
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

Dans les communes de plus de 10 000 habitants, les personnes recensées sont facilement géolocalisables grâce au Répertoire d'immeubles localisés (RIL) qui sert notamment à la constitution de la base de sondage des adresses enquêtées par le recensement. À l'inverse, dans les communes métropolitaines de moins de 10 000 habitants (dénommées aussi « petites communes »), les données collectées sur les individus et les logements lors du recensement de la population ne contiennent pas d'information sur leur géolocalisation précise. Les coordonnées géographiques (x,y) des unités enquêtées en petites communes sont ainsi inconnues.

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.

Cet article présente la manière dont sont articulées deux méthodes utilisées depuis plusieurs années à l'Insee pour géolocaliser des données. Pour tirer le meilleur parti des deux méthodes de géolocalisation, l'enjeu principal est de pouvoir les comparer alors qu'elles proposent chacune un système différent de scores de qualité.

La première partie de cet article présente brièvement les deux méthodes de géolocalisation mobilisées pour identifier les coordonnées (x,y) des logements des individus enquêtés dans les petites communes en vue de détailler, dans la seconde partie, une comparaison des deux méthodes et de présenter la règle de décision retenue pour les combiner.

Les deux méthodes de géolocalisation utilisées consistent à rapprocher les données du recensement d'un référentiel géolocalisé constitué à partir de fichiers fiscaux (cadastre). Ces deux méthodes diffèrent par la manière d'opérer ce rapprochement :

- la première méthode s'appuie sur les éléments d'adressage (numéro, type de voie, libellé de voie, complément d'adresse, commune). 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 deuxième méthode repose sur des caractéristiques non-nominatives des individus résidant dans chaque bâtiment. Les données sur les date de naissance, sexe, commune de naissance et commune de résidence des habitants d'une adresse permettent d'établir une correspondance entre un bâtiment enquêté par le recensement et un bâtiment du référentiel. Cette méthode est appelée « appariement probabiliste ».

Dans les deux cas, la qualité de la géolocalisation est *in fine* qualifiée selon une note propre à chaque méthode. Un système de notation harmonisé est alors défini pour permettre de comparer les deux méthodes. La qualité de la géolocalisation est ainsi décrite en 4 catégories : bonne, moyenne, mauvaise ou non géolocalisé.

L'étude se concentre alors sur l'analyse des résultats des deux méthodes de géolocalisation respectives pour les observations enquêtées par l'enquête annuelle de recensement 2017. L'objectif est de définir une règle de décision entre les deux méthodes, basée sur les notes qualité, afin de retenir la géolocalisation de meilleure qualité pour chaque observation.

De façon générale, la géolocalisation par l'adresse est légèrement plus performante : 76 % des individus enquêtés obtiennent une géolocalisation de bonne qualité contre 72 % avec la méthode par appariement probabiliste. Les résultats sont identiques à 10 mètres près dans 85 % des cas des adresses géolocalisées avec une bonne qualité par les deux méthodes. Lorsque la géolocalisation par l'adresse est de bonne qualité, c'est cette méthode qui est privilégiée (76 % des cas). Ensuite, si la géolocalisation par l'adresse n'est pas de bonne qualité mais que celle par appariement probabiliste est de bonne qualité ou de qualité moyenne, c'est cette dernière qui est retenue (22 % des cas). On observe en effet que la méthode de géolocalisation par appariement probabiliste est plus performante dans les situations où les éléments d'adressage ne sont pas très précis, par exemple, lorsqu'ils se limitent au nom d'un lieu-dit. On retient ensuite la géolocalisation à l'adresse de qualité moyenne dans les cas restants où l'appariement probabiliste n'aboutit pas ou ne fournit qu'une géolocalisation de mauvaise qualité (0,5 % des cas). Enfin, dans les cas résiduels où les deux méthodes fournissent des résultats de mauvaises qualité ou n'aboutissent pas, la géolocalisation des observations est interpolée : on approxime les coordonnées des dernières observations en utilisant le district de collecte, le rang d'adresse et la géolocalisation (de bonne qualité) des observations « voisines ».

Bibliographie

[1] Jabot P., Treyens P-E., « Appariement de l'enquête CARE par identification du plus proche écho », *Journées de méthodologie statistiques de l'Insee (JMS 2018)*, juin 2018.