Analyse spatiale et territoriale de données de recensement

Formation Carthageo-Geoprisme 2021 / 1ere journée

C.GRASLAND

15/10/2020

Section 1

Données RPLS

Source

Le répertoire des logements locatifs des bailleurs sociaux (RPLS) a pour objectif de dresser l'état global du parc de logements locatifs de ces bailleurs sociaux au 1er janvier d'une année. Il est alimenté par les informations transmises par les bailleurs sociaux. La transmission des informations pour la mise à jour annuelle du répertoire des logements locatifs est obligatoire. Les données sont ensuite géolocalisées à l'adresse et mis à disposition des utilisateurs sur le **site du ministère de la transition écologique**

Les fichiers sont disponibles en général par régions mais livrés par départements dans le cas de l'Ile de France. Nous allons utilisé ici le fichier du 1er janvier 2020 accessible à l'adresse suivante

https://www.statistiques.developpement-durable.gouv.fr/le-parc-locatif-social-au-1er-janvier-2020-0

Métadonnées

Le fichier de données brutes au format .csv est accompagné d'un document excel précisant le code des variables et la façon dont elles ont été obtenues.

B	C	D	E
Complément d'identification du bâtiment autre	**COMPLGEO**		
Lieu dit	**LIEUDIT**		
Code logement situé en Quartier prioritaire de la politique de la ville	**QPV**	1 = Logement en QPV 2 = Logement hors QPV	*Cette donnée est issue de la déclaration des bailleurs et non de la géolocalisation.*
Code type de construction	**TYPECONST**	I = Individuel C = Collectif E = Etudiant NC = Non conforme	**Non conforme :** valeur renseignée autre que I, E ou C
Code nombre de pièces	**NBPIECE**	1 = 1 pièce 2 = 2 pièces 3 = 3 pièces 4 = 4 pièces 5 = 5 pièces 6 = 6 pièces 7 = 7 pièces 8 = 8 pièces 9 = 9 pièces ou plus 0 = Non renseigné NC = Non conforme	**Non conforme :** valeur renseignée autre que 1 à 9
Surface habitable (m²) - pour calcul	**SURFHAB**	Exemple : 50 0 : Non renseigné Non conforme	• Pour les logements dont la surface habitable est renseignée: ***valeur surface*** **Non conforme :** lorsque la surface habitable est un nombre de plus de • Pour les logements dont la surface habitable est non renseignée : **0 : Non renseigné**
Année d'achèvement de la construction	**CONSTRUCT**	*Exemple : 1950* Non renseigné Non conforme	**Non conforme :** Pour les logements construits avant 1000
Année de première mise en location du logement	**LOCAT**	*Exemple : 2010* Non renseigné Non conforme	**Non conforme :** Pour les logements construits avant 1850
Année d'entrée du logement dans patrimoine locatif du bailleur	**PATRIMOINE**	*Exemple : 2010* Non renseigné Non conforme	**Non conforme :** Pour les logements construits avant 1850

Spatialisation

Le fichier indique pour chaque logement sa localisation précise en terme d'adresse mais aussi d'étage dans un immeuble. A partir de ces données qualitatives, l'INSEE a procédé à un géocodage qui aboutit à la création de deux champs :

- coordonnées de latitude et longitude non projetées
- coordonnées de position en projection Lambert officielle

Selon les analyses on peut utiliser l'une ou l'autre de ces coordonnées. Mais la meilleur solution consiste à **créer un fichier de type sf (spatial features)** en coordonnées WGS94 qu'on pourra ensuite reprojeter dans le système de son choix.

Stratégie

Avant toute exploitation du fichier il est fortement recommandé d'analyser en détail les métadonnées et de définir une stratégie d'analyse.

1. **choisir une première zone d'étude** de petite taille et localisée de préférence dans un espace que l'on connaît bien.
2. **choisir des variables intéressantes** dont l'on connaît bien la signification et dont on a analysé en détail les métadonnées
3. **vérifier la qualité des données** en regardant notamment le nombre de valeurs manquantes, le dégré de précision, etc.
4. **sélectionner des données auxiliaires** issues d'autres sources que l'on souhaite croiser avec celles du RPLS en s'assurant de leur compatibilité (espace, temps, définition, ...)
5. **Ajouter les coordonnées spatiales** et stocker le résultat dans un fichier de type sf comportant les indications de projection.

Importation du fichier .csv

Le fichier initial au format .csv est importé à l'aide de la fonction `fread()` du package `data.table` car elles est rapide et relativement robuste face aux erreurs de codage. Il faut tout de même préciser le type d'encodage avec `encoding ="UTF-8"` ainsi que le caractère décimal avec `dec=","`.

```
 library(data.table)
rpls <- fread("data2021/94/geoloc2020_detail_IDF_dep_94.csv",
              encoding = "UTF-8",
              dec=",")

rpls<-as.data.frame(rpls)

saveRDS(rpls,"data2021/94/RPLS2020.RDS")
```

Le résultat est converti en `data.frame()` puis stocké dans un fichier .RDS.

Examen du fichier .RDS

On importe le fichier enregistré au format RDS et on vérifie sa taille avec dim() et sont ype avec class()

```
don <- readRDS("data2021/94/RPLS2020.RDS")
dim(don)
```

```
## [1] 177902     73
```

```
class(don)
```

```
## [1] "data.frame"
```

Il s'agit bien d'un tableau de données classique de type data.frame. Il comporte 177902 lignes (chacune correspondant à un logement) et 73 variables (décrites dans les métadonnées).

Choix de la zone d'étude

On décide de limiter notre analyse dans un premier temps à quatre communes voisines présentant des profils différents. On peut tester leur profil de respect de la loi SRU en 2019 sur l'application suivante : https://www.ecologie.gouv.fr/sru/

- **Bonneuil-sur-Marne (94011)**: large excédent ($> 25\%$, pas de pénalité)
- **Chennevières-sur-Marne (94019)** : léger déficit (22.76%, 58 k€)
- **Sucy-en-Brie: déficit (94071)** : (19.93% , 150k€)
- **Ormesson-sur Marne (94055)** : très fort déficit (2.29%, 665 k€)

On relève leur code INSEE afin de pouvoir faciliter l'extraction des données.

Choix des variables

On va se limiter ici à un très petit nombre de variables

variables de localisation

- result_id : code de l'adresse
- result_label : label de l'adresse
- LIBCOM : nom de la commune
- DEPCOM : code de la commune
- latitude : coordonnées latitude
- longitude: coordonnée longitude
- X : coordonnée projetée (EPSG = 2154)
- Y : coordonnée projetée (EPSG = 2154)

variables thématiques

- CONSTRUCT : année de construction
- SURFHAB : surface habitable en m2
- NBPIECE : nombre de pièces

Extraction du fichier

On applique la double sélection des individus et des variables en nous servant des fonctions `filter()`et `select()`du package dplyr.on aboutit ici à un fichier de 8139 lignes et 11 variables.

```
sel <- don %>%
  filter(DEPCOM %in% c("94011","94019",
                       "94071","94055")) %>%
  select(result_id, result_label,
         DEPCOM, LIBCOM,
        latitude,longitude,X,Y,
        CONSTRUCT,SURFHAB,NBPIECE)

dim(sel)
```

```
## [1] 8139   11
```

Recodage et typage

Certaines variables doivent être recodées ou changées de type afin de faciliter leur exploitation ultérieure par R.

```
sel$DEPCOM <- as.character(sel$DEPCOM)
sel$LIBCOM <- as.factor(sel$LIBCOM)
sel$PLG_IRIS <- paste(sel$DEPCOM,sel$PLG_IRIS, sep = "")
sel$SURFHAB <- as.numeric(sel$SURFHAB)
```

Résumé rapide

On analyse rapidement les variables thématiques choisies

CONSTRUCT	SURFHAB	NBPIECE
Min. :1902	Min. : 13.00	Min. :1.000
1st Qu.:1966	1st Qu.: 55.00	1st Qu.:2.000
Median :1969	Median : 68.00	Median :3.000
Mean :1977	Mean : 64.58	Mean :3.176
3rd Qu.:1992	3rd Qu.: 77.00	3rd Qu.:4.000
Max. :2019	Max. :172.00	Max. :6.000

Sauvegarde du fichier

On sauvegarde le fichier obtenu au format .RDS afin de garder le formatage des variables :

```
saveRDS(sel,"data2021/94/sel_logt.RDS")
```