Le portail de données européennes sur le recensement : Census Hub
Description et précautions d'utilisation

I. Description et contexte

En 2014, l'ensemble des États membres de l'Union Européenne, ainsi que l'Islande, le Liechtenstein, la Norvège et la Suisse ont répondu au premier règlement européen concernant les recensements de la population et du logement.

Ce règlement, signé en 2008, vise à harmoniser les données de recensement fournies par les États membres. Il précise les informations demandées, pour l'année de constat 2011.

Les résultats sont disponibles, via un portail unique d'interrogation des données, appelé Census Hub.

Le portail offre de nombreuses informations, en trois volets :

- les données relatives au recensement (Census data) ;
- les données relatives à la qualité (Data on quality) ;
- les métadonnées descriptives (Metadata).

1. Les données relatives au recensement :

Les données disponibles sont des tableaux multi-dimensionnels, dont le contenu a été pré-défini par Eurostat.

Au total, 60 tableaux de données sont disponibles, portant sur les individus, les ménages (ou logements) et les familles.
La majorité des données sont de niveau « régional » (NUTS2, nomenclature des unités territoriales statistiques).
Quelques tableaux sont disponibles à un niveau géographique plus fin : « départemental » (NUTS3) ou « municipal » (LAU2, Local Administrative Units).

De nombreuses informations sont disponibles : elles portent sur les caractéristiques socio-démographiques des individus, leur liens familiaux, les caractéristiques des logements…

Les thèmes disponibles sont les suivants :

- pour les individus :
  - Age
  - Sexe
  - Zone géographique (pays, région, département, ou commune)
  - Taille de la localité
  - Lieu de résidence habituelle un an avant le recensement
  - Situation au regard de l'activité du moment
  - Secteur (branche d'activité économique)
  - Profession
  - Situation dans la profession
  - Emplacement du lieu de travail
  - Niveau d'instruction (niveau le plus élevé atteint)
  - Position dans la famille
  - Position dans le ménage
  - Situation matrimoniale légale
  - Pays de citoyenneté
  - Pays/lieu de naissance
  - Année d'immigration dans le pays depuis 1980
  - Année d'immigration dans le pays depuis 2000

- pour les ménages et les logements :
  - Taille du ménage privé
  - Type de ménage privé
  - Installations permettant de se laver
  - Norme de densité (nombre de pièces par occupant du logement)
  - Conditions de logement
  - Nombre d'occupants
  - Nombre de pièces
  - Régime d'occupation des logements classiques
  - Régime de propriété
  - Logements par époque de construction
  - Type de local d'habitation
  - Logements par type de bâtiment
  - Type de chauffage
  - Lieux d'aisances
  - Modalités de jouissance du logement par le ménage
  - Système d'adduction d'eau

- Pour les familles :
  - Taille du noyau familial
  - Type de noyau familial

2. Les données relatives à la qualité et les métadonnées :

Les métadonnées et les données sur la qualité sont très utiles pour comprendre la façon dont sont construites les données, et évaluer la comparabilité entre les pays.
Elles permettent d'évaluer la pertinence des informations, en mesurant la représentativité de l'information délivrée (sur-couverture, sous-couverture, part d'information provenant directement collectée…), ou sa variance (coefficients de variation associés à l'aléa de sondage).

Les métadonnées décrivent en particulier le processus de collecte des informations, le traitement des données, les éventuelles divergences avec les concepts demandés par Eurostat ou les causes d'imprécision...

II. Précautions d'emploi - Différences avec la diffusion du recensement sur insee.fr

La source de données Census Hub peut être utilisée pour des comparaisons européennes.
L'harmonisation des concepts utilisés a nécessité des travaux spécifiques : les données disponibles sur le Census Hub peuvent différer de celles présentes sur insee.fr
Les données insee.fr sont les données qui font foi, pour les populations légales, ou toute autre analyse statistique sur les données françaises uniquement.

Les précautions signalées ci-après ne concernent que les données françaises du Census Hub. Chaque État membre de l'Union européenne a signalé des précautions d'utilisation spécifiques dans les métadonnées et les données relatives à la qualité présentées sur le Census Hub. Il convient de s'y référer pour toute exploitation des données du Census Hub.

1. Utilisation des données 2010 et pondérations spécifiques 2011

Afin de respecter les échéances européennes et d'assurer la qualité des données, l'Insee a utilisé les informations issues du recensement 2010. Ces données ont ensuite été harmonisées avec les populations légales 2011.
Ainsi, les données disponibles sur le Census Hub relatives à la France ne sont pas rigoureusement identiques à celles présentes sur insee.fr, mais elles sont comparables au niveau de la structure de la population, et les ordres de grandeur sont ceux de 2011, ce qui permet des comparaisons européennes pertinentes.

2. Nomenclatures différentes (profession, niveau d'instruction)

Par ailleurs, les données du Census Hub diffèrent de celles du recensement diffusé sur insee.fr pour certaines variables.
Les informations relatives à la profession et au niveau d'instruction sont diffusées en nomenclatures européennes, celles-ci diffèrent de celles habituellement utilisées sur insee.fr.
Pour la profession, c'est la nomenclature ISCO ou CITP (classification internationale type des professions, 10 postes) qui est utilisée, et pour le niveau d'instruction c'est la nomenclature ISCED ou CITE (classification internationale type de l'éducation, 6 postes).
L'utilisation de ces nomenclatures a nécessité des estimations spécifiques, qui peuvent rendre les résultats fragiles à un niveau fin (mentionnés par la lettre U dans le Census Hub).

Les modélisations effectuées ont utilisé des données issues de l'enquête Emploi, qui ont permis de s'assurer de la qualité des données au niveau national.

3. Modélisations (période d'achèvement, résidence antérieure)

D'autres variables ont nécessité des modélisations, pour s'adapter aux catégories demandées par Eurostat. Il s'agit de la résidence antérieure (pour les individus) et de la période d'achèvement de la construction (pour les logements).

L'indicateur de résidence antérieure conserve les trois modalités qui étaient proposées avant 2011, mais la question « Où habitiez-vous il y a cinq ans ? » est devenue « Où habitiez-vous il y a un an ? ». Faute d'information plus précise sur la mobilité, l'imputation s'appuie sur le principe simplificateur suivant : il n'y a pas plus d'une mobilité au cours de cinq années consécutives. Ainsi, les personnes qui résidaient dans le logement il y a cinq ans sont déclarées résidentes dans ce même logement il y a un an. Dans le cas où aucun individu du logement recensé ne se trouvait dans le logement il y a cinq ans, on convient de s'appuyer sur l'année d'emménagement figurant dans la feuille de logement (question 8). Si une personne ne se trouve dans aucune de ces deux situations, l'imputation s'effectue de manière aléatoire en considérant que l'emménagement a pu avoir lieu à n'importe quelle date située entre la date de collecte du recensement et cinq années auparavant. À quelques aménagements près liés à des considérations de calendrier de collecte, la probabilité d'être résident dans le logement recensé un an avant la collecte est alors voisine de 80 % (quatre chances sur cinq). Un traitement spécifique complexe est appliqué aux enfants nés entre le 1er janvier *n-5* et le 1er janvier *n-1*, *n* désignant l'année de collecte : l'algorithme s'appuie sur la composition du ménage et sur les valeurs imputées au préalable aux adultes du ménage. L'indicateur reste non renseigné pour les personnes résidant en collectivité et pour les enfants nés l'année précédant la collecte.

La période d'achèvement de la construction concerne toutes les catégories de logements. L'exercice d'imputation consiste à gérer le changement de nomenclature en affectant à tout logement qui a été déclaré construit entre une année *a* et une année *b*, une année de construction tirée au hasard entre *a* et *b*. On en déduit immédiatement la tranche de construction du logement dans la nouvelle nomenclature. La procédure aléatoire affecte bien entendu la même tranche de construction à tous les logements d'un même immeuble. Une difficulté apparaît pour les logements construits avant 1949, en l'absence de borne inférieure de tranche dans la nomenclature initiale. On utilise alors les données de l'enquête Logement 2006 pour positionner le logement par rapport à l'année 1919, en passant par une étape intermédiaire d'imputation aléatoire qui affecte le logement à une tranche propre à l'enquête Logement. En effet, cette dernière distingue les logements construits avant 1871, ceux construits de 1871 à 1914 et ceux construits de 1915 à 1948. L'affectation du logement à l'une de ces trois tranches repose sur des statistiques de structure du parc de logements en tenant compte d'une typologie de communes composée de trois groupes de communes. Le cas des DOM a été traité selon la même méthodologie, mais en utilisant les données de l'enquête Logement 2006 propre aux DOM.

4. Concepts différents (activité)

Enfin, le règlement européen sur les recensements de la population requérait la fourniture de variables d'activité au sens du Bureau international du travail (BIT). Pour produire de telles variables à partir des données du recensement dont le questionnaire n'est pas construit en vue de cette finalité, un traitement économétrique spécifique a dû être développé.

L'activité au sens du BIT attribue à toute personne de 14 ans ou plus une des trois modalités suivantes : actif occupé, chômeur, inactif. La procédure d'imputation s'appuie fondamentalement sur les données de l'enquête Emploi, qui est la seule enquête permettant d'obtenir une activité individuelle conforme à la définition du BIT. L'idée générale consiste à modéliser l'activité à l'aide d'un ensemble de variables bien corrélées à l'activité et que l'on peut collecter, d'une part à l'occasion du recensement et d'autre part auprès des individus de l'échantillon de l'enquête Emploi. Ces variables, dites explicatives, permettent de mettre en place un système de prédiction de l'activité, individu par individu. La procédure la plus technique concerne les individus de moins de 75 ans vivant en ménage ordinaire. Elle est composée de trois étapes, que l'on détaille ci-dessous.

La première étape consiste à sélectionner les variables significativement explicatives de l'activité au sens du BIT. Ces variables doivent être présentes à la fois dans le recensement et dans le questionnaire de l'enquête « Emploi ». Les techniques statistiques permettent de vérifier que la variable disponible la plus explicative de l'activité au sens du BIT est la déclaration spontanée d'activité. Celle-ci traduit le ressenti exprimé par l'enquêté en matière d'activité : dans le bulletin individuel du recensement, c'est par nature le sens de la question 10 « *Quelle est votre situation principale ?* » posée à tout individu recensé de 14 ans et plus. C'est aussi une question que l'on trouve formulée, de manière quasi identique, à la fin du questionnaire de l'enquête Emploi. Néanmoins, parce que les conditions de collecte sont très différentes entre les deux enquêtes (recensement d'une part, enquête Emploi d'autre part), il a fallu apporter des aménagements à la déclaration spontanée d'activité collectée dans l'enquête Emploi, de façon à la rendre homogène à l'information du recensement. Au-delà de cette variable, il s'avère que le sexe, la catégorie d'âge et le niveau de diplôme permettent d'améliorer la prédiction de l'activité au sens du BIT : ces informations seront mobilisées dans l'étape suivante.

La seconde étape donne lieu à des calculs de probabilité. Puisque le sexe, l'âge et le diplôme sont des caractéristiques qui influencent l'activité, on effectue des croisements *ad hoc* des modalités de ces trois variables et on construit ainsi dix catégories de population en ne conservant que les croisements discriminants. Pour chacune de ces catégories, on estime à partir de l'échantillon Emploi les probabilités de passage de chacune des trois modalités de la déclaration d'activité spontanée vers chacune des trois modalités de l'activité au sens du BIT. *In fine*, on obtient un jeu de 90 probabilités de changement de situation, dites « de transition ».

La troisième étape modifie ces probabilités afin que l'on respecte une contrainte forte : en effet, il s'agit de faire en sorte qu'après imputation, on retrouve « à très peu de chose près » les effectifs de chômeurs, d'actifs et d'inactifs, région par région (la Corse étant regroupée avec la région PACA). Cette exigence relève d'une recherche de cohérence entre les statistiques diffusées, d'une part en exploitant le recensement, d'autre part en exploitant l'enquête Emploi. On y répond en cherchant à perturber au minimum les probabilités de transition issues de l'étape 2, tout en introduisant la dimension géographique parce que les nouvelles probabilités doivent désormais dépendre de la région. Un programme d'optimisation mathématique permet de trouver le jeu de probabilités de transition adéquat respectant les contraintes fixées, en distinguant exactement 2079 probabilités. *In fine*, on constate que les taux de chômage par région obtenus à partir de l'enquête Emploi se retrouvent à quelques centièmes de points de pourcentage près lorsqu'on exploite les données imputées du recensement.

L'imputation individuelle finale s'obtient en activant une loterie qui affecte au hasard l'une des trois modalités au sens du BIT, en respectant les probabilités de transition issues de l'étape 3 étant données la déclaration spontanée d'activité du recensement et les valeurs des trois autres variables explicatives retenues. Dans les DOM, la même méthodologie a été appliquée, mais à partir des données Emploi propres aux DOM. L'imputation concernant les personnes de 75 ans et plus à consister à les considérer toutes comme inactives. Enfin, faute d'information mieux adaptée, on a dû

se résoudre à imputer l'activité spontanée issue de la question 10 à tout individu résidant en communauté.

5. Précautions générales

L'utilisateur des données imputées, quelle que soit la variable et quelle que soit la méthode, doit impérativement garder en mémoire que **ces données imputées dépendent fortement du modèle retenu**. Il y a donc par construction une **erreur de modèle** qui affecte toutes les statistiques utilisant ces données. Cette erreur peut être substantielle dès qu'on s'intéresse à des sous-populations, surtout si elles sont spécifiques (exemple : les inactifs avec la variable de lieu de résidence antérieure). Il convient donc d'être très vigilant à l'occasion des exploitations statistiques de ces variables et si possible **de s'abstenir de produire des études fines de corrélation ou portant sur des populations petites ou particulières**.