

Alsace : Périurbanisation

Stammtisch scientifique du 4 nov. 2009
Les données existantes.

Mlle. Julie VERGOZ
julie.vergoz@apr-strasbourg.org

Présents :

- Serge AUVRAY, INSEE ;
- Gérard BELLITTO, Cercle Gutenberg ;
- Jean-Marc BIRY, CAUE 67 ;
- Rémy GREBIL, SGARE ;
- Mayette GREMILLET, INSEE ;
- François GUYON, APR ;
- Stéphane HEIM, Université de Strasbourg ;
- René KAHN, Université de Strasbourg ;
- Frédéric KOSMAN ;
- Jean-Paul MASQUIDA, APR ;
- Moïse MAYO, INSEE ;
- Henri NONN, APR ;
- Jiyeon PARK, Université de Strasbourg ;
- Jacques POSTIC, INSEE ;
- Renaud SADARNAC, CUS ;
- Jérôme THIEN, Université de Strasbourg ;
- Laurence VATON, Consultante indép. développement durable ;
- Julie VERGOZ, APR ;
- Jean-Paul VILLETTE, Université de Strasbourg ;
- Philippe WALTER, Conseil Général 68 ;
- Raymond WOESSNER, APR.

Après une courte présentation du projet par Raymond Woessner et un appel à participation pour les prochains stammtsichs scientifiques, Bernard Aubry a présenté son travail. Celui-ci consiste à mettre à disposition une base de données qu'il a établie sous forme de plusieurs fichiers afin de permettre à tous de réaliser les statistiques de son choix. Il s'agira d'une base de données de 200 variables concernant les communes alsaciennes. Ces variables seront simples, transparentes et pourront évoluer en fonction des besoins de chacun. Elles sont explicitées dans le document ci-joint. M. Aubry a mis en avant le problème de qualité et de cohérence de certaines données. Celles-ci seront explicitement listées afin de mettre en garde l'utilisateur.

Il a été mis en avant la volonté de pouvoir insérer des dates charnières dans les statistiques. De même, M. Masquida a mis en avant la nécessité de définir un modèle pour le périurbain et donc de définir des critères pouvant aboutir à une hiérarchie des communes.

L'idée d'un partenariat avec la DIACT fut soulevée car cet organisme travaille actuellement sur la définition des villes mondes et villes européennes. Comme d'après M. Woessner une métropole se construit avec son cadre périurbain, il a été décidé de contacter la DIACT.

Le prochain stammtsich scientifique fut présenté : il s'agira de l'approche sociologique du périurbain. Il aura lieu le 2 décembre.

Rédaction :
Julie VERGOZ
5 novembre 2009

Alsace : Périurbanisation

Stammtisch scientifique du 4 nov. 2009
Essai de création d'une base communale,
par Bernard Aubry

Contact :
Mlle. Julie VERGOZ
[julie.vergoz@apr-
strasbourg.org](mailto:julie.vergoz@apr-strasbourg.org)

(version du 23 novembre 2009)

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'accès aux 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 les 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 donner 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 :

<http://www.recensement.insee.fr/accesDonneesTelechargeables.action>

En simplifiant, on dispose de trois catégories de fichiers :

- Des chiffres-clés préétablis pour chaque commune ou pour des zonages spécifiques (canton, aire urbaine...) :

<http://www.recensement.insee.fr/accesChiffresCles.action>

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 tableaux détaillés par thème :

<http://www.recensement.insee.fr/basesTableauxDetailles.action>

Il s'agit de fichiers 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 ceux qui souhaitent effectuer des 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 minuscules localités. A priori, chaque chiffre n'a en soi 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 les croisements de variables qu'il souhaiterait.

Parmi ces fichiers, deux d'entre eux permettent de suivre les évolutions temporelles : 1) pour les deux derniers recensements (RP1999 et RP2006), un ensemble de 8 tableaux. C'est en puisant dans ces fichiers que nous avons constitué l'essentiel de la base communale proposée. 2) un nombre limité de données issues de ce que l'Insee appelle maintenant le fichier harmonisé des recensements (ex fichier Saphir). Les variables retenues permettent de suivre les tendances lourdes qui se sont produites au cours de la période 1968-1999. Nous avons prolongé, quand c'était possible, la série jusqu'en 2006 de façon disposer d'un indicateur couvrant une période de quatre décennies.

- Des fichiers-détails : <http://www.recensement.insee.fr/fichiersDetail.action>

En l'occurrence, chaque ligne du fichier, (donc chaque enregistrement) correspond à un individu ou à un logement recensé. S'agissant du RP2006, on a accès actuellement à plusieurs fichiers. L'un concerne les logements (le niveau géographique le plus fin est la commune), deux autres l'ensemble des individus (l'un avec la région - distinction rural/urbain -, l'autre avec le canton comme niveau le plus fin). Enfin, un ensemble de 5 fichiers correspondent aux déplacements (mobilité et migrations).

Avantage majeur : il est possible a priori d'effectuer tous les croisements souhaités des variables présentes dans le fichier.

Inconvénients : 1) la lourdeur. Il faut d'abord télécharger un fichier (txt) qui comprend selon les cas la métropole tout entière ou un ensemble de régions (chaque fichier représente alors environ 20% du territoire). Ensuite il faut isoler la région d'intérêt. Chaque fichier est volumineux (550000

enregistrements pour l'Alsace pour le seul fichier individus), ce qui rend son exploitation peu accessible à l'utilisateur lambda. 2) Les contraintes réglementaires. On ne peut disposer à la fois d'une finesse dans les variables structurelles et géographiques, c'est pourquoi le fournisseur ne propose pas un fichier unique, comme cela plairait à l'utilisateur, mais un ensemble de fichiers. Il faut donc apprendre à s'adapter aux contraintes liées à cette multiplicité des bases. Le fichier individuel régional rend possible des analyses extrêmement fines en matière d'emploi (profession, activité détaillées) et par origine (nationalité détaillée), cependant que le fichier cantonal ne permet de réaliser que des exploitations beaucoup plus sommaires. Ni l'un ni l'autre ne sont d'un grand secours pour l'analyse de la périurbanisation. En revanche à partir des fichiers de mobilité intégrant la commune, il est possible d'effectuer des croisements de diverses variables : outre le sexe, l'âge, etc., on dispose notamment de la PCS en 6 postes et de l'activité économique en 5 postes, ce qui autorise déjà des études relativement élaborées dans le domaine de l'analyse locale.

Dans le cadre de notre travail de construction d'un fichier communal pour l'étude des villes et de leur environnement, les fichiers détails servent surtout de référence, notamment pour contrôler les totalisations. En l'état actuel, il apparaît donc que ces fichiers détails qui imposent l'usage de logiciels puissants sont plutôt réservés à des études approfondies.

Remarque 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 la mettre à disposition du plus grand nombre 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 donc 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 permettraient de mieux comprendre la société. Un effort majeur de pédagogie devrait être engagé en un moment où de plus en plus il devient évident que les chiffres sont décrédibilisés (cf. Claude Thélot, quand il évoque l'« *affaiblissement* » de la statistique publique¹).

Quant à l'accès aux informations élémentaires, donc soumises à des règles sévères de diffusion, il n'est possible de façon efficace que si chacun a la possibilité de composer lui-même des commandes en ligne sous forme d'interrogations directes des plus gros fichiers de données gisant dans les réserves des instituts nationaux de statistique, le système fixant lui-même en sortie la sélection des données diffusables. Cela se fait déjà couramment dans certains pays d'Europe du nord. Les commandes en ligne sont parfois payantes.

¹ Extrait de l'introduction de l'ouvrage *Echantillonnage et méthodes d'enquêtes*, ouvrage collectif sous la direction de Pascal Ardilly, coll. Sciences Sup, Paris Dunod, 400 p. (2004).

Reste le cas particulier des chercheurs. D'ores et déjà ceux-ci, dans des conditions contractuelles, peuvent être amenés à se brancher directement sur les fichiers détails de l'Insee et à les traiter jusqu'aux limites de leur possibilité informative.

La base proposée

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

Les fichiers sont thématiques. Chaque base est décrite et fait l'objet d'une ou de plusieurs applications possibles, parmi de nombreuses autres. Il va de soi que les fichiers peuvent être facilement fusionnés puisqu'ils ont en commun une variable géographique unique, la commune.

Deux bases comprennent une ligne par communes :

- 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 de logements vacants et des résidences secondaires.

Les autres fichiers comprennent deux lignes par commune (une ligne pour chacun des deux derniers recensements). Pour l'instant on dispose de trois fichiers :

- Le fichier POP : il intègre des données sur l'âge, la formation.
- Le fichier ACTEMP : il rassemble des informations sur la population active (comptée au lieu de résidence) et l'emploi (compté au lieu de travail).
- 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 dispose d'une sélection de données issues du fichier harmonisé des recensements (ex fichiers Saphir) pour la seule période 1968-1999 (HARM). Les changements de définition intervenus depuis l'entrée en vigueur du nouveau recensement font que le point 1999 est dédoublé puisque, pour une même variable, les chiffres peuvent différer parfois sensiblement selon qu'ils correspondent à la définition en cours en 1999 et en 2006.

Remarques importantes

- 1) **Une base modeste.** Celle-ci a pour vocation de fournir une série de données de référence qui permettent d'aborder un nombre important de thèmes d'analyse. L'approfondissement de chacun de ces thèmes passera

nécessairement par l'appel de données complémentaires issues soit du site, soit des données propres à l'utilisateur. Ainsi la base GEO sera alors enrichie par la description de zonages ad hoc. Pour les études sur la périurbanisation on ajoutera la composition communale des aires urbaines et des agglomérations à différentes dates. On pourra aussi télécharger une sélection de chiffres issus d'autres bases : du recensement certes, mais aussi du fichier CLAP (Connaissance localisée de l'appareil productif), du fichier de l'équipement des communes, etc.

- 2) **Décimales à foison.** Les utilisateurs des données du nouveau recensement seront peut-être étonnés de constater que les résultats sont donnés avec des décimales (en grand nombre !). D'aucuns ironisaient naguère quand on leur apprenait que le nombre d'enfants par femme était tantôt de 1,82, tantôt de 2,05. Les mêmes devront accepter que, dans la commune X, le nombre d'actifs occupés est de 224,36587... et le nombre de résidences principales de 195,4256... En effet les méthodes de recensement ont évolué depuis 2004. La procédure est différente selon que la commune a moins de 10000 ou plus de 10000 habitants. Les plus petites communes sont interrogées de manière exhaustive selon un cycle de 5 ans. En revanche pour les plus grandes communes, la collecte se fait sur une base aléatoire d'adresses sachant que chaque année 8,5% des logements sont interrogés. De cette façon, 14% de la population est observée chaque année, soit 70% au terme du cycle de 5 ans.

Des redressements complexes sont donc nécessaires pour reconstituer la population et leurs caractéristiques au début de l'année 2006.

- 3) **De nouvelles définitions.** Tirer parti des données des recensements n'est donc pas toujours aisé, surtout pour les non initiés. La société change et les définitions des objets d'investigation évoluent dans le temps, ce qui rend alors plus difficile l'interprétation des données. La nouvelle procédure de recensement a conduit non seulement à changer les méthodes de collecte sur le terrain, mais encore elle a aussi introduit des changements notables dans les définitions. Deux de ces changements peuvent troubler particulièrement les utilisateurs des données. L'un porte sur la population active. La définition retenue en 2006 est plus extensive que dans le passé puisqu'elle intègre un plus grand nombre de personnes ayant une activité professionnelle même partielle (étudiants, inactifs, retraités).

L'autre concerne la mesure de l'âge. Dans le passé, l'âge d'une personne était calculé par différence de millésime : en 1999 un enfant de 1 an était considéré comme un enfant né en 1998. De ce fait la première génération (0 à moins d'un an) était incomplète puisqu'elle était formée des seuls enfants nés entre 1^{er} janvier et la date du recensement (début mars). Dorénavant la première génération (0<1an) est complète puisqu'elle comprend aussi les enfants qui n'ont pas atteint la date de leur premier anniversaire à la date du recensement. Autrement dit, dans le passé, il s'agit d'un âge défini par différence de millésime (année de recensement mois année de naissance), alors qu'à présent il s'agit d'un âge révolu. Deux conséquences importantes :

- 1) le suivi des âges quinquennaux est fortement perturbé, surtout la tranche des 0-4 ans,

- 2) on sera très vigilant à l'examen des pyramides des âges et en cas de calcul de l'âge moyen ou de l'âge médian notamment.
- 4) **Exploitation principale ou complémentaire : des données différentes pour un même élément?** Le problème n'est pas nouveau. Les utilisateurs sont souvent perturbés quand ils constatent que pour une même variable, définie de la même façon, à une même date, pour un même lieu, les chiffres qui la caractérisent sont différents, parfois très différents. Ces divergences proviennent du fait qu'il y a deux procédures successives d'exploitation des données du recensement. Certaines variables, celles relatives à la famille, à la catégorie socioprofessionnelle, le secteur d'activité, ne sont connues que par ce qu'on appelle l'exploitation complémentaire (par opposition à l'exploitation principale). Il s'agit d'un traitement spécifique des bulletins. Prenons un exemple emblématique sur la localisation du lieu de travail. Les personnes interrogées indiquent le nom et l'adresse de l'établissement dans lequel ils travaillent. S'ils indiquent seulement Peugeot-Mulhouse, ils seront classés avec pour lieu de travail Mulhouse dans l'exploitation principale (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. Ces divergences, a priori moins fréquentes au RP 2006 que dans les recensements précédents, conduisent à privilégier les chiffres issus des exploitations complémentaires. Il existe cependant des contre-exemples (pour les salariés des établissements de l'armée notamment), ce qui contribue d'autant plus à accentuer le brouillage informationnel.
- 5) **Iris 2000.** Quand c'est possible, on traitera des informations au niveau infra-communal (Iris2000). Le recul historique sera moindre que pour les données communales (introduction du zonage en 1990) et la batterie des variables disponibles sera plus limitée. Le programme de diffusion du RP2006 prévoit une mise en ligne prochaine d'un jeu de données infra communales.

La statistique, science de l'incertain

Jusqu'à présent les chiffres des recensements avaient la caractéristique d'être des nombres entiers. S'ils donnaient l'illusion de la précision, ils avaient du moins le mérite d'être, du moins pour certains d'entre eux, mémorisables, et de servir de référence pour les vérifications. Maintenant, selon la façon dont les données sont totalisées, les résultats seront différents. Il faut s'y faire et apprendre à utiliser les chiffres pour ce qu'ils peuvent apporter. Plus que jamais la statistique devient donc la science de l'incertain : l'essentiel est de savoir déterminer des indicateurs significatifs, adaptés au problème étudié.

Bientôt tout un chacun pourra accéder aux plus gros fichiers et devenir en quelque sorte un statisticien en herbe. Le travail du praticien ne s'en trouvera pas pour autant dévalué : il ne perdra en rien de son intérêt, bien au contraire. Le statisticien professionnel devra mettre ses compétences au service de la valorisation de celle des autres, moins avertis des subtilités de la discipline.

Une large diffusion de la statistique publique est donc assurément une très bonne chose, pour autant que ceux à qui elle est destinée sauront en faire le meilleur usage. Les produits mis à disposition sont-ils prêts à l'emploi ? l'utilisateur dispose-t-il d'un service d'assistance adéquat ? On a des raisons d'être sceptique. Une certaine émulation des institutions publiques devrait cependant permettre l'amélioration des services rendus, puisque le citoyen peut comparer la quantité et la qualité des données mises en ligne par son propre pays à celles proposées au-delà des frontières par les administrations homologues.