Association de Prospective Rhénane

Pôle Européen de Gestion et
*d'Economie ; **APR, bureau 202***
61, avenue de la Forêt Noire
67000 STRASBOURG (France)
Tel : 00.33 3 90 24 21 62.
E-Mail : contact@apr-strasbourg.org

Contribution à la communication au sein de l'APR

Stammtisch du mercredi 4 novembre 2009 sur le thème de la périurbanisation :

A propos des données statistiques disponibles

M. Bernard AUBRY

Essai de création d'une base communale

La diffusion des données publiques, et notamment celles issues des recensements de la population passe maintenant par internet. Pour des raisons évidentes, la nouvelle procédure présente des avantages indéniables. Mais la disparition des publications classiques sur support papier - fascicules bleu, jaune, orange et vert - trouble les utilisateurs habitués à trouver des données cohérentes d'un recensement à l'autre et plus encore ceux qui n'ont eu encore l'opportunité d'entrer dans le monde numérique. De leur côté, les meilleurs surfeurs, s'ils n'auront aucune difficulté à télécharger les données, en auront assurément à mettre en forme et à traiter les milliers de chiffres qui leur seront proposés. Ils ne pourront réellement en tirer parti que s'ils possèdent déjà une certaine pratique du traitement des données démographiques et sociales. C'est que la multiplicité des variables qui apparaissent effraie, et la complexité des libellés qui les désignent ne donnent pas aux fichiers l'abord convivial que l'on souhaiterait. Il arrive aussi très souvent que les chiffres pourtant censés caractériser la même entité à une même date sont différents quand ils ne sont pas contradictoires. Le coût d'entrée dans les fichiers des recensements est manifestement conséquent.

C'est précisément l'objet de ce travail que de proposer un ensemble de données de base très faciles à utiliser sous la forme de tableaux simples qui libèrent l'utilisateur des affres évoquées plus haut. Chacun des fichiers réunit pour chacune des quelque 900 communes d'Alsace un ensemble de variables. En tout, ce sont environ 200 données communales. L'information contenue dans ces fichiers est certes limitée eu égard à l'extrême richesse potentielle que représentent les silos de données que possède l'Insee, mais elle permet déjà de donner à tout un chacun, professionnel ou non, une première vision des phénomènes étudiés. En l'occurrence, pour aborder l'étude de la périurbanisation qui est l'un des thèmes actuels sur lesquels investit l'APR, on pourra présenter rapidement une série d'indicateurs de cadrage. Les fichiers ainsi construits représentent en quelque sorte un sésame pour accéder à de plus grandes bases : chacun pourra greffer aux données communales retenues celles qui seront susceptibles d'alimenter la recherche thématique engagée. Ce pourront être des bases issues des recensements de la population comme de toute enquête statistique ayant la commune comme unité statistique.

Ce que l'on trouve sur le site de l'Insee :

En simplifiant, on dispose de trois catégories de fichiers :

- Des fichiers présentant des **chiffres-clés** préétablis pour chaque commune ou pour des zonages spécifiques (canton, aire urbaine…). L'intérêt de ces fichiers est qu'ils sont bien adaptés à la consultation. Les tableaux présentés sont assortis de graphiques. Pour obtenir un chiffre donné et le comparer au même en 1999, c'est parfait.

- Des **fichiers thématiques communaux** en format Excel (en règle générale, une ligne par commune). Ces fichiers fournissent pour chaque localité des milliers de variables. Ils sont destinés à faire l'objet de traitements statistiques passant par des regroupements et des calculs d'indicateurs adéquats. En effet, ces fichiers sont formés pour l'essentiel de très petits nombres portant souvent sur de très petites communes. Ils n'ont alors individuellement aucun intérêt.
  Avantage de ces fichiers : les informations sont plus nombreuses que celles fournies naguère sur papier (fascicules verts). Inconvénients : les informations proposées sont insuffisantes aux attentes des chargés d'étude et a fortiori des chercheurs. En effet les fichiers sont verrouillés : l'utilisateur n'a pas la possibilité d'opérer des croisements de variables qu'il souhaiterait.
  Deux familles de fichiers de ce type rendent possible les évolutions temporelles. D'une part un ensemble de six fichiers comprennent des chiffres portant sur les deux derniers recensements (RP1999 et RP2006). C'est en puisant dans ces fichiers que nous avons constitué l'essentiel de la base communale proposée. A priori on ne se posera pas trop de questions sur la cohérence temporelle des données de cette base, puisque l'on a tout lieu de faire confiance aux statisticiens qui ont opéré les choix. D'autre part, un ensemble de huit fichiers issus de ce que l'Insee appelle maintenant le fichier harmonisé des recensements (ex fichier Saphir[1]). Les variables, encore peu nombreuses, permettent de suivre des évolutions portant sur la période 1968-1999. Nous avons chaque fois que c'est possible prolongé la série jusqu'en 2006 de façon à disposer d'un ensemble d'indicateurs cohérents couvrant une période de quatre décennies.

- Des **fichiers-détails**, c'est-à-dire des fichiers pour lesquels chaque ligne (chaque enregistrement) correspond à un individu recensé.
  Avantage majeur :
  il est possible a priori d'effectuer tous les croisements souhaités des variables de la base.
  Inconvénients :
  1) la lourdeur. Ces fichiers (550000 lignes pour le seul RP 2006 relatif à l'Alsace) ne sont pas destinés à l'utilisateur lambda. En effet, on ne peut télécharger à chaque fois une seule région, mais un ensemble de régions couvrant environ 20% du territoire national. Encore ne dispose-t-on que de la population résidant en Alsace. Si l'on veut connaître en détail la population travaillant en Alsace, comme c'est souvent le cas, il faut télécharger l'ensemble des cinq énormes fichiers, puisqu'il faut bien pouvoir repérer toutes les personnes qui résident dans les autres régions et qui viennent travailler dans la nôtre. Même chose, mais avec des conséquences plus grandes encore, si l'on veut étudier en détail les migrations définitives ou les migrations naissance-résidence.
  2) Les contraintes réglementaires. Actuellement le niveau géographique de diffusion le plus fin est la région (à l'intérieur d'une région la seule distinction géographique est le rural/urbain).

[1] *1 – La création du fichier Saphir (à la DR de l'Insee-Alsace) a bénéficié de conventions avec le Ministère de l'équipement. Une exploitation spécifique du fichier sur le thème de la périurbanisation a été réalisée en étroite relation avec la DAU (Direction de l'architecture et de l'Urbanisme) et a conduit à la publication d'un Atlas statistique des villes nouvelles d'Ile-de-France (mai 1995).*

Dans le cadre de notre travail de construction d'un fichier communal pour l'étude des villes et leur environnement, les fichiers détails ne peuvent donc servir qu'à des fins de contrôle (totalisations au niveau régional).
En l'état actuel, il apparaît donc que ces fichiers détails sont plutôt réservés à certains types d'étude (comparaisons interrégionales par exemple) et que leur exploitation est plutôt l'affaire des professionnels.

Remarques sur la diffusion des données publiques.

Après avoir été longtemps proposée comme un bien marchand, la statistique publique est maintenant considérée comme *res nullius*. Elle appartient au citoyen et c'est dans cette perspective que les instituts nationaux de statistique doivent s'efforcer de mettre à disposition du plus grand nombre les chiffres collectés sur fonds public. Ils doivent le faire sous la forme la plus conviviale possible. Soulignons à ce propos que le principe n° 15 du Code des bonnes pratiques de la statistique européenne exige l'**accessibilité et la clarté.** Il y a encore beaucoup de chemin à parcourir avant que toutes les composantes de la société (des partenaires sociaux aux citoyens de base) ne puissent accéder « confortablement » aux données qui leur permettraient de mieux comprendre leur environnement et de contribuer à élever le niveau du débat social. Un effort majeur de pédagogie devrait donc être engagé en un moment où il devient de plus en plus évident que les chiffres sont en voie de décrédibilisation (cf Claude Thélot, quand il évoque l' « *affaiblissement* » de la statistique publique.

On comprend bien que s'agissant des données dites sensibles, l'Insee ne soit pas autorisé à faire figurer des informations fines à un niveau géographique élémentaire. En revanche on regrette qu'il ne soit pas encore en mesure de proposer un système informatique permettant à l'utilisateur d'obtenir via internet des résultats portant sur les dites données sensibles pour des zones géographiques d'une taille suffisamment grande, ce qui écarterait toute critique d'ordre réglementaire. Cela se fait déjà couramment dans certains pays d'Europe du nord (les commandes en ligne peuvent d'ailleurs être parfois payantes). S'agissant du thème de la périurbanisation, on disposerait de résultats détaillés (sexe, âge, formation…) sur les étrangers et les immigrés non par commune, mais pour les différentes catégories territoriales utiles aux besoins de l'étude.

– – – – – – – – –

La base proposée

La base telle qu'elle est prévue comprend en fait un ensemble de fichiers Excel volontairement de taille réduite (les seules communes d'Alsace), mais qui regroupent l'essentiel des informations comparables pour les deux derniers recensements (1999 et 2006).

Les fichiers sont thématiques. Chaque fichier devrait être accompagné d'informations sur l'origine et la signification des variables. Les commentaires devraient être limitées à l'essentiel et être modifiés ou augmentés en fonction des suggestions faites par les utilisateurs. Il faudrait également associer à chaque fichier un ou plusieurs exemples d'applications possibles. Il va de soi que les fichiers pourront être facilement fusionnés entre eux puisqu'ils ont en commun une variable géographique unique, la commune.

Deux bases comprennent une ligne par commune :

Le fichier GEO : il décrit la commune du point de vue de son appartenance à des territoires (unité urbaine…).

Le fichier HIST : il comprend des données de population depuis 1954 ainsi que les naissances et les décès de chaque période intercensitaire. Ce fichier donne une perspective historique du solde naturel et du solde migratoire apparent. Il indique aussi le nombre des logements vacants et des résidences secondaires à chaque recensement.

Les autres fichiers comprennent deux lignes par commune (une ligne par recensement). Pour l'instant on propose trois fichiers :

1. Le fichier POP : il intègre des données sur l'âge, la formation.
2. Le fichier ACTEMP : il rassemble des informations sur la population active (comptée au lieu de résidence) et l'emploi (compté au lieu de travail).
3. Le fichier MEN : il comprend une sélection de variables sur les logements, les ménages et les familles.

En plus de ces données propres à 1999 et 2006, on crée un fichier unique comprenant une sélection de données issues du fichier harmonisé des recensements (ex fichiers Saphir) pour la seule période 1968-1999. Les changements de définition intervenus depuis l'entrée en vigueur du nouveau recensement font que les données de 1999 pour une même variable peuvent différer – parfois sensiblement – des données de 2006.

Remarques importantes

1) Il est inutile d'alourdir exagérément cette base. Cette base a vocation à fournir une série de données de référence qui permettent en soi d'aborder un nombre considérable de thèmes d'analyse. L'approfondissement de chacun de ces thèmes passera nécessairement par l'appel de données complémentaires. Soit qu'il s'agisse de données propres à l'utilisateur. Ainsi s'agissant de la périurbanisation, la base GEO sera enrichie par la description de zonages ad hoc, à savoir la composition communale quand elle existe des aires urbaines et des agglomérations à différentes dates anciennes. Soit qu'il s'agisse de bases téléchargeables issues ou non du recensement. Ainsi pour une meilleure approche de l'emploi on fera appel au fichier CLAP (Connaissance localisée de l'appareil productif). On pourra aussi télécharger des informations sur l'équipement des communes, sur l'occupation de l'espace, etc.

2) Les utilisateurs des données du nouveau recensement seront peut-être perturbés lorsqu'ils constateront que les résultats sont donnés avec des décimales (en très grand nombre !). D'aucuns ironisaient naguère quand les statisticiens leur apprenaient que le nombre d'enfants par femme était par exemple de 1,82. Les mêmes utilisateurs seront maintenant bien obligés d'accepter que le nombre d'actifs occupés est de 224,36587… dans la commune de X et le nombre de résidences principales y est de 195,4256... C'est que les méthodes de recensement ont évolué et, depuis 2004, la collecte se fait selon des procédures différentes dans les communes de moins et dans celles de plus de 10000 habitants. Dans les plus petites communes, toutes interrogées à raison de 20% d'entre elles chaque année, la collecte est quasi analogue à ce qu'elle était auparavant. Dans les plus grandes, la collecte se fait par tirage au sort à raison de 8,5% des logements chaque année. Des redressements complexes sont donc nécessaires pour reconstituer la population et leurs caractéristiques au début de l'année 2006, sachant que les collectes ont eu lieu dans les premières semaines des années 2004, 2005, 2006, 2007 et 2008.

3) Tirer parti des données des recensements n'est donc pas toujours aisé. Nombreuses sont les variables pour lesquelles l'on diffuse simultanément deux chiffres à la même date. Les écarts, qui ne sont pas toujours négligeables, sont dus au fait que l'exploitation

statistique des informations collectées se fait en deux étapes successives. La première étape correspond à ce qu'on appelle l'exploitation principale (on parlait naguère d'exploitation exhaustive). La seconde correspond à l'exploitation complémentaire (on disait exploitation par sondage). Dans le premier cas la transcription des informations se fait directement, dans le second cas elle se fait par application d'un système de traitement (système expert) qui calcule la modalité à affecter en fonction des informations lues sur le bulletin du recensement. Ainsi, la catégorie socioprofessionnelle, le secteur d'activité, l'appartenance à une famille ne sont connus que dans le cadre de l'exploitation complémentaire. Prenons un exemple. Les personnes interrogées indiquent le nom et l'adresse de l'établissement dans lequel ils travaillent. S'ils indiquent sur leur bulletin une information sommaire comme le fait de travailler à Peugeot-Mulhouse, ils auront dans l'exploitation principale Mulhouse pour lieu de travail (image de l'information fournie sur le bulletin). En revanche dans l'exploitation complémentaire qui se fait par appel du fichier Sirene (fichiers des entreprises et des établissements) ils seront affectés dans la commune de Sausheim. Les divergences sur le lieu de travail ainsi observées peuvent être très importantes. Elles compliquent considérablement la tâche des chargés d'étude qui s'arrachent parfois les cheveux à tenter d'y voir clair.

Ces divergences semblent moins fréquentes au RP 2006 que dans les recensements précédents. D'une façon générale, on est conduit à privilégier les chiffres issus des exploitations complémentaires. Il existe cependant des contre-exemples (pour les salariés des établissements de l'armée notamment).

1. Parmi les raisons qui rendent délicates les comparaisons historiques, on en retient au moins deux. L'une concerne la mesure de l'âge. La date de référence n'est plus la date du recensement (début du mois de mars), mais l'un des premiers jours de janvier. Par conséquent, en 2006, un enfant de 1 an est – approximativement - un enfant né au cours de l'année 2004 (âge révolu atteint au début de l'année 2006). En 1999, un enfant de 1 an était né en 1998 (calcul de l'âge par différence de millésime). Conséquence importante : la génération des enfants de 0 an était incomplète en 1999 (environ 2 mois sur 12) alors qu'elle est complète en 2006. Et quand on caractérise les générations quinquennales comme c'est courant de le faire, le groupe des 00-04 ans n'est pas du tout comparable de 1999 à 2006 (il ne l'est pas non plus pour les groupes d'âge plus élevés, mais les écarts sont alors moins importants). Conséquence : on sera très vigilant en cas de calcul de l'âge moyen ou de l'âge médian !

   L'autre différence porte sur la population active. La définition retenue en 2006 est plus extensive puisqu'elle intègre certaines personnes (étudiants, inactifs, retraités) qui ne l'étaient pas au recensement précédent.

2. Les études urbaines ne peuvent être approfondies sont si le niveau géographique le plus fin du territoire bâti est la commune. Dans le passé, le seul découpage disponible pour l'analyse statistique était le canton urbain. Depuis 1990, l'information est diffusée au niveau de l'iris 2000. Une base dans ce découpage sera téléchargeable de façon imminente.

La statistique, science de l'incertain

La statistique que l'on disait déjà science de l'incertain le devient plus encore aujourd'hui. Jusqu'à présent les chiffres des recensements avaient la propriété d'être des nombres entiers.

L'un des avantages, et non des moindres, était que l'on pouvait se raccrocher à quelques chiffres solides et bien mémorisés, des chiffres servant de référence. Certes le caractère entier des chiffres donnait l'illusion de la précision. Rien ne dit que les chiffres du nouveau recensement sont moins fiables que ceux des précédents. Il est probable qu'ils seront plutôt de meilleure qualité, même si l'on ne dispose pas encore de tout le recul nécessaire pour en juger. Les méthodes changent. Il faut s'y faire et travailler autrement.

A l'avenir, le travail du statisticien ne perdra en rien de son intérêt, bien au contraire.

Bernard AUBRY

