Représentation cartographique des indicateurs socio-démographiques à partir des données carroyées à 200 mètres de l'INSEE

Sylvain Potier

10 avril 2015

Introduction

Ce document présente succintement la méthode utilisée au Département Observation et Analyse de l'Agence Régionale de Santé Pays de la Loire pour traiter les données brutes mises à disposition par l'INSEE, pour les représenter graphiquement et les exploiter.

Il ne se substitue pas aux éléments officiels sur la présentation et la description des données, dont les références sont disponibles à la fin de ce document.

1 Les données et les indicateurs

Les données carroyées à 200 mètres, fournies par l'INSEE, sur la France métropolitaine, représentent un ensemble d'indicateurs socio-démographiques. Ces données, au format alphanumérique, peuvent être représentées cartographiquement, soit sous forme vectorielle, soit sous forme rasterisée.

Pour le format vectoriel, un carroyage projeté dans le système de coordonnées européen INSPIRO-Compatible ETRS 1989 LAEA (code EPSG 3035) permet de représenter les données attributaires, à l'aide d'une jointure sur le champ "idINSPIRE".

Pour la forme rasterisée (format image .tif), l'identifiant "idINSPIRE" permet de déterminer les coordonnées du coin sud-ouest (bas-gauche) de chaque carreau, dans le système de coordonnées ETRS 1989 LAEA (code EPSG 3035).

Le traitement et l'affichage de l'ensemble des carreaux (plus de 2,2 millions) au format vectoriel dans un logiciel SIG sont grands consommateurs de mémoire vive et de temps. De ce fait, l'option retenue a été, pour chaque indicateur, de réaliser les traitements des valeurs brutes au préalable, puis de répartir ces valeurs en 5 classes, et enfin de les représenter au format raster, plus léger à manipuler.

2 Les traitements et la répartition des valeurs en 5 classes à l'aide de la méthode du k-means

2.1 Le calcul des indicateurs

Chaque carreau contient le nombre d'individus résidant dans le carreau. Des rectangles, contenant chacun plusieurs de ces carreaux, disposent des données brutes correspondant aux indicateurs socio-démographiques. Un identifiant commun, l'identifiant du rectangle ("IdK"), permet de déterminer à quel rectangle appartient chaque carreau.

Le mode opératoire défini par l'INSEE préconise, grâce à cet "identifiant du rectangle", de calculer, pour chaque indicateur, la valeur par carreau, grâce à la règle simple :

$$Indicateur_{Carreau} = \frac{Indicateur_{Rectangle} \times NbIndividus_{Carreau}}{NbIndividus_{Rectangle}}$$

Remarques :
- Les indicateurs ont été calculés puis arrondis à l'entier le plus proche
- Afin d'atténuer les écarts, le logarithme népérien a été utilisé pour déterminer la répartition en classes du nombre d'individus par carreau, à l'aide de la méthode du k-means
- Un indicateur présente des valeurs qui ne sont pas réparties en 5 classes comme pour les autres indicateurs : la somme des revenus fiscaux par Unité de Consommation (UC)

2.2 La répartition en 5 classes : méthode du k-means

Afin de disposer de la référence nationale et éviter les effets de frontières dûs à l'échantillonnage sur une ou plusieurs régions, les calculs ont été réalisés sur l'ensemble de la France métropolitaine. Le logiciel libre R a été utilisé, et la méthode du k-means a été préférée à la méthode de Jenks (Seuils naturels), pour une question technique de temps de calcul.

Le paramétrage de la méthode du k-means pour répartir les valeurs en cinq classes sous R a été réalisé comme suit :
- Cette méthode permet de déterminer les centres des cinq classes
- Paramètre « nstart = 10» : 10 ensembles aléatoires sont réalisés pour déterminer les centres de classes (valeur à mettre en adéquation avec la taille de l'effectif et les temps de calcul)
- La borne inférieure, non-incluse, de la première classe est considérée à zéro
- La borne supérieure de la dernière classe est égale à la valeur maximale de la série étudiée
- Les bornes supérieures des 4 premières classes sont déterminées, grâce à la formule :

$$Bornesup_{Classe1} = \frac{Centre_{Classe1} + Centre_{Classe2}}{2}$$

3 Les données produites et leur exploitation

3.1 La table des valeurs calculées

Pour chaque indicateur, une table .dbf est créée, contenant, pour chaque carreau, la valeur de l'indicateur, le centre de la classe et la valeur de sa borne supérieure. Cette table .dbf permet donc de représenter cartographiquement l'indicateur grâce à la couche des carreaux au format vectoriel.

3.2 Création des rasters et projection des données à la volée

La création a été réalisée à l'aide de la fonction "Raster" sous le logiciel libre R, en tenant compte de l'étendue des données. Les rasters créés contiennent donc une information géographique dans le système ETRS 1989 LAEA (code EPSG 3035), permettant d'obtenir une image orthonormée, compatible et reprojetée à la volée lors de la superposition avec des données en RGF93 (EPSG 2154).

Pour les carreaux ne contenant pas de valeur (cours d'eau, zones non-habitées), la valeur -32768 leur est attribuée avant la création du raster. En effet, lors de la création du raster, cette valeur est alors considérée comme "Nulle" et traitée comme vide.

3.3 Les fichiers mis à disposition

Les fichiers représentant chaque indicateur sont nommés selon le schéma suivant :
Indicateur_Info-complémentaire.extension

Le répertoire «Originaux_INSEE» contient les données téléchargées depuis le site de l'INSEE et les couches de carreaux au format vectoriel. Le répertoire «CaroNat» contient les données calculées pour leur représentation aux formats vectoriel ou raster.

Cinq fichiers illustrent chaque indicateur :

- **Indicateur.dbf** : table attributaire, permettant de visualiser l'ensemble des données calculées concernant cet indicateur sous forme tabulaire. Elle peut être jointe à la couche vectorielle des carreaux.
- **Indicateur_val.tif** : raster représentant, pour chaque carreau, la valeur calculée réelle de l'indicateur illustré
- **Indicateur_bs.tif** : raster représentant, pour chaque carreau, la répartition en 5 classes déterminée par la méthode du k-means. Chaque carreau contient la borne supérieure de la classe à laquelle il appartient, permettant d'indiquer ces valeurs limites dans une légende associée.
- **Indicateur_bs.tif.vat.dbf** : table attributaire liée au raster Indicateur_bs.tif, créée à l'aide d'ArcView, afin d'aider à la représentation graphique des données. Cette table contient les 5 bornes supérieures déterminées par la méthode du k-means.
- **Indicateur_val.tif.vat.dbf** : table attributaire liée au raster Indicateur_val.tif, créée à l'aide d'ArcView, afin d'aider à la représentation graphique des données. Cette table contient les valeurs uniques calculées pour cet indicateur.

Les fichiers Caro.QGS et Caro.MXD permettent de visualiser les rasters à l'aide des logiciels Quantum GIS 2.6 et ArcView 10.0 respectivement.

3.4 Les fichiers représentant les indicateurs ont les noms suivants :

- **1indMen** : Nombre total de ménages d'une personne
- **5indMen** : Nombre total de ménages de 5 personnes et plus
- **BasrMen** : Nombre total de ménages dont le revenu fiscal par unité de consommation est en-dessous du seuil de bas revenu
- **CollMen** : Nombre total de ménages en logement collectif
- **IndAge1** : Nombre total d'individus âgés de 0 à 3 ans
- **IndAge2** : Nombre total d'individus âgés de 4 à 5 ans
- **IndAge3** : Nombre total d'individus âgés de 6 à 10 ans
- **IndAge4** : Nombre total d'individus âgés de 11 à 14 ans
- **IndAge5** : Nombre total d'individus âgés de 15 à 17 ans
- **IndAge6** : Nombre total d'individus âgés de 25 ans et +
- **IndAge7** : Nombre total d'individus âgés de 65 ans et +
- **IndAge8** : Nombre total d'individus âgés de 75 ans et +
- **NbIndiv** : Nombre d'individus résidant dans le carreau
- **NbMen** : Nombre de ménages résidant dans le carreau
- **Occ5Men** : Nombre total de ménages présents depuis 5 ans ou plus dans leur logement actuel
- **PropMen** : Nombre total de ménages propriétaires
- **SurfMen** : Surface cumulée des résidences principales, en mètres carrés
- **IndSrf** : Somme des revenus fiscaux par unité de consommation winsorisés des individus

Références

- La page INSEE dédiée à ces données
- La documentation complète sur les données carroyées
- La documentation synthétique
- La méthode k-means à l'aide du logiciel R