La valeur des données géographiques

Christophe Terrier

Je précise au préalable mon point de vue du monde : les données dont je parle sont essentiellement de nature socio-démographique ; elles sont « géographiques » lorsqu'elles sont rattachées à un cadre géographique (lieu ou territoire). On peut bien sûr imaginer un autre point de vue pour lequel la véritable donnée serait ce que je considère comme un cadre. Cette différence de point de de vue devrait faire la richesse de notre débat.

Les données ne sont pas données, elles sont construites.

Les données de nature socio-démographique sont en général obtenues par la statistique. Or celle-ci ne consiste pas seulement à compter des individus ; elle nécessite d'avoir au préalable défini les cases dans lesquelles on va les dénombrer. Cette étape consistant à bien dire de quoi on parle est absolument fondamentale et une absence de consensus à ce stade ne pourra jamais être compensée ultérieurement, même par une grande qualité des enquêtes statistiques.

Les concepts

Cette construction nécessite que l'on s'accorde sur des concepts. S'il y n'y a pas, du moins en France, de divergence sur la notion d'individu en tant que personne humaine, tout se complique lorsqu'on aborde des constructions plus complexes comme le ménage ou l'entreprise. Il y a une trentaine d'années, l'Insee a entrepris de regrouper les nombreuses informations recueillies au travers des diverses enquêtes dans des « bases de données », et en particulier dans la Bdcom, base de données communales qui fût sa première véritable base de données géographiques. On s'est alors rendu compte que les concepts n'étaient pas harmonisés d'une enquête à l'autre, chaque concepteur adoptant le point de vue le mieux adapté au but recherché par son enquête, sans souci de cohérence avec une connaissance universelle. C'est ainsi que l'on dénombrait par exemple douze concepts différents de « ménage » et même vingt neufs définitions différentes du « chiffre d'affaires ». Il n'est en général pas possible de redonner a posteriori une cohérence à des informations collectées sur des bases conceptuelles différentes : c'est donc un très gros effort d'harmonisation des concepts qui a du être entrepris. C'est évidemment une opération de long terme. Elle ne peut pas produire de résultats rétroactifs et bien souvent, une évolution du concept gêne ou même empêche l'étude des séries longues.

Le concept de ménage est assez exemplaire pour mériter qu'on s'y attarde un peu. On essaie par ce concept de caractériser un ensemble d'individus qui vivent ensemble. On est passé de la notion ancienne de « feu » à la notion de « famille » pour se stabiliser après la dernière guerre mondiale sur la notion de « ménage » regroupant les individus vivant ensemble dans un même logement. On ne détaillera pas ici ces différents concepts ni leur évolution mais on évoquera une remise en cause profonde de l'actuel concept de ménage par un groupe de chercheurs démographes dans le cadre d'un groupe de travail animé par l'Unité de Méthodologie Statistique de l'Insee. En lisant l'excellent rapport de ce groupe, on ne peut qu'être convaincu que l'évolution des modes de vie a rendu inopérant ce concept sur lequel sont basées tant d'analyses.

On avait déjà émis bien des soupçons sur la notion de « chef de famille » qui correspondait à une autre époque où la place de la femme était « au foyer ». C'est maintenant, entre autres remises en question, le concept de « domicile » qui ne semble plus correspondre à la réalité contemporaine. Or il s'agit d'une donnée essentielle pour l'analyse géographique. Il est désormais jugé nécessaire de parler de « résidence habituelle », étant entendu qu'une même personne peut avoir plusieurs résidences habituelles entre lesquelles elle partage sa vie. Les gardes partagées des enfants, les lieux de travail ou d'étude éloignés, les multirésidences diverses occasionnées par certaines activités ou favorisées par la retraite sont autant de facettes de ce nouvel éparpillement des lieux de vie d'une même personne. Évidemment cette notion de « résidence habituelle » est plus difficile à cerner que la simple « résidence principale » classique qui avait le bon goût d'être unique. On est donc partagé entre un concept simple, relativement facile à obtenir mais ne reflétant plus la réalité de la société moderne et un concept mieux adapté mais plus complexe à enquêter et donc plus coûteux à obtenir.

On se retrouve actuellement avec une cohabitation de deux concepts opposés : les résultats

du recensement de la population sont toujours basés sur la notion de « résidence principale », que d'ailleurs on désigne sous le nom « résidence habituelle », ce qui complique encore un peu la compréhension tandis que d'autres enquêtes « ménages » de l'Insee -trop rares car trop coûteuses- s'efforcent de cerner la « résidence habituelle », éventuellement multiple, de l'individu. Le maintien de l'ancien concept pour le recensement permet d'entretenir l'illusion de pouvoir faire des analyses historiques. J'émets personnellement quelques réserves devant certaines interprétations des résultats des dernières enquêtes de recensement[1].

Les nomenclatures

Les nomenclatures forment le deuxième élément de la construction des données. Sans s'appesantir sur cette étape essentielle visant à définir des catégories et des regroupements, on soulignera l'existence, ici aussi, de la difficulté à mesurer l'évolution historique d'un phénomène. Si l'on veut bien rendre compte de la situation présente, il est souvent nécessaire d'y adapter la nomenclature ; et si l'on modifie la nomenclature entre deux points de mesure, il devient difficile d'effectuer des comparaisons.

Les nomenclatures territoriales, maillages ou zonages, méritent ici une attention particulière puisque l'on parle de données géographiques. Tantôt zonage de pouvoir (commune, département...), formant souvent l'ancrage géographique de la donnée, tantôt zonage de savoir (zone d'emploi, aire urbaine...), construit pour donner un sens à la donnée, aucune des nomenclatures actuelles n'est totalement invariante dans le temps. L'hétérogénéité de ces nomenclatures, par exemple entre les pays de l'Europe, nécessitent le recours à des techniques particulières de traitement – par exemple, le lissage- lorsque l'on veut procéder à des analyses géographiques débordant le cadre national.

La qualité des données

Divers éléments interviennent dans la qualité des données. L'un d'eux, particulièrement important pour les analyses géographiques, vient de l'appareil de mesure. La mesure peut être uniforme sur l'ensemble du territoire considéré : c'est en général le cas, pour la France, des données provenant de la statistique publique. Mais si la donnée est collectée indépendamment par chaque territoire – ce qui est le cas, avec la décentralisation, pour certaines données- il y aura un risque certain de biais géographique induit par des qualités différentes de collecte. Une mention particulière doit être faite pour les statistiques européennes qui sont collectées par chacun des services statistiques des pays membres mais dans le cadre d'une coordination plus ou moins poussée selon les sujets.

Un autre risque de biais, qui n'épargne pas les données géographiques, provient de la source de l'information. Les données peuvent être obtenues par des enquêtes adaptées ; elles peuvent également être obtenues par l'utilisation des sources administratives existantes. Cette dernière façon de faire est largement encouragée par le Conseil National de l'Information Statistique car elle permet d'éviter d'importuner par des enquêtes spécifiques des personnes ou des entreprises pour obtenir des informations qu'elles ont déjà fourni dans le cadre de leurs obligations administratives. Ces données doivent en général être utilisées avec des précautions particulières car elles ont été collectées avec des finalités souvent assez différentes de celles poursuivies par le chercheur. L'erreur classique – hélas trop fréquente- consiste par exemple à réaliser des études géographiques sur l'emploi à partir des données des organismes collecteurs de fonds. La finalité de ces organismes étant de collecter auprès des entreprises les prélèvement sociaux obligatoires, les conventions et les modes opératoires utilisés sont adaptées à cette finalité mais pas toujours compatibles avec une analyse localisée de l'emploi.

Les données ne sont pas données, elles sont transmises.

L'information sur l'information : les métadonnées

L'utilisateur des bases de données géographiques n'est en général ni le concepteur ni le

1Voir en ligne sur Cybergéo l'article **Démographie et mouvement**, contribution au débat sur le recensement de la population en continu http://cybergeo.revues.org/index23008.html

fabricant de ces données. Au delà du travail de conception -évoqué précédemment au travers des concepts et des nomenclatures- un travail de transmission est nécessaire. Il s'agit de bien informer sur la signification et les limites des informations produites : c'est l'information sur l'information – encore désignée sous le nom de « métadonnées »- sur laquelle de très gros efforts ont été réalisés par la statistique publique mais il reste à espérer que les utilisateurs des données les lisent vraiment.

Le prix de l'information

Toute information a un coût, qu'elle soit collectée par enquête ou par un traitement approprié de données administratives. Elle peut être transmise sous une forme gratuite ou payante, à un public large ou restreint, selon des formes rendant son usage plus ou moins aisé. Les politiques publiques n'ont pas toujours été constantes en ce domaine. Rappelons simplement qu'à l'heure actuelle une très grande masse de données « géographiques » est disponible gratuitement en ligne sur le site Insee.fr sous diverses formes (tableaux, cartes et données individuelles), le tout accompagné d'une importante méta-information. Soulignons cependant que si cette mise à disposition gratuite en ligne met l'information à portée de tout bon géographe un peu formé, il reste encore du chemin et du travail d'inter-médiation à faire pour rendre cette information utilisable par les politiques et les décideurs.

Les données géographiques et l'informatisation : une histoire liée.

L'histoire des bases de données -et celle des bases de données géographiques- est relativement récente. Il y a quarante ans, l'Insee était encore une des rares institutions françaises à s'être dotée de moyens de traitement automatique de l'information, d'abord mécanographiques puis informatiques. Le support des données était essentiellement la carte perforée, ultérieurement remplacée par la bande magnétique, supports surtout adaptés au traitement séquentiel. Corrélativement, les équipes étaient spécialisées, voire cloisonnées, ce qui favorisait la discordance des concepts utilisés pour les différentes enquêtes : ne devant pas mêler ni comparer les données provenant de sources multiples, il n'y avait pas de raison de chercher à les harmoniser. Par ailleurs le traitement informatique des données était l'affaire des « informaticiens » qui transmettaient aux « statisticiens » des tableaux de chiffres imprimés sur des liasses de papier. Les informations étaient ensuite mises à disposition sous forme d'annuaires statistiques imprimés.

La situation a commencé à évoluer lorsque les moyens de traitement se sont développés. Quelques centres informatiques universitaires ou inter-universitaires ont vu le jour. Quelques administrations, quelques grandes communes d'avant-garde se sont également équipées d'ordinateurs. Ces partenaires ont alors souhaité disposer de données brutes et non plus seulement de tableaux sur papier. Deux types de problèmes sont apparus : le premier, déjà évoqué, provenait des incohérences constatées entre des informations contenues provenant de sources statistiques différentes. Pour s'attaquer à ce problème il a fallu entreprendre un vaste travail – qui reste toujours d'actualité – pour harmoniser les concepts, mettre en cohérence les sources ainsi que pour construire et mettre à disposition une documentation claire sur la signification des données, leur mode d'élaboration et leur champ de validité. Le deuxième type de problème concernait la confidentialité des données individuelles contenues dans les fichiers. Le problème ne se posait pas tant que les informations n'étaient traitées que par des statisticiens tenus par le secret professionnel. Son émergence a suscité un cadre législatif et entrainé la création de la Cnil, Commission Nationale Informatique et Liberté.

C'est à cette époque que l'Insee a mis en chantier de grandes bases de données. Parmi elles la BDCOM, banque de données communales, qui avait vocation à regrouper un grand nombre d'informations statistiques sur les communes, intégrée quelques années plus tard dans le projet SEDDL, système d'étude et de diffusion des données locales.

Dans le même temps les utilisateurs de l'information, statisticiens ou chercheurs, s'emparaient de l'outil informatique et se dotaient de logiciels adaptés à leurs métiers. Et puis ce fut l'accès direct aux ordinateurs « en temps réel » et puis... Aujourd'hui chacun dispose sur son bureau d'un ordinateur dont la puissance est supérieure à celle de l'ordinateur avec lequel on traitait la totalité du recensement de la population à la fin des années 60. L'accès à internet s'est généralisé et

l'on peut accéder gratuitement en ligne à une quantité de données géographiques dépassant sans doute la capacité d'assimilation d'un individu normal.

Les utilisateurs des bases de données géographiques

Bien que ces données soient accessibles à tous, il semble nécessaire de disposer d'un certain bagage intellectuel pour les utiliser. Il est donc naturel que les « clients » de ces bases de données géographiques soient des chercheurs ou appartiennent à des organismes d'étude travaillant pour des collectivités locales. Mais l'apparition des bases de données géographiques et le développement d'outils adaptés à leur traitement et à leur valorisation -notamment cartographique- s'est accompagné de l'émergence d'un nouveau courant doté d'un grand dynamisme, le géomarketing. Ces géomarketeurs et leurs sociétés se sont montrés beaucoup plus acharnés que les chercheurs géographes et autres étudieurs pour obtenir des données géographiques de qualité et leurs pressions -parfois même les procès qu'ils ont intenté à l'Insee- ont sans doute joué un rôle important dans l'élaboration de l' actuelle offre en matière de données géographiques.

L'avenir des données géographiques

L'évolution technique ne s'est pas arrêtée aux ordinateurs individuels et à internet. Les Iphones et autres smartphones sont maintenant dans toutes les mains et se promènent dans toutes les rues. Ils permettent de recevoir des informations, de consulter des bases de données, mais aussi de fournir des informations, soit activement par l'envoi volontaire d'un message, soit passivement en permettant la géolocalisation de leur porteur. S'ouvre avec ces appareils tout un monde nouveau d'élaboration d'informations géolocalisées. Des opérateurs mondiaux - en particulier Google avec Google Maps et Google Earth – mettent à disposition de tous les référentiels et les outils pour rendre visible au monde entier ces informations géolocalisées.

Si l'on veut que ces informations puissent être mobilisées pour la connaissance publique et citoyennes, il serait temps de s'y mettre. Les utilisateurs commerciaux, eux, sont très actifs pour exploiter ces nouvelles mines d'informations. Tom-Tom vient de publier un palmarès de la circulation dans les villes de France. Orange travaille avec les sociétés d'autoroute pour mesurer la fluidité du trafic...

encadré

La géolocalisation des données

La géolocalisation des données a fortement évolué sous la double poussée de l'évolution technique et de l'émergence des problèmes urbains. Le recensement de la population fournit un bon exemple de cette évolution. Au moment de la collecte, l'adresse précise du logement et des personnes qui l'habitent est évidemment connue et relevée puisqu'il faut s'assurer que l'on recense tout le monde et sans double compte. Cette information[2] a toujours été couverte en France par le secret statistique pour la protection du citoyen. Elle restait à usage interne de contrôle de la collecte et n'était pas saisie dans les fichiers destinés au traitement statistique. Seules étaient produites et publiées des informations statistiques au niveau géographique de la commune. Nombreuses étaient cependant les directions régionales de l'Insee qui complétaient ces informations communales par des tableaux donnant la population des hameaux et des écarts. Le besoin d'informations infra-communales dans les villes a conduit à établir des statistiques au niveau des îlots, correspondant en gros aux pâtés de maisons. On a pour cela mis en place dans chaque direction régionale de l'Insee une équipe de cartographie[3] pour établir -en général sur la base des plans cadastraux- la cartographie des îlots. Cette cartographie était utilisée pour l'organisation de la collecte et la répartition des secteurs entre les agents recenseurs. Elle était introduite dans les fichiers de

2 Notons que si l'adresse était connue, elle n'en était pas pour autant géolocalisée en (x,y). Ce n'est que dans la période récente que l'on dispose d'outils permettant de géolocaliser un lieu à partir de son adresse.

3 L'histoire de cette cartographie et des relations entre l'Insee, le cadastre et l'IGN mériterait un long chapitre

traitement et servait de base à l'établissement de statistiques par îlots. Ultérieurement la Cnil, inquiète – notamment à cause de la montée en puissance du géomarketing - de l'usage pouvant être fait de ces données par îlot pour « profiler » les territoires, recommanda de ne plus publier de statistiques sur une si petite maille territoriale. L'Insee élabora alors un regroupement d'îlots en IRIS, compromis entre une attente d'informations à un niveau géographique fin et une exigence de protection des individus.

Un nouveau pas a été franchi avec l'introduction des nouvelles modalités de recensement. Pour les communes de moins de 10 000 habitants, rien n'est changé ou presque : le recensement est effectué de façon exhaustive tous les 5 ans, un roulement étant établi pour enquêter à tour de rôle un cinquième des communes chaque année. Pour l'instant et bien que les techniques existent, on n'a pas souhaité géolocaliser l'information à la collecte. Dans les communes de plus de 10 000 habitants, l'enquête de recensement est réalisée chaque année sur un échantillon tournant établi de façon à fournir, au bout de 5 ans, une information représentative au niveau de la commune et des îlots. Cet échantillon tournant est établi sur la base d'un répertoire d'immeubles dans lequel tous les immeubles de la commune sont géolocalisés. D'un point de vue géographique, l'avancée est que la géolocalisation est maintenant directement associée à l'information élémentaire ; le recul est que la collecte n'est plus exhaustive et donc l'information moins précise à un niveau géographique fin. Cette nouvelle donne de la géolocalisation n'est pour l'instant pas entièrement valorisée compte tenu des recommandations de la Cnil. Elle a cependant permis d'établir – ce qui était réclamé depuis longtemps par les géographes- des statistiques sur des mailles carrées d'un kilomètre de côté. Cette information carroyée[4] ne porte pour l'instant que sur la population mais devrait ultérieurement être enrichie.

Pour ce qui est de l'information statistique hors recensement, la principale avancée découle de la possibilité désormais accessible de géolocaliser les adresses. Toute information basée sur un répertoire comprenant les adresses – ce qui est le cas de Sirene pour les entreprises et les établissements- est donc potentiellement géolocalisable. Pour les enquêtes « ménages », la technique consistant à doter le matériel utilisé pour la collecte assistée par ordinateur (CAPI) d'une puce GPS permettant de localiser le lieu de l'enquête, n'est pas utilisée à l'Insee. Au chapitre des nouveautés on notera que pour l'enquête Transports réalisée en 2005, un échantillon de personnes se sont vu confier pendant quelques jours un GPS destiné à enregistrer tous leurs déplacements[5].

De façon générale on peut affirmer que dorénavant - compte tenu des avancées technologiques - les limites à la géolocalisation des données ne sont plus techniques mais sociétales. La France, très en pointe dans le domaine de la protection des informations individuelles, se retrouve par voie de conséquence, en plus mauvaise position en ce qui concerne l'accessibilité à une information finement localisée.

---

4 On trouvera sur le site de l'Insee, outre les tableaux statistiques et la carte
http://www.insee.fr/fr/ppp/bases-de-donnees/donnees-detaillees/duicq/accueil.asp?page=donnees_carroyees.htm
l'explication détaillée sur la méthode utilisée pour carroyer l'information
http://www.insee.fr/fr/ppp/bases-de-donnees/donnees-detaillees/duicq/accueil.asp?page=doc/donnees_carroyees_doc.htm
Ce travail étant réalisé par l'Insee dans le cadre d'une concertation européenne, on trouve à la fin de cette documentation un renvoi sur le site du Forum Européen pour la GéoStatistique (EFGS) : http://www.efgs.info.

5 On trouvera un point plus complet sur l'usage des nouvelles technologies pour le recueil d'informations sur les flux de personnes dans **Flux et afflux de touristes : les instruments de mesure, la géomathématique des flux** Christophe Terrier, article paru dans la revue Flux - 2005 et accessible en ligne sur
http://www.christophe-terrier.fr/ct2-textes/05terrier47-62.pdf