insee.fr
jeanne.pages@insee.fr

Mots-clés : géolocalisation, appariements

Domaine concerné : Intégration de données – Appariements et fusion de sources

Résumé

Ce document présente les travaux menés pour géolocaliser l'ensemble des individus du recensement de la population (RP) sur le champ des communes de moins de 10 000 habitants ("petites" communes) de métropole. Ces unités ne sont en effet pas géolocalisées dans le cadre du processus actuel de production du RP, à la différence des adresses des communes de 10 000 habitants ou plus.

Deux méthodes coexistent à l'Insee pour géolocaliser le recensement de la population dans les communes de moins de 10 000 habitants. Elles consistent toutes deux à rapprocher les données du recensement d'un référentiel géolocalisé constitué à partir de fichiers fiscaux (Fidéli, Fichier démographique des logements et des individus). Ces deux méthodes diffèrent par la manière d'opérer ce rapprochement. La première méthode, mise en œuvre par l'application Geoloc, s'appuie sur les éléments d'adressage (numéro, type de voie, libellé de voie, complément d'adresse, commune). La deuxième méthode repose sur des caractéristiques non-nominatives des individus résidant dans chaque bâtiment (date de naissance, sexe, commune de naissance et commune de résidence).

Dans les deux cas, la qualité de la géolocalisation est qualifiée selon une note propre à chaque méthode. Afin de pouvoir comparer les deux méthodes, des améliorations ont été apportées aux deux systèmes de notation – l'objectif étant de mieux identifier pour chaque méthode les cas où la coordonnée obtenue est de très bonne qualité. La nouvelle notation permet donc de mieux cibler les coordonnées de bonne qualité de part et d'autre afin de choisir entre les deux méthodes de manière plus pertinente. Un ensemble de règles de choix a ensuite été défini afin de déterminer pour chaque adresse quelle coordonnée serait retenue, en se basant sur la qualité relative des coordonnées des deux méthodes ainsi que sur

des critères de disponibilité des données (la géolocalisation par Géoloc fournit des données définitives plus tôt que celle par appariement sur les caractéristiques individuelles). À l'issue de cet arbitrage, des coordonnées sont attribuées aux adresses résiduelles non géolocalisées par l'une ou l'autre des deux méthodes par un processus d'interpolation, basé sur les adresses géolocalisées du district de collecte correspondant.

Abstract

This paper presents the work carried out by Insee to geocode all individuals in Census data who live in municipalities with less than 10,000 inhabitants (the rest of the individuals are already geocoded though a buildings register covering larger municipalities).

Two geocoding methods coexist at INSEE for this purpose : they both use geographic coordinates from the land register as a reference. To match the Census data with the register, the GA method relies on street addresses (number, type of street, street name, municipality), and the GI method relies on non-nominal characteristics of individuals residing in each building (date of birth, sex, municipality of birth and municipality of residence).

In both cases, the quality of the geocoding is assessed through a score. The first step consisted in making the two methods comparable. We improved the two scoring systems in order to better identify when the proposed coordinates are of excellent quality for each method. We were then able to choose one of the two proposed sets of coordinates when they were initially both of good quality. We defined a set of decision rules, so as to select the best possible coordinates for each individual. When both methods fail to propose a coordinate of sufficient quality, the coordinates are interpolated using the coordinates of buildings nearby.

Introduction

Ce document présente les travaux menés par l'Insee pour géolocaliser l'ensemble des individus du recensement de la population (RP) sur le champ des communes de moins de 10 000 habitants ("petites" communes). Ces unités ne sont en effet pas géolocalisées dans le cadre du processus actuel de production du RP, à la différence des communes de 10 000 habitants et plus ou des communautés, dont le calcul de population mobilise un répertoire d'ores et déjà géolocalisé.

Or, dans la perspective d'études ayant une dimension spatiale (par exemple une analyse des trajets domicile-travail ou de la structure socio-démographique des habitants en quartier prioritaire de la politique de la ville), il est nécessaire de disposer de données géolocalisées sur tout le territoire. Cette nécessité s'exprime aussi désormais au travers de la réponse aux règlements européens sur le Census 2021, dans la mesure où les États-Membres doivent fournir des données sur des carreaux de 1km de côté. Des travaux ont donc été menés et ont abouti en 2021 à une nouvelle méthode de géolocalisation des lieux de résidence issus du recensement de la population dans les communes de moins de 10 000 habitants, dans l'optique d'inscrire cette opération de géolocalisation du recensement dans un rythme annuel.

Deux méthodes coexistent à l'Insee pour géolocaliser le RP dans les communes de moins de 10 000 habitants. Toutes deux utilisent comme référentiel les coordonnées géographiques issues du cadastre, par l'intermédiaire du fichier Fidéli.

La méthode de géolocalisation par appariement sur les éléments d'adressage (GA) se base sur l'adresse postale (numéro, voie, complément d'adresse) pour relier chaque adresse à une coordonnée. L'application Géoloc de l'Insee met en œuvre cette méthode de géolocalisation par l'adresse pour produire des statistiques infra-communales à partir de plusieurs sources utilisées au sein du système statistique public (demandeurs d'emploi en fin de mois, bénéficiaires des prestations légales versées par les CAF, bénéficiaires du régime général de l'assurance maladie, etc.)

La méthode de géolocalisation par appariement sur les caractéristiques des individus (GI) se base sur les caractéristiques des habitants pour relier chaque adresse à une coordonnée. Cette méthode est utilisée à l'Insee pour géolocaliser le recensement de la population dans l'objectif de calculer des indicateurs au niveau des Quartiers prioritaires de la politique de la ville (QPV).

Ces deux méthodes étaient jusqu'ici utilisées de manière indépendante l'une de l'autre. L'objectif de ces travaux est de combiner les deux processus afin d'obtenir, pour chaque adresse du recensement, la coordonnée de la meilleure qualité possible.

La première partie présente les deux méthodes de géolocalisation GA et GI, notamment la manière dont les coordonnées sont calculées et la façon dont ces coordonnées sont qualifiées à l'aide d'une note. Des aménagements ont été effectués par rapport aux méthodes originelles afin de les faire dialoguer au mieux : ils ont porté principalement sur l'amélioration de la pertinence de la note qualité.

La deuxième partie expose la démarche de géolocalisation retenue pour la géolocalisation des données du recensement : à l'aide des notes fournies par chaque méthode, on choisit la meilleure coordonnée proposée par l'une ou l'autre méthode, si celle-ci est de qualité satisfaisante. Dans le cas contraire, on définit une méthode d'imputation des coordonnées. Les principaux résultats ayant conduit à adopter cette démarche sont également présentés.

Les analyses de cette étude portent sur l'enquête annuelle de recensement (EAR) 2017, qui représente environ 2,4 millions d'adresses et 6,5 millions d'individus sur le champ des ménages ordinaires en petite commune.

Encadré : Éléments de contexte sur le recensement de la population

Le recensement de la population français est un recensement rotatif : il s'appuie sur une enquête à grande échelle conduite annuellement (appelée "Enquête annuelle de recensement" ou EAR) et portant sur une fraction seulement de la population. La population légale est calculée chaque année pour chaque commune à partir de l'agrégation de cinq EAR successives, d'informations provenant des répertoires des immeubles localisés et des communautés et de sources d'origine fiscale. La méthode de recensement diffère selon la taille de la commune. Les communes de moins de 10 000 habitants sont re-

censées exhaustivement tous les cinq ans, de façon tournante : des méthodes d'estimation spécifiques sont utilisées pour calculer la population légale les années où aucune collecte n'a eu lieu. Les communes de plus de 10 000 habitants sont recensées tous les ans, mais par sondage, sur un échantillon représentant environ 8 % des logements de la commune. Des estimations de populations sont calculées à partir de cet échantillon (représentant 40 % des logements de la commune sur un cycle complet de 5 EAR).

1. Présentation des méthodes de géolocalisation existantes

1.1 Géolocalisation par appariement sur les éléments d'adressage (GA)

Le processus de géolocalisation par appariement sur les éléments d'adressage (méthode GA) permet d'attribuer des coordonnées à des fichiers d'adresses postales en comparant celles-ci avec un référentiel d'adresses, construit notamment à partir de sources fiscales. Il comprend une première phase qui apparie de façon automatique les adresses du fichier source avec celles du référentiel. À l'issue de cette première phase, certaines adresses ne sont pas appariées car elles sont ambiguës ou absentes du référentiel. Il est alors possible de les géolocaliser manuellement lors d'une deuxième phase dite de "reprise" qui mobilise une équipe de gestionnaires. Cette méthode est utilisée dans de nombreux contextes au sein du système statistique public français : géolocalisation du répertoire des entreprises, de fichiers d'allocataires de l'assurance maladie, etc. À noter que cette méthode de géolocalisation est en général moins performante dans les petites communes que dans les grandes, en raison d'une moins bonne normalisation des adresses postales en zone rurale (par exemple : pas de numéro de voie, alors qu'il y a plusieurs habitations sur la même voie).

1.1.1. Référentiel utilisé

Le référentiel d'adresses utilisé par la méthode GA, pour la géolocalisation de l'EAR 2017 dans le cadre de cette étude, contient 25 millions d'adresses (dont 16 millions en "petites" communes) et leurs coordonnées associées. Il est construit à partir de Fidéli 2015, pour les adresses en petite commune : les coordonnées géographiques x, y sont celles de la parcelle cadastrale dans laquelle se trouve le bâtiment[1].

Une adresse du référentiel a ainsi généralement une unique coordonnée géographique x, y. Cependant, certaines adresses non normalisées peuvent avoir pour une même adresse plusieurs coordonnées géographiques x, y : par exemple un lieu-dit (ex : « Le Bourg ») où chaque x, y correspond à chacune des adresses fiscales sans pour autant pouvoir les différencier au niveau du libellé de l'adresse. À l'opposé, plusieurs adresses peuvent partager la même coordonnée x, y quand elles se situent sur une même parcelle cadastrale.

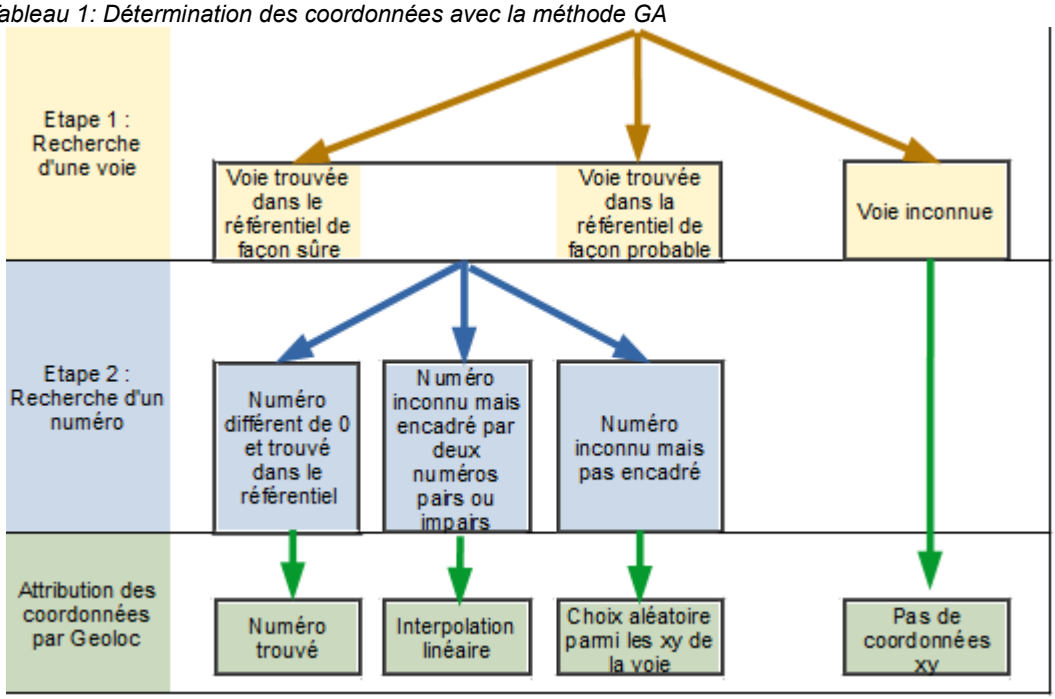
1.1.2. Méthode d'appariement

L'appariement avec la méthode GA se déroule en deux phases :

1 Le référentiel d'adresses de Géoloc repose principalement sur le RIL dans les communes de plus de 10 000 habitants et sur la géolocalisation des adresses de Fideli dans les communes de moins de 10 000 habitants. Il est complété ou corrigé à la marge par les travaux de gestionnaires de géolocalisation.

1. Appariement automatique : les adresses de l'EAR sont rapprochées de celles du référentiel sur la base du code commune, du libellé de voie (ou lieu-dit) et du numéro dans la voie. Dans un premier temps, la méthode GA recherche les adresses du fichier source dans le référentiel en commençant par le code commune, puis le libellé de voie, sans utiliser le numéro. La voie peut être soit trouvée et sûre, soit trouvée et probable, soit inconnue. Si la voie est sûre ou probable, l'adresse correspondante dans cette voie du référentiel est recherchée en utilisant cette fois le numéro renseigné dans le fichier source (Tableau 1).

Tableau 1: Détermination des coordonnées avec la méthode GA



2. Reprise manuelle : les observations qui ne sont pas codées après l'appariement automatique et la capitalisation peuvent être traitées manuellement par des gestionnaires. Étant donné la volumétrie importante qu'elle représenterait, aucune reprise manuelle n'a été effectuée pour la géolocalisation de l'EAR.

À l'issue de cette étape, 93,7 % des adresses ont été appariées avec le référentiel, c'est-à-dire que l'on a retrouvé au moins la voie de façon probable.

1.1.3. Géolocalisation des adresses non-appariées

Quand une adresse n'est pas appariée à l'issue de ce processus, des coordonnées lui sont imputées par tirage aléatoire au sein de la commune : la commune est divisée en carreaux de 100 m de côté, et un de ces carreaux est tiré au sort avec une probabilité qui dépend du nombre de logements qu'il contient (la

probabilité est nulle si le carreau ne contient aucun logement). Les coordonnées attribuées à l'adresse correspondent au centre du carreau de 100 m de côté ainsi tiré au sort.

Les coordonnées de la méthode GA obtenues par imputation ne sont utilisées pour la géolocalisation du recensement que de manière très résiduelle (cf paragraphe 2.1.1.).

1.1.4. Qualité de la géolocalisation

Une note qualité synthétise la manière dont les coordonnées d'une adresse ont été obtenues (Tableau 2).

Tableau 2: Répartition des notes qualité avec la méthode GA pour la géolocalisation de l'EAR 2017

Note qualité initiale méthode GA	Nombre d'adresses	Part d'adresses	Nombre d'individus	Part d'individus
11. Voie Sûre, Numéro trouvé	1 826 414	75,8%	5 019 575	76,9%
12. Voie Sûre, Position aléatoire dans la voie	183 931	7,6%	492 929	7,6%
21. Voie probable, Numéro trouvé	138 119	5,7%	358 762	5,5%
22. Voie probable, Position aléatoire dans la voie	108 215	4,5%	265 438	4,1%
33. Voie inconnue, Position aléatoire dans la commune	152 627	6,3%	389 278	6,0%
TOTAL	2 409 306	100,0%	6 525 982	100,0%

Source : EAR 2017, individus géolocalisés avec la méthode GA.
Champ : Individus des ménages résidant dans les petites communes de métropole enquêtées lors de l'EAR 2017 (groupe de rotation n°4).
Lecture : Parmi les 6 525 982 individus des ménages de petite commune de l'EAR 2017, 5 019 575 (76,9 %) sont géolocalisés avec une voie sûre et un numéro sûr.

À noter que la qualité de la géolocalisation avec la méthode GA dépend du territoire étudié. Typiquement, dans les territoires de faible densité où la part de communes sans nom de voirie est importante, la géolocalisation est moins performante. Au niveau départemental, la part d'individus dont les coordonnées sont positionnées aléatoirement dans la commune (note 33) oscille entre moins de 1 % (Val-de-Marne, Essone, Bas-Rhin) et 46 % (Corse-du-Sud).

1.2. Géolocalisation par appariement sur les caractéristiques des individus (GI)

Cette méthode a été développée par l'Insee pour géolocaliser les individus du recensement en petite commune afin de calculer des indicateurs au niveau des Quartiers prioritaires de la politique de la ville (QPV). Elle a également été utilisée pour géolocaliser des enquêtes tirées dans le RP, comme l'enquête Cadre de Vie et Sécurité (CVS).

1.2.1. Principe de la méthode

La géolocalisation par appariement sur les caractéristiques des individus (GI) consiste à utiliser les caractéristiques des *individus* qui résident à une certaine *adresse* (date de naissance, commune de naissance

et sexe). La notion d'*adresse* dans le cadre de cette méthode fait référence à celle d'un immeuble ou d'un bâtiment, identifié dans les EAR par un code immeuble, plutôt qu'une référence à une adresse postale (même si les deux notions se recoupent la plupart du temps). On rapproche ainsi les *adresses* (au sens *bâtiment*) de l'EAR de celles d'un référentiel géolocalisé ad-hoc constitué à partir des fichiers fiscaux (dans lesquels ces informations individuelles sont également disponibles) et de Fidéli (qui contient les coordonnées géographiques, issues du cadastre)[2]. Une adresse de l'EAR sera alors jugée proche d'une adresse donnée dans le référentiel si les deux adresses comportent des individus avec les mêmes caractéristiques. Si l'on retrouve dans les deux adresses des individus avec la même combinaison sexe, date de naissance, commune de naissance, il est très probable que ces adresses se correspondent : elles sont appariées.

En pratique, à chaque adresse de l'EAR est associé un ou plusieurs échos potentiels issus du référentiel, chacun assorti d'un score qualifiant l'appariement. L'écho retenu est celui ayant le score le plus élevé. Pour minimiser le risque d'erreur, on cherche à s'assurer que les immeubles de l'EAR situés dans un même district de collecte et donc relativement proches géographiquement sont aussi relativement proches géographiquement dans le référentiel.

La production du référentiel géolocalisé ad-hoc nécessaire à la mise en œuvre de cette méthode est longue. En particulier, il ne sera pas possible en régime courant d'utiliser le référentiel au 1er janvier N pour géolocaliser l'EAR N, mais seulement celui au 1er janvier N-1, ce qui dégrade légèrement la qualité de la géolocalisation (voir Annexe 2). Dans ce rapport, la méthode GI a été étudiée avec un référentiel de l'année N (Fideli 2017 pour géolocaliser l'EAR 2017), afin de comparer les deux méthodes sur des bases théoriques, indépendamment des questions de calendrier.

1.2.2. Géolocalisation des adresses non appariées

À la suite de l'étape précédente, la quasi-totalité des adresses (94,7%) de l'EAR est appariée avec une adresse dans le référentiel ad-hoc et est donc géolocalisée. Les coordonnées des autres adresses sont obtenues par imputation. La méthode GI ayant été développée spécifiquement dans le contexte des EAR, le traitement du reliquat d'adresses non-appariées s'appuie sur le district de collecte ou sur le rang d'adresse[3] :

- Interpolation entre adresses appariées : les coordonnées d'une adresse non-appariée sont obtenues en calculant le barycentre des coordonnées des adresses appariées qui l'encadrent, en termes de rang d'adresse. À noter que des critères de distance sont retenus ici : l'interpolation n'a lieu que si les adresses encadrantes sont distantes de moins de 400 m si elles sont dans la même rue, de moins de 200 m si elles sont dans des rues différentes.

- Barycentre du district de collecte : en cas d'échec de l'interpolation, l'adresse est placée au barycentre des adresses appariées de son district.

2 La méthode GI a été développée avant la création de Fidéli. A partir de Fidéli 2, le référentiel ad-hoc de la méthode GI sera uniquement issu de Fidéli 2.

3 En amont de la collecte du recensement, les agents recenseurs numérotent les adresses qu'ils vont enquêter. Cette numérotation suit l'ordre de la tournée de l'agent recenseur : dans la grande majorité des cas, les rangs d'adresse consécutifs correspondent à des adresses adjacentes.

Les coordonnées de la méthode GI obtenues par imputation (interpolation ou barycentre du district) ne sont pas utilisées (cf paragraphe 2.1.1.). En revanche, le principe d'interpolation entre adresses appariées a été repris et adapté (cf paragraphe 2.1.2.).

1.2.3. Qualité de la géolocalisation

Une note synthétise la manière dont les coordonnées d'une adresse ont été obtenues (Tableau 3).

Tableau 3: Répartition des notes qualité avec la méthode GI pour la géolocalisation de l'EAR 2017

Note qualité initiale méthode GI	Nombre d'adresses	Part d'adresses	Nombre d'individus	Part d'individus
1. Apparié avec un fichier fiscal dont les coordonnées sont de bonne qualité	2 278 012	94,6%	6 234 465	95,5%
2. Apparié avec un fichier fiscal dont les coordonnées sont de qualité moyenne	2 364	0,1%	7 875	0,1%
3. Apparié avec un fichier fiscal dont les coordonnées sont de mauvaise qualité	1 177	0,0%	3 708	0,1%
4. Non apparié (imputé)	110 834	4,6%	231 208	3,5%
5. Non apparié (pas de coordonnées)	16 919	0,7%	48 726	0,7%
TOTAL	2 409 306	100,0%	6 525 982	100,0%

Source : EAR 2017, individus géolocalisés avec la méthode GI (référentiel de l'année N).
Champ : Individus des ménages résidant dans les petites communes de métropole enquêtées lors de l'EAR 2017 (groupe de rotation n°4).
Lecture : Parmi les 6 525 982 individus des ménages de petite commune de l'EAR 2017,6 2 34 465 (95,5%) sont appariés avec un fichier fiscal dont les coordonnées sont de bonne qualité ;

Les notes 1 à 3 caractérisent des adresses appariées avec le fichier fiscal. La coordonnée du fichier fiscal est considérée comme de bonne qualité lorsque celle-ci provient de la parcelle cadastrale. Elle est considérée de qualité moyenne ou mauvaise quand elle a été imputée grâce aux coordonnées des autres adresses de la voie, de la parcelle cadastrale, voire de la commune.

Les notes 4 et 5 caractérisent des adresses non appariées avec le fichier fiscal. Leurs coordonnées sont interpolées lorsque cela est possible (note 4), à l'aide des adresses encadrantes ou au barycentre du district. Lorsqu'aucune autre coordonnée n'est appariée dans le district, la méthode GI ne fournit pas de coordonnée (note 5).

Cette notation a deux principaux défauts :

- les trois premières notes reflètent la qualité de la coordonnée utilisée, mais ne reflètent pas la qualité de l'appariement[4]. Si une adresse du recensement est appariée avec une adresse des fichiers fiscaux dont la coordonnée est de bonne qualité, mais que l'appariement s'effectue sur des critères imparfaits, la note qualité sera excellente.

4 Plus précisément, la qualité de la coordonnée correspond à la qualité d'un précédent appariement entre données fiscales et cadastre, fournie par Fidéli, mais elle ne reflète pas la qualité de l'appariement entre les données fiscales géolocalisées et l'EAR.

- de ce fait, les notes sont trop concentrées : beaucoup d'adresses (94,6 %) obtiennent une note de géolocalisation excellente, ce qui ne reflète pas la qualité réelle des coordonnées proposées.

1.3. Amélioration de la pertinence des notes qualités

L'objectif de cette étude est de combiner les deux méthodes de géolocalisation existantes : pour chaque adresse on souhaite retenir la meilleure coordonnée parmi celles fournies par les méthodes GA et GI.

Dans cette perspective, il est crucial d'avoir des notes qualité pertinentes, et qui permettent de comparer les deux méthodes à armes égales. Il est intéressant de limiter la meilleure note aux cas où la coordonnée est effectivement excellente.

L'objectif de ce travail n'est pas d'améliorer la géolocalisation fournie par chaque méthode, mais de mieux cibler quelles adresses sont bien géolocalisées avec chaque méthode. Les travaux menés ont ainsi permis d'améliorer la pertinence des notes qualité de la méthode GI, et dans une moindre mesure de la méthode GA.

1.3.1. Évaluation de l'amélioration des notes qualité

Le fait d'avoir deux méthodes de géolocalisation distinctes donne des outils pour évaluer la pertinence de leur notation. Les deux méthodes GI et GA font appel toutes deux à un référentiel alimenté par un même fichier géolocalisé (Fidéli, dont la géolocalisation est issue du cadastre) : les coordonnées présentes dans les référentiels sont donc les mêmes, mais les méthodes d'appariement utilisent des informations totalement différentes. Grâce à cela, on peut estimer la qualité de l'appariement de chaque méthode en comparant les coordonnées obtenues avec l'une et l'autre méthode. Si les coordonnées fournies par les deux méthodes pour une adresse donnée sont identiques, alors que l'on a utilisé pour l'appariement dans un cas l'adresse et dans l'autre les caractéristiques des habitants de l'adresse, il est très probable que l'appariement soit correct.

La méthode d'évaluation de la qualité repose donc sur la distance moyenne observée entre les deux coordonnées pour chaque croisement de notes.

Tableau 4: Distance entre les coordonnées obtenues avec chaque méthode en fonction des notes qualité initiales

Note qualité initiale méthode GA	Variable	Note qualité initiale méthode GI					
		1. Apparié avec un fichier fiscal dont les coordonnées sont de bonne qualité	2. Apparié avec un fichier fiscal dont les coordonnées sont de qualité moyenne	3. Apparié avec un fichier fiscal dont les coordonnées sont de mauvaise qualité	4. Non apparié (imputé)	5. Non apparié (pas de coordonnées)	TOTAL
11. Voie Sûre, Numéro trouvé	Nombre d'individus	*4 821 704*	*5 499*	*660*	*162 136*	*29 576*	*5 019 575*
	Distance moyenne en mètres entre coordonnées GI et GA	20	68	986	155		24
12. Voie Sûre, Position aléatoire dans la voie	Nombre d'individus	*462 563*	*1 018*	*987*	*22 753*	*5 608*	*492 929*
	Distance moyenne en mètres entre coordonnées GI et GA	242	232	663	584		256
21. Voie probable, Numéro trouvé	Nombre d'individus	*338 754*	*520*	*213*	*13 231*	*6 044*	*358 762*
	Distance moyenne en mètres entre coordonnées GI et GA	128	224	1 140	549		142
22. Voie probable, Position aléatoire dans la voie	Nombre d'individus	*248 916*	*119*	*423*	*12 040*	*3 940*	*265 438*
	Distance moyenne en mètres entre coordonnées GI et GA	302	372	739	820		322
33. Voie inconnue, Position aléatoire dans la commune	Nombre d'individus	*362 528*	*719*	*1 425*	*21 048*	*3 558*	*389 278*
	Distance moyenne en mètres entre coordonnées GI et GA	1 674	1 509	1 441	1 527		1 649
TOTAL	Nombre d'individus	*6 234 465*	*7 875*	*3 708*	*231 208*	*48 726*	*6 525 982*
	Distance moyenne en mètres entre coordonnées GI et GA	149	236	1 055	380		157

Source : EAR 2017, individus géolocalisés avec les méthodes GI (référentiel de l'année N) et GA ;
Champ : Individus des ménages résidant dans les petites communes de métropole enquêtées lors de l'EAR 2017 (groupe de rotation n°4) ;
Lecture : Parmi les 6 234 465 individus des ménages de petite commune de l'EAR 2017 qui ont une bonne note qualité initiale avec la méthode GI, 4 821 704 ont également une bonne note qualité initiale avec la méthode GA ; la distance moyenne entre les deux coordonnées est dans ce cas de 20 mètres.

Pour évaluer la pertinence de la notation de la méthode GA, on compare, pour les cas où la méthode GI obtient la meilleure note, les distances entre la coordonnée GA et la coordonnée GI, et ce pour chaque note de la méthode GA. On adopte une démarche symétrique pour évaluer la pertinence de la méthode GI.

Ainsi, parmi les adresses qui obtiennent la note maximale avec la méthode GI (colonne rouge), lorsque la coordonnée obtenue avec la méthode GA obtient une note de 11 (resp. 12/21/22/33), celle-ci est en moyenne à 20 mètres (resp. 242/128/302/1674 mètres) de la coordonnée obtenue avec la méthode GI.

On peut ainsi estimer que les adresses qui ont obtenu la note 11 avec la méthode GA sont sensiblement mieux localisées que celles qui ont obtenu les notes 12, 21 et 22, ces dernières étant sensiblement mieux géolocalisées que celles qui ont obtenu la note 33.

L'objectif des deux parties suivantes est de créer des groupes de notes plus pertinents et homogènes. En particulier, il s'agit d'identifier, au sein des "bonnes" notes de chaque méthode, d'éventuels sous-groupes dont la qualité serait moindre. On peut évaluer l'amélioration en vérifiant que la distance moyenne entre les coordonnées de bonne qualité diminue.

1.3.2. Amélioration de la note qualité de la méthode GA.

Avec la méthode GA, lorsqu'une adresse contient un libellé de voie mais pas de numéro de voie (ex : route de Castenau), si le libellé de voie est connu dans le référentiel, la note qualité maximale est attribuée (11. Voie sûre, numéro trouvé). Or en pratique la géolocalisation de ces adresses est de qualité moindre que lorsqu'il existe un numéro de voie (ex : 15, route de Castelnau).

Pour cette raison, la note 11 a été scindée en deux sous-notes : 11a (les adresses ayant un numéro dans le recensement) et 11b (les adresses n'ayant pas de numéro dans le recensement).

Tableau 5: Nouvelle note synthétique méthode GA

Note qualité détaillée méthode GA	Nombre d'individus	Part d'individus	Distance moyenne entre coordonnées GA et GI	Nouvelle note synthétique méthode GA
11a. Voie Sûre, Numéro trouvé (avec numéro)	4 761 026	76%	18	1. Bon
11b. Voie Sûre, Numéro trouvé (sans numéro)	60 678	1%	137	2. Moyen
21. Voie probable, Numéro trouvé	338 754	5%	128	
12. Voie Sûre, Position aléatoire dans la voie	462 563	7%	242	3. Mauvais
22. Voie probable, Position aléatoire dans la voie	248 916	4%	302	
33. Voie inconnue, Position aléatoire dans la commune	362 528	6%	1 674	4. Non apparié
TOTAL	6 234 465	100%	149	

Source *: EAR 2017, individus géolocalisés avec la méthode GA ;*
Champ : Individus des ménages *résidant dans les* petites communes *de métropole* enquêtées lors de l'EAR 2017 (groupe de rotation n°4) *qui ont la note "1. Apparié avec un fichier TH de bonne qualité" avec la méthode GI. ;*
Lecture : Parmi les 6 234 465 individus des ménages de petite commune de l'EAR 2017 qui ont *la note "1. Apparié avec un fichier TH de bonne qualité" avec la méthode GI*, *4 761 026 obtiennent la note "11a. Voie sûre, numéro sûr (avec numéro)" avec la méthode GA. Dans ce cas, la nouvelle note synthétique de la méthode GA est "1. Bon". La distance moyenne avec la coordonnée de la méthode GI lorsque cette dernière est bonne est de 18 mètres.*

Cela permet d'isoler 60 678 individus auparavant classés en note 11 et dont la coordonnée est manifestement de moins bonne qualité (puisqu'elle est à 137 mètres en moyenne de la coordonnée de la méthode GI lorsque cette dernière est de bonne qualité). Sur la base de cette comparaison on propose une nouvelle note synthétique pour la méthode GA (dernière colonne du tableau 5).

Avec la nouvelle note synthétique de la méthode GA, 76 % des individus obtiennent une coordonnée de bonne qualité au niveau national. Au niveau départemental, cette proportion varie entre 96 % dans le Bas-Rhin (67) et 3 % en Corse-du-Sud (2A).

Grâce à cette nouvelle note qualité synthétique, on cible mieux les meilleures performances de la méthode GA : les adresses qui obtiennent la meilleure note avec les deux méthodes sont distantes de seulement 18 mètres en moyenne (Tableau 6).

Tableau 6: Distance entre les coordonnées obtenues avec chaque méthode en fonction de la note qualité initiale de la méthode GI et de la nouvelle note synthétique de la méthode GA

Nouvelle note synthétique méthode GA	Variable	Note qualité initiale méthode GI					
		1. Apparié avec un fichier fiscal dont les coordonnées sont de bonne qualité	2. Apparié avec un fichier fiscal dont les coordonnées sont de qualité moyenne	3. Apparié avec un fichier fiscal dont les coordonnées sont de mauvaise qualité	4. Non apparié (imputé)	5. Non apparié (pas de coordonnées)	TOTAL
1. Bon	Nombre d'individus	*4 761 026*	*5 482*	*557*	*159 478*	*28 160*	*4 954 703*
	Distance moyenne en mètres entre coordonnées GI et GA	18	68	994	144		22
2. Moyen	Nombre d'individus	*399 432*	*537*	*316*	*15 889*	*7 460*	*423 634*
	Distance moyenne en mètres entre coordonnées GI et GA	129	224	1 074	601		146
3. Mauvais	Nombre d'individus	*711 479*	*1 137*	*1 410*	*34 793*	*9 548*	*758 367*
	Distance moyenne en mètres entre coordonnées GI et GA	263	247	685	666		279
4. Non apparié	Nombre d'individus	*362 528*	*719*	*1 425*	*21 048*	*3 558*	*389 278*
	Distance moyenne en mètres entre coordonnées GI et GA	1 674	1 509	1 441	1 527		1 649
TOTAL	Nombre d'individus	*6 234 465*	*7 875*	*3 708*	*231 208*	*48 726*	*6 525 982*
	Distance moyenne en mètres entre coordonnées GI et GA	149	236	1 055	380		157

Source : EAR 2017, *individus géolocalisés avec les méthodes GI (référentiel de l'année N) et GA ;*
Champ : Individus des ménages *résidant dans les* petites communes *de métropole* enquêtées lors de l'EAR 2017 (groupe de rotation n°4) .
Lecture : Parmi les 6 234 465 individus des ménages de petite commune de l'EAR 2017 qui ont une bonne note qualité avec la méthode GI, 4 *761 026* ont également une *nouvelle note synthétique pour la méthode GA égale à "1. Bon"* ; la distance moyenne entre les deux coordonnées est dans ce cas de *18* mètres.

1.3.3. Amélioration de la note qualité de la méthode *GI*.

Création d'un score de qualité de l'appariement

Comme évoqué plus haut, la note qualité initiale de la coordonnée obtenue par méthode GI n'utilise pas d'information sur la qualité de l'appariement, mais uniquement sur la qualité des coordonnées fournies par Fidéli.

Pourtant, l'appariement entre l'adresse du recensement et l'adresse du référentiel ad-hoc se fait parfois avec une grande certitude (par exemple quand sur deux individus dans l'adresse du recensement, on en retrouve deux dans le référentiel avec exactement les mêmes caractéristiques : même date de naissance, même commune de naissance, même sexe), et parfois avec moins de précision (par exemple quand sur deux individus dans l'adresse du recensement, on en retrouve un seul dans le référentiel, avec des caractéristiques proches mais pas totalement identiques de celles du recensement).

Nous avons ainsi créé une note qualité de l'appariement qui contient quatre modalités :

- Si l'adresse a été obtenue en appariant uniquement avec un ou plusieurs individu(s) ambigu(s)[5], la qualité de l'appariement est « 3. Ambigus ».

- Si l'adresse a été obtenue en appariant moins de la moitié des individus de l'adresse, la qualité de l'appariement est « 2. Mauvaise ».

- Si l'adresse a été obtenue en appariant au moins la moitié des individus de l'adresse :

* Soit il y a deux personnes à l'adresse et on en apparie une sur deux, auquel cas la la qualité de l'appariement est « 2. Mauvaise »

* Soit il y a deux personnes à l'adresse et on apparie les deux, auquel cas la qualité de l'appariement est « 1. Bonne »

* Soit il y a plus de deux personnes à l'adresse, auquel cas la qualité de l'appariement est « 1. Bonne ».

Création d'un score global de qualité de la coordonnée GI

À partir de ce score de qualité de l'appariement, on peut construire un score "global" qui décrit la qualité de la coordonnée obtenue par la méthode GI. Ce score tient compte du fait qu'on ait apparié ou non avec le référentiel, de la qualité des coordonnées utilisées issues de Fidéli et de la qualité de l'appariement.

Nouvelle note synthétique de la méthode GI

- Si l'immeuble de l'EAR a été apparié avec un immeuble du référentiel ayant une coordonnée de bonne qualité :

* Soit la qualité de l'appariement est bonne, auquel cas le score global est "1. Bon"

* Soit la qualité de l'appariement est mauvaise, auquel cas le score global est "2. Moyen"

* Soit l'appariement ne repose que sur des individus ambigus, auquel cas le score global est "3. Mauvais"

- Si l'immeuble de l'EAR a été apparié avec un immeuble du référentiel ayant une coordonnée de qualité moyenne ou mauvaise, le score global est "3. Mauvais"

- Si l'individu n'a pas pu être apparié et a été interpolé, le score global est "4. Non apparié (imputé)"

- Si l'individu n'a pas pu être imputé, le score global est "5. Non apparié (pas de coordonnées)"

5 Les individus "ambigus" sont, de manière simplifiée, des individus qui ont un « jumeau » dans leur commune de résidence : l'EAR collecte des informations sur deux individus qui ont exactement les mêmes caractéristiques (sexe, date de naissance, commune de naissance, commune de résidence).

Tableau 7: Nouvelle note synthétique méthode GI

Note qualité initiale méthode GI	Score de qualité de l'appariement	Nombre d'individus	Part d'individus	Distance moyenne en mètres entre les coordonnées GA et GI	Nouvelle note synthétique méthode GI
1. Apparié avec un fichier fiscal dont les coordonnées sont de bonne qualité	1. Bon	3 597 700	73%	12	1. Bon
	2. Mauvais	1 044 570	21%	24	2. Moyen
	3. Ambigus	118 756	2%	149	3. Mauvais
2. Apparié avec un fichier fiscal dont les coordonnées sont de qualité moyenne	Tous	5 482	0%	68	
3. Apparié avec un fichier fiscal dont les coordonnées sont de mauvaise qualité	Tous	557	0%	994	
4. Non apparié (imputé)	Tous	159 478	3%	144	4. Non apparié (imputé)
5. Non apparié (pas de coordonnées)	Tous	28 160	1%		5. Non apparié (pas de coordonnées)
TOTAL	TOTAL	4 954 703	100%	22	TOTAL

Source : EAR 2017, individus géolocalisés avec la méthode GI (référentiel de l'année N).
Champ : Individus des ménages *résidant dans les* petites communes *de métropole* enquêtées lors de l'EAR 2017 (groupe de rotation n°4) qui ont *la nouvelle note synthétique "1. Bon" avec la méthode GA.*
Lecture : Parmi les *4 954 703* individus des ménages de petite commune de l'EAR 2017 qui ont la *nouvelle note synthétique "1. Bon" avec la méthode GA*, *3 597 700* obtiennent la note qualité initiale "1. Apparié avec un fichier *fiscal* dont *les coordonnées sont* de bonne qualité" avec la méthode GI et ont un score de qualité de l'appariement avec la méthode GI "1. Bon". Dans ces cas, la nouvelle note synthétique de la méthode *GI* est "1. Bon" ; la distance moyenne avec la méthode GA est de 12 mètres.

Avec la nouvelle note synthétique de la méthode GI, 73 % des individus qui ont une nouvelle note « 1. Bon » avec la méthode GA obtiennent une coordonnée de bonne qualité au niveau national. Au niveau départemental, cette proportion varie entre 80 % dans le Bas-Rhin (67) et 49 % en Corse-du-Sud (2A).

Grâce à la nouvelle note qualité de la méthode GI, on cible mieux les meilleures performances de la méthode GI : les adresses qui obtiennent la meilleure note avec les deux méthodes sont distantes de seulement 12 mètres en moyenne.

Tableau 8: Distance entre les coordonnées obtenues avec chaque méthode en fonction de la note qualité initiale de la méthode GI et de la nouvelle note synthétique de la méthode GA

Nouvelle note synthétique méthode GA	Variable	Nouvelle note synthétique méthode GI					
		1. Bon	2. Moyen	3. Mauvais	4. Non apparié (imputé)	5. Non apparié (pas de coordonnées)	TOTAL
1. Bon	Nombre d'individus	*3 597 700*	*1 044 570*	*124 795*	*159 478*	*28 160*	*4 954 703*
	Distance moyenne en mètres entre coordonnées GI et GA	12	24	149	144		22
2. Moyen	Nombre d'individus	*299 710*	*89 729*	*10 846*	*15 889*	*7 460*	*423 634*
	Distance moyenne en mètres entre coordonnées GI et GA	116	141	432	601		146
3. Mauvais	Nombre d'individus	*532 207*	*159 960*	*21 859*	*34 793*	*9 548*	*758 367*
	Distance moyenne en mètres entre coordonnées GI et GA	248	274	558	666		279
4. Non apparié	Nombre d'individus	*267 299*	*85 693*	*11 680*	*21 048*	*3 558*	*389 278*
	Distance moyenne en mètres entre coordonnées GI et GA	1 676	1 658	1 693	1 527		1 649
TOTAL	Nombre d'individus	*4 696 916*	*1 379 952*	*169 180*	*231 208*	*48 726*	*6 525 982*
	Distance moyenne en mètres entre coordonnées GI et GA	140	162	326	380		157

Source : EAR 2017, *individus géolocalisés avec les méthodes GI (référentiel de l'année N) et GA*
Champ : Individus des ménages *résidant dans les* petites communes *de métropole* enquêtées lors de l'EAR 2017 (groupe de rotation n°4)
Lecture : Parmi les *4 696 916* individus des ménages de petite commune de l'EAR 2017 qui ont *une nouvelle note synthétique égale à "1. Bon" pour la méthode GI*, *3 597 700* ont également *une nouvelle note synthétique égale à "1. Bon" pour la méthode GA* ; la distance moyenne entre les deux coordonnées est dans ce cas de *12* mètres.

La création de nouvelles notes qualité synthétiques pour chaque méthode de géolocalisation a permis de mieux identifier les adresses qui sont bien géolocalisées avec chacune des méthodes. Elle a permis de créer des catégories plus homogènes que les scores initiaux.

Ces nouvelles notes permettent de choisir de manière plus pertinente pour chaque adresse entre la coordonnée proposée par la méthode GA ou la coordonnée proposée par la méthode GI.

2. Coordonnées utilisées pour la géolocalisation du recensement

Dans cette partie on présente de quelle manière les coordonnées de chaque individu du recensement seront obtenues. Pour chaque individu, on a le choix entre :

- retenir la coordonnée proposée par la méthode GA (idéalement si celle-ci est de bonne qualité)
- retenir la coordonnée proposée par la méthode GI (idéalement si celle-ci est de bonne qualité)
- calculer une coordonnée à partir d'éléments extérieurs, comme les coordonnées des adresses proches (si l'on pense que les méthodes GA et GI fournissent des coordonnées de mauvaise qualité).

2.1. Choix des coordonnées pour la géolocalisation du recensement

2.1.1. Règle de décision pour le choix des coordonnées

En fonction des notes obtenues par les coordonnées de chaque méthode, on définit une règle pour savoir quelles coordonnées choisir.

Règle de décision pour le choix des coordonnées

- lorsque la coordonnée proposée par la méthode GA est de bonne qualité, on prend la coordonnée de la méthode GA (en vert foncé dans le tableau 9 : ligne 1)
- lorsque la coordonnée proposée par la méthode GA n'est pas de bonne qualité :
  * si la coordonnée de la méthode GI est de qualité bonne ou moyenne, on prend la coordonnée de la méthode GI (en violet foncé et violet clair : colonnes 1 et 2, à l'exclusion de la ligne 1)
  * sinon, si la coordonnée de la méthode GA est de qualité moyenne, on prend la coordonnée de la méthode GA (en vert clair : ligne 2 dans le tableau, à l'exclusion des colonnes 1 et 2)
  * sinon (lorsque les coordonnées la méthode GA et de la méthode GI sont mauvaises), on interpole (en blanc : colonnes 3, 4 et 5, à l'exclusion des lignes 1 et 2), cf paragraphe 2.1.2.

NB : Dans les cas où il n'y a aucun individu apparié dans le district, on reprend la géolocalisation aléatoire dans la commune proposée par la méthode GA (note "4. Non apparié").

Tableau 9: Choix des coordonnées (en nombre d'individus)

Nouvelle note synthétique méthode GA	Variable	Nouvelle note synthétique méthode GI					
		1. Bon	2. Moyen	3. Mauvais	4. Non apparie (interpolé)	5. Non apparie (pas de coordonnées)	TOTAL
1. Bon	Nombre d'individus	3 597 700	1 044 570	124 795	159 478	28 160	4 954 703
	Part d'individus	*55,1%*	*16,0%*	*1,9%*	*2,4%*	*0,4%*	*75,9%*
2. Moyen	Nombre d'individus	299 710	89 729	10 846	15 889	7 460	423 634
	Part d'individus	*4,6%*	*1,4%*	*0,2%*	*0,2%*	*0,1%*	*6,5%*
3. Mauvais	Nombre d'individus	532 207	159 960	21 859	34 793	9 548	758 367
	Part d'individus	*8,2%*	*2,5%*	*0,3%*	*0,5%*	*0,1%*	*11,6%*
4. Non apparié	Nombre d'individus	267 299	85 693	11 680	21 048	3 558	389 278
	Part d'individus	*4,1%*	*1,3%*	*0,2%*	*0,3%*	*0,1%*	*6,0%*
TOTAL	Nombre d'individus	4 696 916	1 379 952	169 180	231 208	48 726	6 525 982
	Part d'individus	72,0%	21,1%	2,6%	3,5%	0,7%	*100,0%*

Choix retenu	Nombre d'individus	Part d'individus
On choisit la méthode GA	**4 988 898**	**76,4%**
dont : avec de bonnes coordonnées	4 954 703	75,9%
dont : avec des coordonnées moyennes	34 195	0,5%
On choisit la méthode GI	**1 434 598**	**22,0%**
dont : avec de bonnes coordonnées	*1 099 216*	16,8%
dont : avec des coordonnées moyennes	*335 382*	5,1%
On interpole ou on impute	**102 486**	**1,6%**
TOTAL	6 525 982	100,0%

Source : EAR 2017, individus géolocalisés avec les méthodes GA et GI (référentiel de l'année N).
Champ : Individus des ménages résidant dans les petites communes de métropole enquêtées lors de l'EAR 2017 (groupe de rotation n°4).
Lecture : Parmi les 6 525 982 individus de l'EAR 2017, 3 597 700 (55,1 %) ont une bonne note globale avec la méthode GA et la méthode GI. Dans ce cas, on privilégie la coordonnée fournie par la méthode GA.

La règle de décision choisie conduit donc à retenir la méthode GA pour 4 988 898 individus (76,4 %), la méthode GI pour 1 434 598 individus (22,0 %), et à interpoler ou imputer les coordonnées pour 102 486 individus (1,6 %).

Un tableau détaillant le choix des coordonnées en nombre d'adresses est présenté en annexe 1.

2.1.2. Critères de choix

La méthode exposée ci-dessus repose sur plusieurs principes, adoptés après des études exploratoires :

- On privilégie GA par rapport à GI lorsque les deux méthodes annoncent une coordonnée de bonne qualité

Dans les cas où les deux méthodes fournissent une coordonnée étiquetée comme de bonne qualité, on privilégie la coordonnée de la méthode GA.

En réalité, ce choix n'est déterminant que pour une petite partie de ces adresses : de manière générale, les coordonnées fournies par les deux méthodes sont souvent proches, mais c'est encore plus le cas lorsque les coordonnées sont de bonne qualité : dans 85,2 % des cas où les deux méthodes fournissent une coordonnée de bonne qualité, les coordonnées proposées par les deux méthodes sont distantes de moins de 10 mètres. Dans 98,2 % des cas les coordonnées sont distantes de moins de 100 mètres.

Le choix de privilégier la méthode GA repose sur plusieurs constats :

- Les évaluations de la méthode GI présentées dans ce rapport ont été effectuées avec un référentiel du même millésime que l'enquête annuelle de recensement (2017). Or, dans l'optique de diffusion des populations carroyées, une EAR N doit être géolocalisée au plus tard en fin d'année N ; mais le référentiel millésimé N utilisé par la méthode GI n'est pas encore disponible à cette période. Pour la production courante, la méthode GI utilisera donc un référentiel de l'année N-1. Il a été estimé que cette année de retard conduisait à une mauvaise géolocalisation d'environ 4 % des individus, à cause des déménagements intervenus entre l'année N-1 et l'année N (voir Annexe 2). Pour la diffusion des populations carroyées, la méthode GI ne proposera donc que cette coordonnée provisoire, de moindre qualité. La méthode GA permet *a contrario* de disposer d'une géolocalisation définitive de l'année N dès la fin N.

- La méthode GA est un service utilisé par de nombreux processus à l'Insee et dans le système statistique public. Le référentiel bénéficie des corrections manuelles opérées pour les autres processus. À l'inverse, la méthode GI est utilisée uniquement pour le recensement ; elle repose sur un processus moins formalisé et ne permet pas de capitaliser sur les opérations passées.

<u>- On privilégie les coordonnées « moyennes » de GI par rapport aux coordonnées « moyennes » de GA</u>

Dans les cas où les coordonnées fournies par les deux méthodes sont moyennes, on privilégie la coordonnée de la méthode GI. Ce choix repose sur la comparaison des distances présentée au tableau 8. Dans le cas où les coordonnées de la méthode GA sont de bonne qualité et celles de la méthode GI de qualité moyenne, la distance moyenne entre les deux est de 24 mètres. Dans le cas où les coordonnées de la méthode GI sont de bonne qualité et celles de la méthode GA de qualité moyenne, la distance moyenne entre les deux est de 116 mètres. Sachant que par ailleurs les coordonnées de bonne qualité semblent relativement précises dans les deux cas (puisque distantes de seulement 12 mètres en moyenne), cela semble indiquer que les coordonnées "moyennes" de la méthode GI sont de meilleure qualité que celles de la méthode GA.

<u>- Périmètre des adresses à interpoler</u>

Lorsque la méthode GA et la méthode GI fournissent des coordonnées de "mauvaise" qualité, on ne retient pas ces coordonnées. On préfère interpoler les coordonnées à l'aide des adresses proches (même principe que l'interpolation présentée au 1.2.2) plutôt que d'utiliser les coordonnées obtenues par les

appariements : 1,6 % des individus sont ainsi interpolés. Ce choix repose sur l'idée que les coordonnées obtenues par interpolation seraient de meilleure qualité que celles obtenues avec un appariement de mauvaise qualité.

Deux scenarii alternatifs ont été étudiés.

- Un premier scenario alternatif consistait à ne retenir que les coordonnées "bonnes" de la méthode GA ainsi que les coordonnées "bonnes" et "moyennes" de la méthode GI. Par rapport au scenario retenu, les notes "moyennes" de la méthode GA n'étaient pas utilisées. Or dans 44 districts, les seules coordonnées proposées sont des coordonnées de la méthode GA de qualité "moyenne" ou "mauvaise". Pour ces districts, on ne disposait donc d'aucune coordonnée. Pour ces 44 districts (8 365 individus) il fallait donc recourir à la géolocalisation aléatoire dans la commune.

- Un second scenario alternatif consistait à retenir toutes les coordonnées issues d'un appariement de l'une ou l'autre méthode. Par rapport au scenario retenu, les notes "mauvaises" de la méthode GA et de la méthode GI étaient utilisées. Ce scenario avait l'avantage de limiter le nombre de districts sans coordonnée à 3 (153 individus). Mais le fait d'utiliser les coordonnées "mauvaises" à la place de l'interpolation dégraderait probablement la qualité de la géolocalisation[6].

Avec la méthode retenue, quatre districts contiennent uniquement des individus sans coordonnée avec l'une ou l'autre méthode : dans ces quatre districts, tous les individus (au nombre de 254) sont géolocalisés aléatoirement dans la commune[7].

2.1.3. Origine et qualité des coordonnées au niveau local

La part des coordonnées obtenues avec chaque méthode varie de façon importante selon les territoires (Tableau 10).

6 Les adresses géolocalisées par la méthode GA avec une qualité moyenne ont une précision de l'ordre de 100 m en moyenne et de 200m pour celles géolocalisées avec une qualité mauvaise. Cette précision de 200 m semble inférieure à ce qu'on peut espérer de l'interpolation car, avec l'interpolation réalisée par la méthode GI, on obtient une précision de 143 m en moyenne (calculée par rapport à la coordonnée trouvée par GA, lorsque GA est bon). Mais comme le champ des adresses interpolées *in fine* ne correspond pas à celui qui permet d'estimer une précision de 143 m, cette mesure du gain de précision avec l'interpolation plutôt que des coordonnées « mauvaise » fournies par GA serait à consolider.

7 Comme il n'y a qu'un seul district dans trois de ces quatre communes, il s'agit en réalité d'une géolocalisation aléatoire dans le bon district. Dans la quatrième commune, il n'y a qu'une seule adresse dans le district concerné.

Tableau 10: Origine des coordonnées par anciennes régions

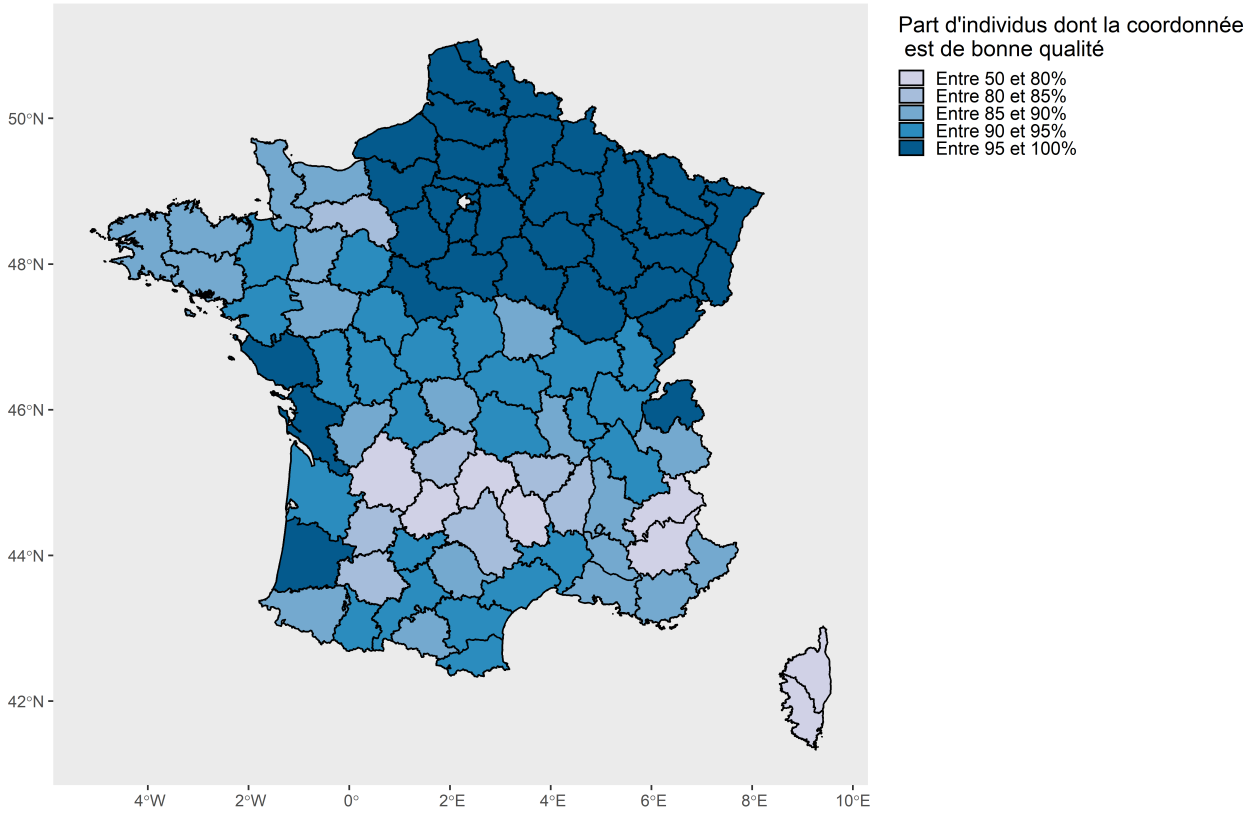
NUTS2 (Anciennes régions)	Nombre d'individus	Part des individus ayant des coordonnées de qualité / origine :				
		GA - Bons	GI - Bons	GI - Moyens	GA - Moyens	Interpolés
Alsace	215 390	95,4%	3,5%	0,8%	0,1%	0,2%
Nord-Pas-de-Calais	415 099	93,2%	4,8%	1,5%	0,2%	0,3%
Lorraine	324 281	91,7%	5,9%	1,6%	0,5%	0,3%
Picardie	275 714	91,5%	5,7%	2,1%	0,3%	0,4%
Ile-de-France	402 281	91,4%	5,7%	2,0%	0,3%	0,6%
Haute-Normandie	219 890	88,4%	8,2%	2,6%	0,3%	0,5%
Champagne-Ardennes	157 306	88,4%	8,4%	2,2%	0,4%	0,5%
France-Comté	168 888	84,5%	10,9%	3,3%	0,4%	0,8%
Centre	309 503	81,0%	13,7%	3,7%	0,6%	1,0%
Poitou-Charante	263 175	78,1%	15,9%	4,1%	0,5%	1,3%
Loire Atlantique	422 657	75,7%	16,9%	5,2%	1,1%	1,1%
Languedoc-Roussillon	317 164	75,6%	15,7%	5,8%	0,6%	2,3%
Bourgogne	228 742	74,1%	19,2%	5,0%	0,4%	1,4%
Rhône Alpes	750 522	72,3%	19,2%	6,5%	0,4%	1,6%
Aquitaine	414 761	65,1%	24,8%	6,8%	0,5%	2,7%
Basse-Normandie	216 891	64,6%	23,9%	6,5%	1,5%	3,5%
Provence Alpes Côte d'Azur	308 420	62,0%	23,8%	9,9%	0,7%	3,7%
Bretagne	433 028	61,9%	27,6%	8,0%	0,6%	1,9%
Midi-Pyrénées	358 839	58,5%	30,3%	8,3%	0,4%	2,6%
Auvergne	195 284	56,1%	32,8%	8,1%	0,3%	2,6%
Limousin	88 284	51,5%	36,6%	8,4%	0,8%	2,7%
Corse	39 863	16,5%	44,0%	25,5%	3,2%	10,6%
TOTAL	6 525 982	75,9%	16,8%	5,1%	0,5%	1,6%

Source : EAR 2017, individus géolocalisés avec les méthodes GA et GI (référentiel de l'année N).
Champ : Individus des ménages résidant dans les petites communes de métropole enquêtées lors de l'EAR 2017 (groupe de rotation n°4).
Lecture : Parmi les 402 281 individus d'Île-de-France, dans 91,4 % des cas on retient une une coordonnée obtenue grâce à la méthode GA de bonne qualité.

Au niveau national, 75,9 % des individus ont des coordonnées de bonne qualité obtenues grâce à la méthode GA. Le taux d'individus avec des coordonnées de bonne qualité monte à 92,7 % en utilisant la méthode GI en complément de la méthode GA. En Corse, seuls 16,5 % des individus ont des coordonnées de bonne qualité obtenues grâce à la méthode GA. En combinant GA et GI, cette proportion monte à 60,5 %.

Le fait d'utiliser de manière conjointe les deux méthodes GA et GI permet donc d'améliorer sensiblement les résultats au niveau national par rapport à la situation où on n'utiliserait qu'une seule méthode, mais au niveau local cet apport est encore plus crucial pour certains territoires, où les adresses sont non normalisées et la méthode GA moins performante.

Figure 1: Part d'individus dont la coordonnée est de bonne qualité au niveau département



Source : EAR 2017, individus géolocalisés avec les méthodes GA et GI (référentiel de l'année 2017).
Champ : Individus des ménages résidant dans les petites communes de métropole enquêtées lors de l'EAR 2017 (groupe de rotation n°4).
Note : les départements de Paris (75) et des Hauts-de-Seine (92) apparaissent en blanc sur la carte car il n'y a pas de petite commune recensée lors de l'EAR 2017 pour ces départements. À noter que la proportion de petites communes varie sensiblement d'un département à l'autre.
Lecture : Dans le Finistère, la part d'individus de petite commune ayant une coordonnée de bonne qualité (obtenue avec la méthode GA ou GI) se situe entre 85 % et 90 %

La part de coordonnées de bonne qualité varie grandement entre les départements (Figure 1). Elle est proche de 100 % dans les départements du nord-est de la France. Elle est de 50 % en Corse-du-Sud.

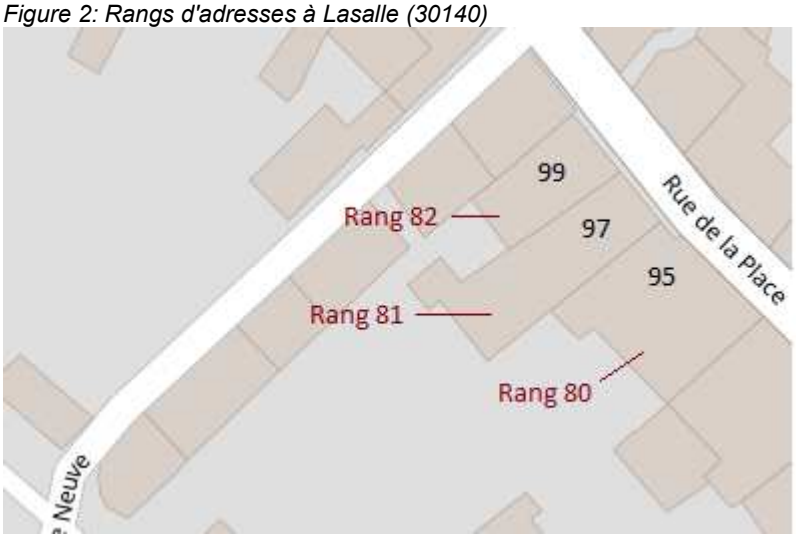
Dans les départements où la part de communes sans nom de voirie est élevée, la méthode GA n'obtient pas de bons résultats. L'adjonction de la méthode GI permet de compléter les résultats dans ces territoires, mais elle ne permet pas d'atteindre totalement le même niveau de qualité que dans les territoires ayant des éléments d'adressage bien identifiables.

2.2. Interpolation

1.2.1. Principe de l'interpolation

En amont de la collecte du recensement, les agents recenseurs numérotent les adresses qu'ils vont visiter. Cette numérotation suit l'ordre de la tournée de l'agent recenseur : dans la grande majorité des cas, les rangs d'adresse consécutifs correspondent à des adresses adjacentes.

Figure 2: Rangs d'adresses à Lasalle (30140)



Source : Open Street Map, version du 17/11/2021.
Champ : adresses de l'EAR 2017 (zoom).
Lecture : L'adresse "95, rue de la place, LASALLE" correspond au rang d'adresse 80 du district. L'adresse "97, rue de la place, LASALLE" correspond au rang d'adresse 81 du district.

Dans le cas où on ne dispose pas de coordonnées de bonne qualité, on peut donc se baser sur les coordonnées des rangs d'adresse précédant et suivant l'adresse à géolocaliser. Sur l'image ci-dessus (Figure 2), on peut estimer la coordonnée du "97, rue de la Place" (rang d'adresse 81) en calculant le barycentre des coordonnées des rangs d'adresse 80 et 82.

2.2.2. Règle d'interpolation

Lorsqu'aucune des deux méthodes ne fournit de coordonnée de qualité satisfaisante, on calcule le barycentre entre les coordonnées des deux rangs d'adresse encadrant l'adresse que l'on souhaite géolocaliser.

Si le rang d'adresse précédent (resp. suivant) n'a pas de coordonnée de qualité satisfaisante, on utilise l'avant-dernier (resp. deuxième suivant) rang d'adresse, et ainsi de suite, au sein du district.

Si aucun des rangs d'adresse précédents (resp. suivants) n'a de coordonnée de qualité satisfaisante, on reprend la coordonnée du rang d'adresse suivant (resp. précédent).

Si aucune coordonnée n'est connue dans le district, on retient la coordonnée aléatoire dans la commune proposée par la méthode GA.

2.2.3. Qualité des coordonnées utilisées

L'interpolation intervient dans des cas où les deux méthodes échouent à fournir une coordonnée de bonne ou de moyenne qualité ; typiquement dans des zones peu denses où les communes n'ont pas adopté de nom de voirie et où le référentiel issu des sources fiscales ne permet pas de retrouver de manière unique les individus recensés. Cela concerne parfois des zones entières et pas uniquement une adresse entourée d'autres adresses bien géolocalisées.

Pour cette raison, l'interpolation se fait parfois à partir de coordonnées correspondant à des rangs d'adresse éloignés. Cela reste malgré tout préférable à une géolocalisation aléatoire dans la commune.

Au final, dans 63 % des cas, les coordonnées utilisées pour interpoler sont à moins de 500 mètres l'une de l'autre (Tableau 11).

Tableau 11: Distance entre les deux coordonnées utilisées pour calculer le barycentre

Distance entre les deux coordonnées utilisées pour interpoler	Nombre d'individus	Part d'individus
Une seule coordonnée utilisée	8 385	8,2%
a. Coordonnées identiques	2 170	2,1%
b. Moins de 100 mètres	26 831	26,2%
c. Entre 100 et 500 mètres	35 242	34,5%
d. Entre 500 et 1000 mètres	12 872	12,6%
e. Entre 1000 et 10000 mètres	16 515	16,2%
f. Plus de 10km	215	0,2%
TOTAL	102 230	100,0%

Source : EAR 2017, *individus géolocalisés avec les méthodes GA et GI (référentiel de l'année N)*
Champ : Individus des ménages *résidant dans les* petites communes *de métropole* enquêtées lors de l'EAR 2017 (groupe de rotation n°4) dont la coordonnée est obtenue par interpolation.
Lecture : Pour *8 385* individus (8,2 % des individus dont l'adresse est obtenue par interpolation), une seule coordonnée est utilisée pour l'interpolation.

2.2.4. Part d'individus interpolés au niveau communal

Au total, 102 486 individus sont interpolés, ce qui représente 1,6 % des individus (Tableau 12). Selon les communes, cette proportion peut varier entre 0 et 100 % d'individus interpolés.

Tableau 12: Part d'individus interpolés par district et par commune

Part d'individus interpolés	Par commune			Par district		
	Nombre de communes	Part de communes	Nombre d'individus interpolés	Nombre de districts	Part de districts	Nombre d'individus interpolés
a. Moins de 1% d'individus interpolés	3 929	56,4%	13 031	11 786	65,1%	7 106
b. Entre 1 et 5% d'individus interpolés	1 846	26,5%	40 599	4 208	23,2%	36 266
c. Entre 5 et 10% d'individus interpolés	766	11,0%	23 680	1 372	7,6%	26 669
d. Entre 10 et 50% d'individus interpolés	374	5,4%	15 691	675	3,7%	22 532
e. Entre 50 et 100% d'individus interpolés	44	0,6%	9 253	55	0,3%	9 657
f. 100% d'individus interpolés	3	0,0%	232	4	0,0%	256
TOTAL	6 962	100,0%	102 486	18 100	100,0%	102 486

Source : EAR 2017, *individus géolocalisés avec les méthodes GA et GI (référentiels de l'année N)*
Champ : Petites c*ommunes de métropole de* l'EAR 2017 (groupe de rotation n°4)
Lecture : *Dans 3 929 communes, il y a moins de 1 % d'individus dont les coordonnées sont obtenues par interpolation. Cela représente 56,4 % des communes. Dans ces communes, 13 031 individus sont interpolés.*

Il y a 421 communes dans lesquelles on interpole plus de 10 % des individus, ce qui représente 25 176 individus.

Conclusion

Cette étude a permis de définir la méthode de géolocalisation du recensement en petites communes à partir de deux méthodes distinctes : la méthode d'appariement sur les éléments d'adressage (GA) et la méthode d'appariement à partir de caractéristiques individuelles des individus (GI). La comparaison des résultats de ces deux méthodes a permis d'améliorer les informations sur la qualité des coordonnées produites, et ainsi de définir une démarche optimale pour choisir la meilleure coordonnée pour chaque adresse du recensement.

Annexes

Annexe 1 : Règle de choix des coordonnées, en nombre d'adresses

Tableau 13: Choix des coordonnées (en nombre d'adresses)

Nouvelle note synthétique méthode GA	Variable	Nouvelle note synthétique méthode GI					
		1. Bon	2. Moyen	3. Mauvais	4. Non apparie (interpolé)	5. Non apparie (pas de coordonnées)	TOTAL
1. Bon	Nombre d'adresses	1 412 456	249 534	59 268	76 001	8 678	1 805 937
	Part d'adresses	*58,6%*	*10,4%*	*2,5%*	*3,2%*	*0,4%*	*75,0%*
2. Moyen	Nombre d'adresses	120 752	22 010	5 180	7 722	2 932	158 596
	Part d'adresses	*5,0%*	*0,9%*	*0,2%*	*0,3%*	*0,1%*	*6,6%*
3. Mauvais	Nombre d'adresses	220 668	40 090	10 524	16 977	3 887	292 146
	Part d'adresses	*9,2%*	*1,7%*	*0,4%*	*0,7%*	*0,2%*	*12,1%*
4. Non apparié	Nombre d'adresses	113 509	22 130	5 432	10 134	1 422	152 627
	Part d'adresses	*4,7%*	*0,9%*	*0,2%*	*0,4%*	*0,1%*	*6,3%*
TOTAL	Nombre d'individus	1 867 385	333 764	80 404	110 834	16 919	2 409 306
	Part d'individus	77,5%	13,9%	3,3%	4,6%	0,7%	*100,0%*

Choix retenu	Nombre d'individus	Part d'individus
On choisit la méthode GA	**1 821 771**	**75,6%**
dont : avec de bonnes coordonnées	1 805 937	75,0%
dont : avec des coordonnées moyennes	15 834	0,7%
On choisit la méthode GI	**539 159**	**22,4%**
dont : avec de bonnes coordonnées	*454 929*	18,9%
dont : avec des coordonnées moyennes	*84 230*	3,5%
On interpole ou on impute	**48 376**	**2,0%**
TOTAL	2 409 306	100,0%

Source : EAR 2017, *individus géolocalisés avec les méthodes GA et GI (référentiels de l'année N).*
Champ : Individus des ménages *résidant dans les* petites communes *de métropole* enquêtées lors de l'EAR 2017 (groupe de rotation n°4).
Lecture : Parmi les *2 409 306 adresses de résidence principales* de l'EAR 2017, *1 412 456* (58,6 %) ont une bonne note globale avec la méthode GA et la méthode GI. Dans ce cas, on privilégie la coordonnée fournie par la méthode GA.

La règle de décision choisie conduit à retenir la méthode GA pour 1 821 771 adresses (75,6 %), la méthode GI pour 539 159 adresses (22,4 %), et à interpoler ou imputer les coordonnées pour 48 376 adresses (2,0 %) [Tableau 13].

Annexe 2 : Utilisation d'un référentiel retardé pour la méthode GI lors de la production courante

La production du référentiel géolocalisé *ad hoc* à la méthode GI est longue. En particulier, il ne sera pas possible en régime courant d'utiliser le référentiel au 1er janvier N pour géolocaliser l'EAR N (ce qui a été fait dans les études exploratoires), mais seulement celui au 1er janvier N-1. Ce décalage conduit en pratique à ce qu'un plus grand nombre d'individus ne soient pas appariés (Tableau 14) : en utilisant le référentiel N-1, 8 % des individus (et 10 % des adresses) ne sont pas appariés, contre 4 % des individus (et 5 % des adresses) quand les fichiers fiscaux ont la même année de validité que l'EAR.

Tableau 14: Nombre d'individus appariés avec la méthode GI, selon que l'on utilise le référentiel au 1er janvier N ou au 1er janvier N-1

		Appariement avec les fichiers fiscaux N-1		
		Individus non appariés	Individus appariés	TOTAL
Appariement avec les fichiers fiscaux N	Individus non appariés	213 970	19 774	233 744
	Part	*3%*	*1%*	*4%*
	Individus appariés	333 167	5 959 071	6 292 238
	Part	*5%*	*91%*	*96%*
	TOTAL	547 137	5 978 845	6 525 982
	Part	*8%*	*92%*	*100%*

Source : EAR 2017, individus géolocalisés avec la méthode GI (référentiel de l'année N et référentiel de l'année N-1).
Champ : Individus des ménages résidant dans les petites communes enquêtées de métropole lors de l'EAR 2017 (groupe de rotation n°4).
Lecture : Parmi les 6 525 982 individus, 6 292 238 (96 %) sont appariés avec la méthode GI en utilisant le référentiel N, et 5 978 845 (92 %) sont appariés en utilisant le référentiel N-1

Ce phénomène est largement imputable aux déménagements. En effet, dans le cas d'un déménagement pendant l'année précédant l'EAR, il est peu probable de trouver l'individu dans le bon logement dans le référentiel de l'année précédente (sauf si une partie du ménage seulement a déménagé). En pratique, 82 % des 333 167 individus qui ne sont pas appariés avec le référentiel N-1 (alors qu'ils l'étaient avec les fichiers fiscaux N) déclarent avoir déménage au cours des 12 derniers mois ; ils ne sont que 6 % parmi les 5 959 07 individus appariés à la fois avec le référentiel N et N-1.

À noter qu'une absence d'appariement ne signifie pas nécessairement une dégradation trop forte de la qualité de la géolocalisation, en raison notamment de la méthode d'interpolation mise en œuvre en cas d'échec de l'appariement. Par ailleurs, parmi les individus qui sont appariés par la méthode GI avec le référentiel N et avec le référentiel N-1, certains peuvent être reliés à des adresses différentes dans les deux

cas. Cette situation se produit notamment lorsqu'un individu a déménagé dans la même commune (Tableau 15)[8].

Tableau 15: Comparaison de la méthode GI avec des référentiels N et N-1 : individus appariés et dans le même carreau

Individus :	Individus appariés avec les deux méthodes	Individus non appariés avec au moins l'une des deux méthodes	TOTAL
Avec des coordonnées identiques	5 726 927	144 630	5 871 557
Part	*88%*	*2%*	*90%*
Avec des coordonnées différentes	232 144	422 281	654 425
Part	*4%*	*6%*	*10%*
TOTAL	5 959 071	566 911	6 525 982
Part	*91%*	*9%*	*100%*

Source : EAR 2017, individus géolocalisés avec la méthode GI (référentiel de l'année N, et référentiel de l'année N-1).
Champ : Individus des ménages de petite commune de métropole de l'EAR 2017.
Lecture : Parmi les 6 525 982 individus, 90 % ont des coordonnées identiques. Parmi eux, 90 % ont des coordonnées identiques : 88 % sont appariés avec les deux méthodes, 2 % ne sont pas appariés avec au moins l'une des deux méthodes.

En tout et pour tout, utiliser le référentiel N-1 pour la méthode GI conduit à modifier les coordonnées de 10 % des individus : dans 6 % des cas, il s'agit d'individus non appariés dans l'un des deux millésimes, dans 4 % des cas il s'agit d'individus appariés dans les deux millésimes mais à des adresses différentes. Au total, le carreau de 1 km x 1 km est modifié par ce phénomène dans seulement 2 % des cas.

8 La méthode GI procède commune par commune : un individu ayant changé de commune ne pourra être retrouvé à tort dans sa commune de résidence antérieure.