Cahiers de démographie locale 2008, Néothèque, coll.
« Dynamiques des populations locales », 2008, pp. 221-234.

Sources administratives et recensement : la fin d'une alternative

Jean-Luc LIPATZ*

Les relations entre recensement de la population et sources administratives ont toujours été pensées comme complémentaires voire quelquefois antagonistes. Aujourd'hui, le sentiment de cette opposition est exacerbé de deux manières. D'un coté, la pression se fait sans cesse plus forte pour une observation détaillée de ce qui se passe à l'intérieur de nos villes, pression parfois même concrétisée par une injonction de la part du législateur. Et parallèlement, le développement d'outils géographiques et statistiques originaux autour d'une mobilisation fine de données de gestion laisse penser qu'il n'y a plus de limites au détail de l'observation.

L'organisation même du recensement actuel semble renforcer cet antagonisme : alors qu'autrefois un recensement avait toujours été une source vite périmée, la nouvelle organisation le place dans un contexte de fourniture d'informations fraîches et régulièrement actualisées, c'est-à-dire sur le même créneau que les sources d'origine administrative. Mais au contraire de celles-ci, il y a des limites à la finesse de l'échelle géographique d'exploitation des résultats. A la complémentarité d'autrefois - les données de gestion assuraient l'entre deux recensements - se substitue donc une simultanéité qui ressemble fort à une confrontation où l'un des adversaires part avec un handicap majeur. Est-ce à dire que les données de gestion seraient amenées à se substituer aux données de recensement ? Que la statistique locale fine reposerait uniquement sur les sources administratives tandis que l'usage du recensement serait réservé à des missions plus élevées ? Une réponse positive - prononcer le divorce entre les deux filières - est cependant bien difficile à faire tant le sacrifice résultant est grand :

* Responsable de la division « Etudes territoriales » à la direction générale de l'Insee. La division « Etudes territoriales » est chargée du développement des sources et des méthodes pour l'analyse des villes à une échelle infra-communale.

- beaucoup d'informations ne sont présentes que dans les recensements ;
- l'introduction d'une dimension géographique fine dans les données de gestion n'est pas gratuite du tout ;
- le passage de la donnée de gestion à la donnée statistique est loin d'être immédiat.

A regarder de près, pourtant, les deux systèmes n'ont jamais été aussi imbriqués. Si l'on examine l'ensemble du dispositif, le recensement repose lourdement sur l'utilisation de données de gestion. Dans sa constitution, la base de sondage des adresses a recours - au moins partiellement - aux fichiers de la taxe d'habitation, aux déclarations de permis de construire, et à toute information extérieure facilement mobilisable : exactement ce qu'offrent les données de gestion. L'établissement des populations légales sera également un savant mélange de données administratives et de données issues des enquêtes annuelles.

Vu du coté adverse, le développement exponentiel à l'INSEE des statistiques localisées à partir de sources administratives aurait été considérablement plus difficile sans les bases de sondage réalisées pour le recensement : celles-ci produisant « gratuitement » des tables de correspondance entre libellés d'adresse et éléments de géo-référencement.

L'expression « prononcer le divorce » prend donc toute sa valeur tant il y a cohabitation, et un divorce semble bien inimaginable. Mais ce qui vient d'être évoqué ne touche qu'à l'univers de la production de données, petit monde restreint à quelques techniciens. Peut-on aller plus loin ? Peut-on imaginer une cohabitation qui produise une information plus riche que la simple somme des informations issues des deux systèmes ?

En première ligne sur ces sujets, la division « Etudes territoriales » de l'INSEE a été fortement impliquée à deux reprises dans des travaux visant à trouver cette synergie manquante. Les deux dispositifs qui vont être évoqués maintenant reposent sur les mêmes idées autour d'ingrédients à peine différents mais à des échelles géographiques complémentaires qui montrent qu'il ne s'agit pas seulement de trouver un palliatif d'une prétendue déficience mais bien de construire une nouvelle voie associant le meilleur des deux systèmes.

L'« investissement zones mixtes »

Mi-2006, le recensement en était à sa troisième enquête annuelle. Avec une telle masse d'informations, il était nécessaire de se poser la question d'une diffusion de résultats du style des « chiffres clés » déjà produits et diffusés sur les plus grandes communes mais, ceci, à l'échelle de territoires géographiquement plus étendus (aires urbaines, communautés d'agglomération, etc.) comprenant les communes de moins de 10 000 habitants n'ayant pas fait parties des trois premières vagues de recensement. La réponse à cette question fut finalement mise à disposition des directions régionales un an plus tard, dans un contexte strictement limité à des travaux en partenariat avec des acteurs locaux. L'outil réalisé, l'« investissement zones mixtes » dans le jargon de ses concepteurs, produit des estimations sur la population des ménages datées au 1er janvier 2005. Il a été conçu d'emblée comme un outil jetable destiné uniquement à la période transitoire avant que la totalité d'un cycle de collecte du recensement soit terminé. Il affronte donc des obstacles spécifiques à ce contexte mais fournit aussi des pistes pour la résolution de difficultés inhérentes au nouveau processus de collecte et qui perdureront en régime de croisière. Cependant, c'est sur le premier aspect, l'incomplétude de la collecte, que le dispositif élaboré montre le mieux comment on peut imaginer une collaboration étroite entre données du recensement et données de gestion.

Concrètement le problème se résume à essayer de construire un chiffre global sur des zones formées d'un ensemble de communes de tailles variables et donc potentiellement enquêtées avec des niveaux de précision différents, à des dates différentes voire même pas encore enquêtées. En effet en 2006 restaient encore deux années de collecte à réaliser. La solution est donc nécessairement une combinaison de solutions à plusieurs problèmes.

- Pour les communes de plus de 10 000 habitants, des estimations existent mais toutes n'ont pas une précision suffisante pour avoir donné lieu à une publication. Dans ce cas il va falloir, soit vérifier que le manque de précision ne compromet pas l'estimation sur une zone plus étendue, soit regarder dans le détail les chiffres obtenus.

- Certaines communes de moins de 10 000 habitants ont été enquêtées en 2005 : leurs données ne posent pas de problème spécifique.
- Certaines communes de moins de 10 000 habitants ont été enquêtées en 2004 ou en 2006 : il va falloir imaginer ce que sera ou ce qu'a été leur situation en 2005. Estimer les évolutions sur un an ne peut se faire sans recours aux données de gestion. Mais, pratiquement, une simple extrapolation ou une simple interpolation fournit le résultat. Comme les données de gestion ne permettent pas d'accéder aux caractéristiques fournies par le recensement, cette interpolation ou extrapolation ne sera appliquée qu'au volume de la population totale des ménages, les structures fines (répartition par âge par exemple) seront reprises directement des enquêtes de recensement.
- Certaines communes de moins de 10 000 habitants n'ont pas été enquêtées : il va falloir les estimer en se passant d'informations récentes issues du recensement. Les données de gestion entrent en piste, avec une logique en deux temps comme précédemment : estimation des volumes et réplication des structures - ici à partir du dernier recensement.

Exploiter les visions décalées

L'idée pour estimer les petites communes non enquêtées repose en fait au départ sur un simple constat d'impuissance : les données de gestion ne pourront jamais remplacer un recensement parce qu'elles décrivent une réalité de gestion qui n'a aucune raison de correspondre à ce que mesurerait un recensement.

Par contre on peut raisonnablement faire l'hypothèse qu'il y a une certaine logique à ce décalage : que les choses s'expriment plus en termes de biais statistique que d'incertitude (de variance) et qu'il est donc possible de formuler une règle un peu systématique permettant de passer d'une réalité à une autre. Cette règle pourrait être un simple facteur de proportionnalité, du style « je sais que 5 % des déclarations fiscales me manquent donc je rajoute 5 % à ce nombre pour obtenir une estimation du nombre de ménages ».

Mais, à regarder de près, de nombreux indices militent pour une règle de transition qui soit sensiblement différente d'un lieu à un autre : l'organisation de la gestion autour de centres ayant chacun la charge d'une portion du territoire et éventuellement des pratiques de gestion subtilement différentes ou des caractéristiques purement géographiques qui influeraient sur la nature du biais entre données de gestion et données de recensement.

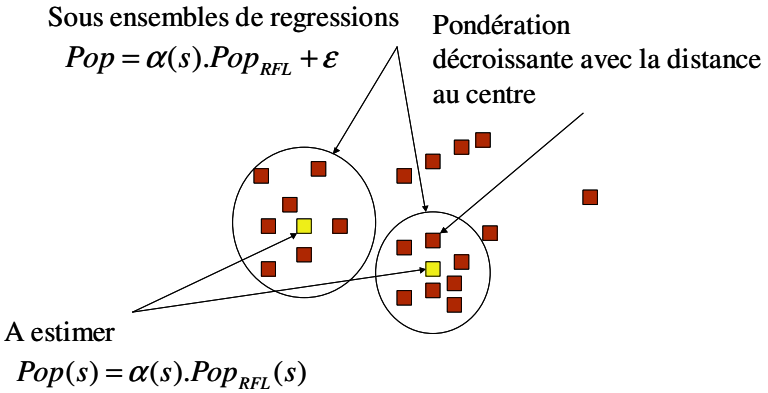
Pour citer un exemple basé sur les déclarations fiscales de revenu, source qui a été privilégiée dans l'outil « zones mixtes », on sait que les propriétaires de résidences secondaires ont des incitations à déclarer celles-ci comme résidences principales alors qu'ils seront néanmoins recensés sur leur lieu d'habitation courant. Comme l'implantation des résidences secondaires n'est pas complètement aléatoire spatialement, il y a fort à parier que suivant les lieux on observera des biais différents. Une description adéquate de la règle de passage entre les données fiscales et celles des ménages du recensement devra donc tenir compte du type de localisation géographique.

La régression géographique pondérée

Une fois admises ces deux hypothèses - l'existence d'une relation entre les données et sa possible dépendance avec le lieu où on se place -, tout n'est plus question que d'outillage, de modélisation de la relation. Certes, il s'agit de travailler sur des communes où par définition on ne connaît que la donnée de gestion (le nombre de ménages fiscaux), mais il existe également toutes les communes qui ont fait l'objet d'une collecte en 2004, 2005, 2006. Et à partir de celles-là, moyennant une petite correction pour se mettre à une date de référence unique (2005), il est possible d'établir la relation existant entre source administrative et recensement. Ensuite, pour revenir à la question de départ, l'estimation cherchée sera produite par simple extrapolation de la relation trouvée aux communes non enquêtées.

Il existe plusieurs manières de traiter ce problème de modélisation en travaillant sur des techniques de régression statistique adaptées à un contexte où l'espace lui-même fait partie des variables explicatives du modèle. Deux pistes ont été explorées : la régression avec intégration de l'auto-corrélation spatiale et la régression géographique pondérée

suggérée par Fotheringham, Brunsdon et Charlton en 1996. C'est cette deuxième piste qui a fini par être choisie en raison de sa plus grande souplesse et de son plus grand pouvoir d'adaptation aux changements géographiques de régime de la relation cherchée. L'idée en est simple : au lieu d'estimer un unique modèle « France entière », on en estime autant qu'il y a de communes manquantes et on réalise cette estimation sur la base de ce qu'on observe dans le voisinage de chaque commune (à une trentaine de kilomètres), en accordant un crédit d'autant plus fort que les observations utilisées sont proches géographiquement du point d'estimation.



Le recensement à l'infra-communal

Le cas de l'infra-communal dans les communes de plus de 10 000 habitants relève d'une situation très similaire à celle des communes de moins de 10 000 habitants, mais exprimée en termes d'adresses enquêtées et non de communes enquêtées. Cette similitude va jusqu'à la nature de la collecte qui est exhaustive dans les deux cas. Dès lors, la tentation d'utiliser les mêmes idées que dans l'outil « zones mixtes » est naturellement assez forte, et d'ailleurs, historiquement, les deux chantiers ont été initiés simultanément. Par contre, la particularité de l'infra-communal est que la collecte ne sera jamais complète : au bout de cinq ans on ne disposera que d'environ 40 % des adresses et donc tout processus d'estimation est, à priori, bon à étudier y compris dans une optique de régime permanent. Au contraire de l'outil « zones mixtes », une fois complètement outillé, ce processus d'estimation sur l'infra-

communal est donc bien destiné à entrer dans une boite à outil permanente destinée aux études locales. Ceci, dans des conditions d'utilisation qui restent cependant encore à définir et qui pourraient aller jusqu'à la production de données millésimées, par exemple pour les besoins d'observation autour de la politique de la ville.

En effet, deux principes de modélisation peuvent être imaginés. Un premier, que l'on pourrait qualifier de « compatible recensement » produirait des données moyennes sur l'ensemble du cycle de collecte, par simple agrégation de l'information sans tenir compte de sa date de recueil : c'est ce principe qui a fait l'objet de l'essentiel des investigations jusqu'ici. Mais il existe une autre possibilité de modélisation intégrant explicitement le temps dans sa formulation : il s'agit alors de produire simultanément une estimation d'une grandeur millésimée et - sous réserve d'une hypothèse assez forte sur la stabilité du biais administratif dans le temps - les taux d'évolution annuels.

Sur deux ans (pour simplifier), cela conduit à estimer simultanément les paramètres de deux équations :

$$PopRP_{2005} = \alpha.PopAdm + \varepsilon$$

$$PopRP_{2004} = \alpha.PopAdm + \beta.PopAdm + \varepsilon$$

Le premier paramètre fournit alors le biais administratif et donc les moyens de produire l'estimation millésimée, tandis que le second paramètre est une combinaison de ce biais et de l'évolution entre les deux années.

Les techniques de résolution des modèles sont les mêmes que dans l'outil « zones mixtes » avec des réglages parfois communs : la loi de décroissance des pondérations avec la distance est une fonction quadratique classique (le « biweight »). Mais les réglages ont été adaptés à la dimension géographique plus restreinte. Dans le cas des travaux sur l'infra-communal, le choix du voisinage résulte d'une optimisation au cas par cas venant d'un arbitrage entre le nombre d'observations utilisées dans la modélisation (un plus grand nombre donnant une meilleure estimation) et l'intégration d'observations de plus en plus lointaines (donc de comportements de moins en moins voisins).

Les données mobilisées sont également différentes dans les deux cas. Alors que l'outil « zones mixtes » travaillait sur des données fiscales

dont la couverture de la population totale est quasiment parfaite, les travaux sur l'infra-communal reposent pour des raisons pratiques sur deux sources : les données des bénéficiaires de l'assurance maladie (CNAM/TS), disponibles sur des millésimes plus récents mais ne couvrant qu'environ 80 % de la population, ainsi que les comptages de logements issus de la base de sondage du recensement. Les données utilisées à l'infra-communal sont donc dès le départ imparfaites, mais ceci ne remet pas en cause la pertinence de la méthode, puisque son principe même est de mesurer l'imperfection des données de gestion et de prendre appui sur une mesure objective de cette imperfection pour suggérer une correction. Que les données de la CNAM/TS soient partielles, insuffisamment localisées à l'adresse, que le RIL ne suive pas fidèlement le nombre de logements dans des zones dont le bâti évolue n'est pas un handicap. La seule hypothèse forte est que ces imperfections ne soient pas de nature aléatoire, or tant ce qui peut sous-tendre les trous de couverture de la CNAM/TS (la structure socio-démographique de la population) que ce qui se cache derrière les défauts du RIL (la rénovation urbaine) sont des phénomènes tellement lourds qu'on peut sans trop de crainte postuler qu'ils ne sont pas uniquement ponctuels.

Pour autant, la question de la validation ne peut pas être complètement évacuée. Et il existe plusieurs moyens sinon de produire une fourchette d'incertitude, au moins de se rassurer sur la pertinence d'opérations qui, au final, peuvent sembler quelque peu magiques.

Il y a d'abord le comportement du modèle sur les localisations enquêtées : vérifier qu'il prédit bien ce qu'on observe en oubliant momentanément qu'on a fait une observation à cet endroit.

Dans le même ordre d'idées, dans le cas de l'outil « zones mixtes », les traitements ont été testés de façon empirique en comparant les estimations produites avec les chiffres réels issus de sources (de gestion) sur lesquels on disposait de séries complètes. Comme il ne s'agissait pas de données de recensement, le message positif recueilli n'est pas suffisant mais à l'inverse les défauts de précision relevés ont dicté quelques règles élémentaires de situations à éviter : zones trop petites, trop hétérogènes, etc..

On peut compléter ces informations par un calcul théorique de précision en s'inspirant de ceux de la théorie des sondages. Et même si ces calculs

sont difficiles à mener en raison de la complexité du contexte de tirage des échantillons du recensement, les indications recueillies convergent bien avec les écarts observés sur les tests empiriques. Enfin, pour les plus difficiles à convaincre, il est toujours possible de mettre à l'épreuve l'hypothèse de stationnarité géographique du biais administratif : la cartographie des corrélations spatiales est un outil précieux pour déterminer les lignes de rupture qui porteront les points de faiblesse de la méthode d'estimation.

Pour ce qui est de l'infra-communal, les résultats obtenus ici et là parlent d'eux-mêmes. Alors que la méthode proposée n'utilise pas une partie des ingrédients standards du recensement (les poids de sondage) elle restitue néanmoins des estimations cumulées par commune complètement compatibles avec celles réalisées avec le recensement seul. Il n'y a qu'un seul détail à mentionner, c'est que dès maintenant, avec une collecte incomplète, elle produit des descriptions spatiales et des quantifications sur des territoires qui ne seront, demain, peut-être pas accessibles au recensement dans son régime de croisière.

Conclusion

Avec l'application de ces méthodes, l'INSEE inaugure une nouvelle manière d'envisager la statistique urbaine et spatiale. La statistique infra-communale a longtemps été pensée comme une simple déclinaison de ce qui se faisait à l'échelle des régions ou des départements : c'est-à-dire comme un simple accroissement de la finesse du découpage géographique. Si pour étudier le fonctionnent interne d'une région on avait recours au maillage communal ou à des zonages d'étude spécifiques comme les aires urbaines, il suffisait pour le cas de l'infra-communal de trouver des objets géographiques jouant le même rôle. Dans cette optique, la diffusion statistique à l'infra-communal s'est organisée au début des années 2000 autour de zonages infra-communaux standard : les IRIS pour produire un maillage complet du territoire infra-communale et quelques familles de quartiers issus du monde de la politique de la ville (Zones Urbaines Sensibles et Zones Franches Urbaines). Ce choix correspondait et correspond toujours à une nécessité : celle d'organiser la production et la diffusion de données sur des objets suffisamment stables dans le temps et de taille suffisamment grande pour garantir la confidentialité des données individuelles. A

contrario l'image produite par ces zonages est souvent trop réductrice voire parfois erronée, aussi une analyse fine des phénomènes de spécialisation territoriale à l'intérieur des villes requiert d'aller au delà d'une première description tracée à grands traits.

L'abandon de l'îlot comme support de la collecte du recensement simultanément avec le développement des outils permettant une géolocalisation fine en coordonnées géographiques au niveau de chaque adresse, placent l'observation urbaine à un moment charnière où s'ouvrent de nouvelles perspectives : celles de la statistique spatiale non zonée. Alors que, classiquement, la vision d'une ville se résumait à la juxtaposition de données sur des mailles séparées, il est maintenant possible d'exprimer celle-ci en termes d'interactions entre entités élémentaires et ceci, quasiment sans restriction sur la finesse du grain utilisé.

Au sein de l'INSEE, la division « Etudes territoriales » travaille depuis plusieurs années dans ce nouvel état d'esprit en élaborant pour les directions régionales de l'INSEE des outils, et les données nécessaires à ces outils, aptes à produire des descriptions fiables des disparités internes aux villes. Localement, ces outils commencent à être mis en œuvre pour le compte de collectivités territoriales dans le cadre de travaux en partenariat avec celles-ci. Au niveau national, les mêmes outils ont été à l'origine des cartes préparatoires à la reflexion sur les nouveaux quartiers de la politique de la ville qui ont été diffusées sur le site de la Délégation Interministerielle à la Ville.

Références bibliographiques

BRUNSDON Chris, FOTHERINGHAM A. Stewart, CHARLTON Martin, 1996, « Geographically Weighted Regression : A Method for Exploring Spatial Non-stationarity », *Geographical Analysis*, 28(4), pp. 281-298.

BRUNSDON Chris, FOTHERINGHAM A. Stewart, CHARLTON Martin, 1996, « The Geography of Parameter Space : An Investigation into Spatial Non-Stationarity », *International Journal of Geographic Information Systems*, 10, pp. 605-627.

Liste des sigles

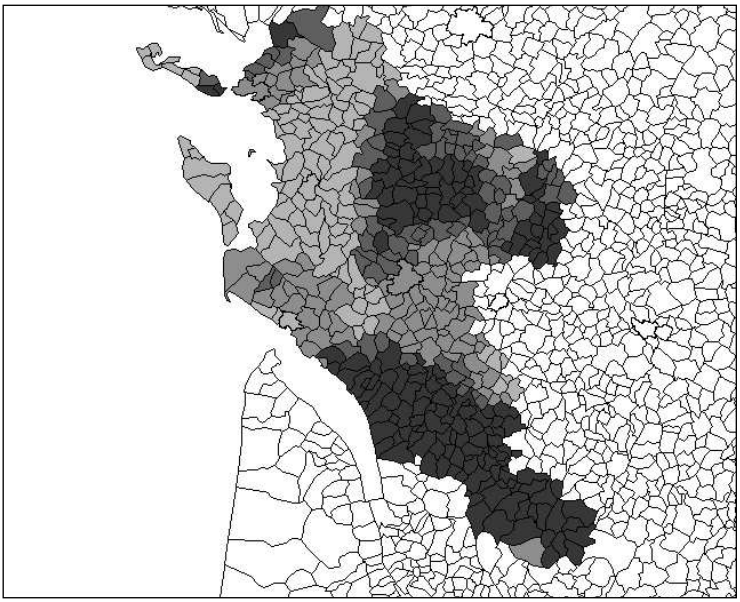
CNAM/TS : Caisse Nationale d'Assurance Maladie des Travailleurs Salariés

INSEE : Institut National de la Statistique et des Etudes Economiques

RIL : Répertoire d'immeubles localisés

ZUS : Zone Urbaine Sensible

Annexe 1. Un exemple de biais géographique : sous-représentation (après lissage) du recensement par rapport aux données fiscales (en gris clair) sur les communes du département de Charente maritime.



Annexe 2. Estimation de la population dans les zones urbaines sensibles de Strasbourg

dc	zus	est_pop
67482	4201030	4303
67482	4201090	7689
67482	4201100	9357
67482	4201120	13412
67482	4201130	11188
67482	4201140	7515
67482	4201180	11436
67482	4201190	1346
67482	HORSZONE	188980

La population totale communale estimée pour les ménages est de 255 227 personnes. Le chiffre « vrai » issu du recensement seul est de 259 399.

Jean-Luc LIPATZ

Sources administratives et recensement : la fin d'une alternative

Alors qu'autrefois un recensement avait toujours été une source vite périmée, la nouvelle organisation le place aujourd'hui dans un contexte de fourniture d'informations fraîches et régulièrement actualisées, c'est-à-dire sur le même créneau que les sources d'origine administrative. Mais au contraire de celles-ci, il y a des limites à la finesse de l'échelle géographique d'exploitation des résultats. La division « Etudes territoriales » de l'INSEE travaille depuis plusieurs années à l'élaboration de nouvelles méthodes basées sur l'association des données du recensement et celles provenant des sources administratives, afin de produire des descriptions fiables des disparités internes aux villes. Il ne s'agit pas seulement de trouver un palliatif d'une prétendue déficience, mais bien de construire une nouvelle voie associant le meilleur des deux systèmes.