02/04/2021

Fractures numériques, sociales et spatiales dans la région Centre-Val de Loire : Rapport Méthodologique

Meyssa BEDDAR
Tom BLACHON
Matthieu SIMOES

Master 1 MéDAS – CNAM

INTRODUCTION :

Selon une étude de l'INSEE, la région Centre-Val de Loire est l'une des régions ayant connu l'une des plus fortes extensions de ses aires périurbaines entre la fin du XXème siècle et le début des années 2000. Demeurant toutefois relativement rural, il s'agit d'un territoire à la situation ambigüe, bordé par l'Ile de France au Nord-Est et par des régions bien moins urbanisées au Sud. Aussi, nous avons décidé de nous intéresser à cette zone géographique en potentielle transition.

Prenant comme point de départ les données de connexion internet dans cette région, nous avons ainsi souhaité étudier la répartition de cette dernière sur le territoire. Dans un second temps, nous avons entrepris de déterminer s'il existait une corrélation entre la qualité des connexions internet et la répartition sociale sur l'ensemble de ce territoire.

Pour mener à bien cette étude, nous nous sommes servis de neuf jeux de données, contenant les informations nécessaires à nos différentes problématiques :

- **Centre-Val-de-Loire-Base-Eligibilité (2021) :** Il s'agit du point central de notre étude. Cette base contient une ligne par foyer dans un immeubles. Ce jeu de données présente notamment des données sur la classe d'éligibilité à internet d'un foyer.

- **Centre-Val-de-Loire-Base-Immeuble (2021) :** Cette base complète la base de données précédente. Elle apporte de l'information sur les immeubles.

- **CONTOURS-IRIS (2020) :** Ce jeu de données au format shapefile regroupe les coordonnées géographiques de chaque IRIS de France.

- **INSEE_LOGEMENT_IRIS-COMMUNES (2015) :** Ce fichier contient de nombreuses informations sur les logements et ménages dans chaque IRIS de France.

- **INSEE_REVENUS_DISPONIBLES_IRIS (2016) :** De même que le fichier logements, cette base donnes des informations sur les revenus de la population au sein de chaque IRIS.

- **INSEE_REVENUS_DISPONIBLES_COMMUNES (2016) :** Identique au fichier précédent, à l'exception de quelques variables. Ce fichier regroupe les informations de revenus au niveau communal.

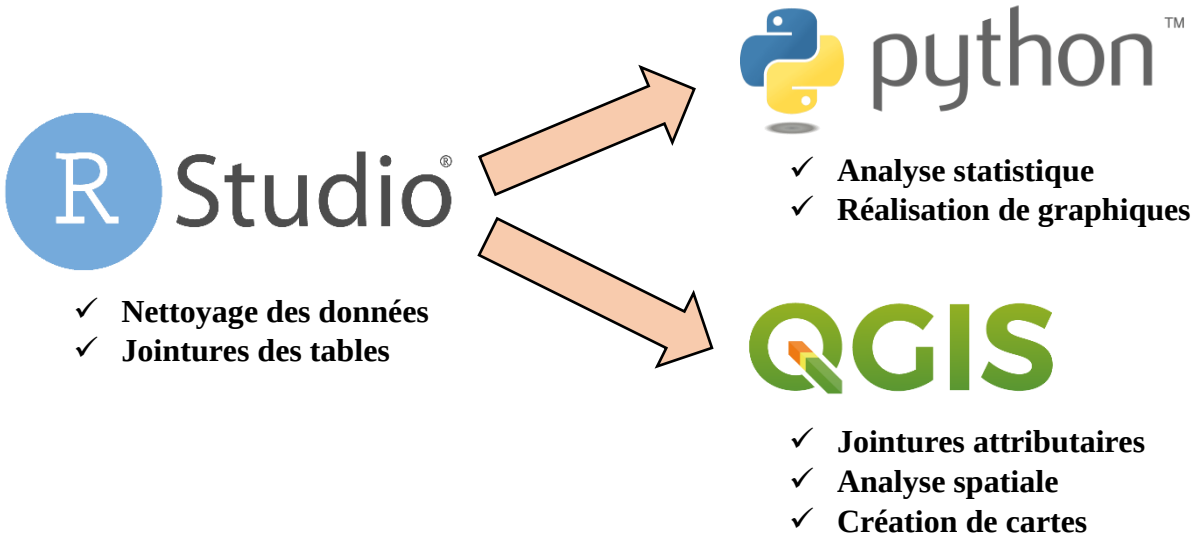
- **Niveau_Diplome_CVL (2017) :** Ce jeu de données présente le plus haut niveau de diplôme obtenu par les habitants de chaque commune du Centre-Val de Loire. Une ligne corresponde à une commune.

- **Tranche_Age_CVL (2017) :** Ce jeu de données présente la répartition des tranches d'âge des habitants de chaque commune du Centre-Val de Loire. Une ligne corresponde à une commune.

- **CSP_Emploi (2017) :** Ce jeu de données présente la répartition des catégories socio-professionnelles des habitants de chaque commune du Centre-Val de Loire. Une ligne corresponde à une commune.

Aussi, nous avons entrepris de réaliser cette étude en deux temps. En premier lieu, ayant une bonne connaissance de la manipulation de données avec le langage R, nous avons décidé de nettoyer et joindre regrouper nos fichiers avec le logiciel Rstudio. Une fois modifiés, nous importons alors nos fichiers « propres » sur l'outil QGIS afin de réaliser des cartes et de procéder à une analyse spatiale des données. Par ailleurs, nous avons procédé à une analyse statistique de nos données en langage Python, pour des raisons d'optimisation.

Dans ce rapport, nous allons donc présenter les étapes de notre approche méthodologique de façon plus détaillé. Ensuite, nous nous aborderons les difficultés que nous avons pu rencontrer durant ce projet, avant d'évoquer les limites et biais de notre étude.



I. DESCRIPTION METHODOLOGIQUE DU TRAVAIL MIS EN ŒUVRE

A. Sur Rstudio et Python

La première partie de notre traitement a été réalisé sur Rstudio, afin de procéder au nettoyage et à la jointure de nos fichiers.

En premier lieu, nous avons importé les données relatives à l'éligibilité et les immeubles en Centre Val de Loire, ainsi que notre fichier « **Contours IRIS** », contenant des données spatiales. Nous avons ainsi pu rattacher nos immeubles à la couches spatiales de nos IRIS. Pour cela, nous avons attribué un système de projection spatial « Lambert 93 » à nos immeubles, afin de pouvoir joindre cette table avec notre couche d'IRIS.

```
##################################################################Immeuble avec info IRIS
setwd("~/Perso/Ecole/Master/Carto/Projet/MeDAS_Centre_Val_de_Loire")
fichier_immeuble <- read_csv(file = "input/Centre-Val_de_Loire_Base_Immeuble_20210115/data/outputs/Centre-Val_de_Loire_Base_
fichier_iris <- st_read(dsn = "input/CONTOURS-IRIS_2-1__SHP__FRA_2020-01-01/CONTOURS-IRIS/1_DONNEES_LIVRAISON_2020-12-00282
                        as_tibble = TRUE, crs = 2154)

immeubles_avec_localisation <- fichier_immeuble %>%
  filter(!is.na(imb_longitude)) %>%
  filter(!is.na(imb_latitude))
# De 1 190 000 immeubles, on passe à 1 189 902, soit 98 lignes sans localisation

immeubles_spatial <- st_as_sf(immeubles_avec_localisation,
                              coords = c("imb_longitude", "imb_latitude"),
                              crs = 3857 # EPSG 3857, WGS84 / Pseudo-Mercator, indiqué dans la doc
)
immeubles_Lambert93 <- immeubles_spatial %>%
  st_transform(2154)

immeubles_iris <- immeubles_Lambert93 %>%
  st_join(fichier_iris)
```

Suite à cela, nous avons pu fusionner notre table des immeubles avec celle de l'éligibilité, grâce à la variable « imb_id », présente dans nos deux jeux de données. Dans cette nouvelle base, appelée « cvl », nous avons alors supprimé les colonnes dont les informations étaient en double après à la fusion. Puis, à l'aide de la fonction *as.factor()* de R, nous avons converti la majorité de nos variables en facteurs, afin de faciliter leur manipulation. Pour les besoins de notre étude, nous avons également créé une nouvelle variable binaire, indiquant les ménages qui possédaient où non un accès à la fibre optique dans leurs foyers.

Dans un second temps, nous avons importé le jeu de données contenant des informations sur les logements de chaque IRIS. Ce jeu possédant plus de 100 variables, nous avons réalisé un tri dans ces dernières afin de ne conserver que les colonnes les plus pertinentes pour notre étude. De même, nous avons supprimé toutes les données ne concernant pas le Centre-Val de Loire. Afin d'obtenir les informations à l'échelle communale et départementale, nous avons ainsi créé deux jeux de données supplémentaires, agrégeant les données des IRIS par communes et par départements.

```
##############################Selection des variables à utiliser et tranformation en var numériques
logement_iris=logement_iris %>% select(c("IRIS","REG","DEP","COM","LIBCOM","TRIRIS","TYP_IRIS","P15_RP","P15_RSECOCC","P15_LOGVAC","P15_MAISON","P15_APPART",
                                         "P15_RP_ACHTOT","P15_RP_ACH19","P15_RP_ACH45","P15_RP_ACH70","P15_RP_ACH90","P15_RP_ACH05","P15_RP_ACH12",
                                         "P15_MEN","P15_PMEN","P15_PMEN_ANEM0002","P15_PMEN_ANEM0204","P15_PMEN_ANEM0509","P15_PMEN_ANEM10P","P15_RP_PROP",
                                         "P15_RP_LOC","P15_NPER_RP","P15_ANEM_RP"))


logement_iris = logement_iris %>%
  filter(logement_iris$DEP ==45 | logement_iris$DEP ==18 | logement_iris$DEP ==28 | logement_iris$DEP ==37 | logement_iris$DEP == 36 | logement_iris$DEP ==41)


var=colnames(logement_iris[,8:29])
for(i in var){
  logement_iris[,i]=as.numeric(logement_iris[,i])
}

logement_comm=logement_iris %>% distinct(COM)
logement_comm$REG=logement_iris$REG[match(logement_comm$COM,logement_iris$COM)]
logement_comm$DEP=logement_iris$DEP[match(logement_comm$COM,logement_iris$COM)]
logement_comm$LIBCOM=logement_iris$LIBCOM[match(logement_comm$COM,logement_iris$COM)]

for(i in var){
  logement_comm[,i]=tapply(logement_iris[,i], logement_iris[,4],sum)
}

logement_dep=logement_iris %>% distinct(DEP)
logement_dep$REG=logement_iris$REG[match(logement_dep$DEP,logement_iris$DEP)]

for(i in var){
  logement_dep[,i]=tapply(logement_iris[,i], logement_iris[,3],sum)
}
```

Nous avons procédé à un traitement identique pour les données relatives aux revenus, dont nous disposions également.

Par la suite, nous avons importé dans notre environnement Rstudio nos trois jeux de données relatifs aux informations sociodémographiques sur la région Centre-Val de Loire : respectivement sur le plus haut niveau de diplôme obtenu, les catégories socioprofessionnelles et les tranches d'âge des habitants. Notons que ces bases de données ont fait l'objet d'un premier nettoyage, directement sur Excel afin de retirer les variables que nous ne souhaitions pas conserver. Par ailleurs, nous avons pivotés les variables restantes afin de n'obtenir qu'une ligne d'information par communes.

```r
age_dep=as.data.frame(unique(age$code_dep))
names(age_dep)[1]="DEP"
var_age=colnames(age[2:8])
for(i in var_age){
  age_dep[,i]=tapply(age[,i],age$code_dep,sum)
}
```

Une fois ces *dataframes* nettoyés et triés, nous avons pu procéder à la jointure de l'ensemble de nos fichiers, pour les trois échelles de nos études. Nous avons donc réalisé une première base de données, intitulée « cvl_iris », en fusionnant notre tableau « cvl » précédemment créé à nos données de logement, de revenu, et à nos informations sociodémographiques. De la même manière, nous avons créé un fichier « cvl_comm » et un fichier « cvl_dep », traitant les mêmes informations, mais aux échelles communales et départementales.

```r
############################## Fusion table CVL avec Revenu iris et Iris logement: Obtenir table d'info à l'iris
cvl_iris=left_join(cvl,logement_iris,by=c("CODE_IRIS"="IRIS"))
cvl_iris$taux_pauvre_60=revenu_iris$DISP_TP6016[match(cvl_iris$CODE_IRIS,revenu_iris$IRIS)]
cvl_iris$Premier_quart=revenu_iris$DISP_Q116[match(cvl_iris$CODE_IRIS,revenu_iris$IRIS)]
cvl_iris$Mediane=revenu_iris$DISP_MED16[match(cvl_iris$CODE_IRIS,revenu_iris$IRIS)]
cvl_iris$Trois_quart=revenu_iris$DISP_MED16[match(cvl_iris$CODE_IRIS,revenu_iris$IRIS)]
cvl_iris=left_join(cvl_iris,age,by=c("COM"="Code.géographique"))
cvl_iris=left_join(cvl_iris,emploi,by=c("COM"="Code.géographique"))
cvl_iris=left_join(cvl_iris,educ,by=c("COM"="code_geo"))

cvl_iris=cvl_iris %>% drop_na("P15_RP")

cvl_iris=cvl_iris %>% relocate("imb_id","REG","DEP","COM","LIBCOM","addr_code_insee","imb_code_insee","addr_nom_commu
cvl_iris=cvl_iris %>% relocate("code_techno_fibre",.after="code_techno")

cvl_iris=cvl_iris %>% select(-c("imb_code_insee","addr_code_insee","addr_nom_commune","NOM_COM","INSEE_COM","TYP_IRIS

head(cvl_iris)
summary(cvl_iris)
```

Cependant, malgré la suppression d'un certain nombre de valeurs manquantes, ces jeux de données demeuraient particulièrement volumineux. Nous avons donc créé un second jeu de données « allégé », pour chaque niveau d'étude (IRIS, communes et départements). Celui-ci ne donne pas les informations détaillées sur les ménages, mais son exportation, notamment pour la création de cartes avec QGIS, nous semblait bien plus optimisé.

Une fois nos jeux de données réalisés et nettoyés, nous avons procédé à leur analyse statistique. Pour des raisons d'optimisation que nous expliciterons plus tard, nous avons privilégié le langage Python pour réaliser cette analyse. À l'aide des librairies « Pandas », « Matplotlib » et « Seaborn » notamment, nous avons ainsi réalisé une série de graphiques venant appuyer et compléter l'analyse spatiale de nos données sur QGIS.

B. Sur QGIS

Une fois le traitement des données sur R achevé et les fichiers exportés, nous avons utilisé le Système d'Information Géographique QGIS pour procéder à une analyse géographique et spatiale de nos données.

Tout d'abord, nous avons commencé par importer notre couche « Concours IRIS » sur le logiciel, ainsi que nos trois jeux de données créés précédemment à l'échelle départementale (« cvl_dep »), communale (« cvl_comm ») et intercommunale (« cvl_iris »).

Puis, à partir de la couche de polygone « **Contours IRIS** », nous avons créé des fonds de carte représentant les IRIS, communes et départements du Centre-Val de Loire. Pour cela, nous nous

avons tout d'abord supprimé, via une requête attributaire sur le code régional, tous les IRIS ne se trouvant pas dans notre région.

Ensuite, nous avons ajouté à cette couche une colonne indiquant le code départemental de chaque IRIS. Nous avons alors fusionné les IRIS en utilisant le code communal et le code départemental, pour créer respectivement une couche des communes et une autre des départements.

Une fois nos fonds de cartes réalisés, nous avons pu y joindre, via la table d'attributs, les jeux de données réalisés sur R, regroupant toutes les informations nécessaires à nos analyses. Ici encore, nous avons pu réaliser les jointures à l'aide du code IRIS, du code communal et du code départemental.

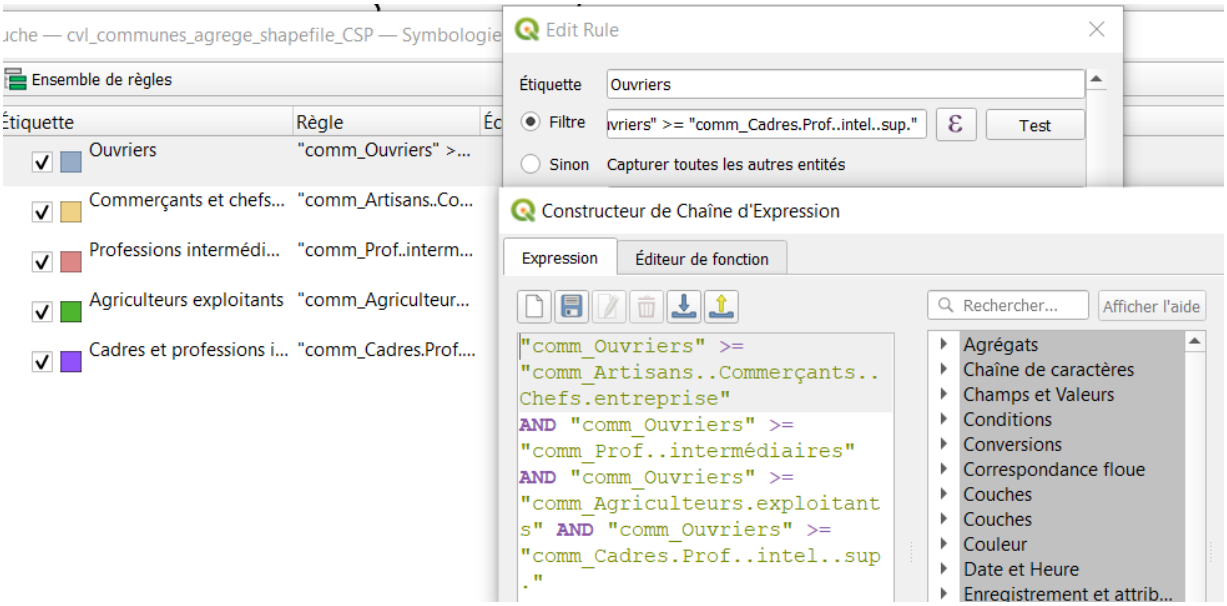
Comme nous avions calculé au préalable sur Rstudio la part de chaque type de connexion internet, nous avons tout d'abord réalisé deux cartes sur les taux de connexion par IRIS. La première représente spécifiquement le taux de connexion fibrée pour chaque IRIS. Pour cela, nous avons privilégié la création d'une graduation selon le pourcentage de connexions fibrées, par tranches de 5% et allant de 0 à 25%. La seconde met quant à elle en valeur le type de connexion majoritaire par IRIS et souligne les zones où la connexion est particulièrement bonne (forte présence de réseau fibré et très haut débit) ou particulièrement mauvaise.

Ensuite, afin d'avoir un aperçu de la densité de la population sur l'ensemble de la région, nous avons souhaité obtenir une carte du nombre de ménages par IRIS. Cette carte, mise en relation avec la précédente, pourrait ainsi nous indiquer si une corrélation est identifiable entre un regroupement important de ménages et une bonne qualité de connexion internet. Pour cela, nous avons à nouveau opté pour une graduation, par nombre de ménage cette fois-ci, allant de 0 à 3920 en fonction des IRIS.

Enfin, nous souhaitions mettre en relation la répartition du taux de connexion fibrée dans la région avec des critères sociodémographiques : le niveau d'éducation, l'âge et le secteur d'activité professionnelle des habitants. Pour cela, nous avons réalisé trois cartes, mettant respectivement en évidence, pour chaque commune :

- Le **plus haut niveau de diplôme** majoritairement obtenu au sein de la commune.
- La **catégorie socio-professionnelle** majoritaire au sein de la commune.
- La **tranche d'âge** la plus représentée au sein de la commune.

Comme nous disposions de plusieurs variables pour chacune de ces informations, nous ne pouvions pas opérer à une classification automatique. Aussi, nous avons dû créer trois nouvelles couches géographiques (une pour chaque critère) auxquelles nous avons attribué un ensemble de règles pour n'afficher systématiquement que la catégorie majoritaire.



Ensemble de règles permettant une hiérarchisation des catégories socio-professionnelles

Une fois ce travail effectué, nous avons simplement eu à afficher sur notre carte les zones où le taux de connexion fibrée était jugé important (entre 15% et 25% de connexion fibrée dans la zone).

Pour finir, nous avons utilisé l'outil de mise en page proposé par QGIS pour ajouter une légende, un titre, une source, une orientation et une échelle à nos cartes, avant de les exporter.

II. DIFFICULTÉS RENCONTRÉES

La première difficulté que nous avons rencontrée concerne la conversion de nos données en données spatiales. En effet, au début de notre étude, nous souhaitions joindre les données de notre fichier « **immeubles** » avec le fichier « **Contours IRIS** ». Pour cela, nous devions faire de notre fichier « **immeubles** » un fichier géographique, puis y joindre les IRIS à l'aide des coordonnées géographiques. Or, nous avons essayé dans un premier temps de réaliser cette opération sur QGIS, mais la jointure était mal effectuée et copiait chaque ligne de notre fichier 10 fois. Ainsi, d'un jeu de données d'un million de lignes, ce dernier en contenait désormais dix millions. Pour régler ce problème, nous avons changé notre angle d'approche en réalisant notre jointure directement sur R et l'opération s'est finalement bien passée.

D'autre part, le principal obstacle auquel nous avons dû faire face durant ce projet est sans doute la question du poids des fichiers. En effet, comme nous l'avons vu précédemment, nous avons fait le choix de regrouper toutes les données relatives à une même échelle entre elles. Or, de ce fait, chaque ligne de nos fichiers de données correspond au plus petit niveau d'étude, à savoir, un ménage au sein d'un IRIS. Le nombre de ménage étant très important, nos fichiers disposent alors de plus de 6 millions de lignes. De même, bien que nous ayons supprimés les variables inutiles à notre étude, nous disposons encore de 70 colonnes de données. Ainsi, après chargement de nos *dataframes* sur Rstudio, les temps pour réaliser des opérations étaient démultipliés et nous avons été confrontés à des problèmes de mémoire.

De ce fait, la conversion de ces fichiers au format shapefile (pour les transformer en couche vectorielle) était particulièrement long et le fichier extrait pesait près de 3Go. Un poids nous empêchant tout traitement sur QGIS. Pire encore, afin de joindre ces données à la couche de polygones de nos IRIS, le nombre de ligne de notre fichier ne nous permettait pas une jointure par table attributaire. Nous souhaitions donc effectuer une jointure spatiale pour palier à ce problème. Or, cela entraine à nouveau des temps de chargements extrêmement longs et une multiplication du poids de notre fichier.

Pour pallier à cette multiplication de problèmes liés à la taille de nos jeux de données, nous avons donc décidé de créer de nouveaux fichiers « résumés ». Ceux-ci regroupent les informations à des échelles bien plus raisonnables (IRIS, communes et départements). Par exemple, dans notre fichier résumé des IRIS, chaque ligne d'information correspond maintenant à un IRIS et non à un ménage. Concernant les données relatives aux ménages, nous avons ainsi fait les sommes des valeurs dont nous disposions au sein d'un IRIS. Sur QGIS, nous avons pu, grâce à cela, réaliser des jointures par tables attributaires, bien moins lourdes et plus efficaces que des jointures spatiales.

Concernant la réalisation des graphiques de notre analyse, nous souhaitions conserver nos jeux de données affinés, malgré leurs poids importants. Pour optimiser le traitement, nous avons ainsi transféré ces fichiers sur *JupyterLab* où nous avons réalisé les graphiques dans le langage Python. En effet, *JupyterLab* fonctionne via un serveur distant et exploite donc moins les ressources de nos machines, permettant une vitesse de calcul bien plus rapide.

III. LIMITES ET BIAIS DE L'ETUDE

Malgré la résolution des problèmes rencontrés, notre étude n'est cependant pas parfaite et comporte certaines limites et biais qu'il nous faut aborder.

Tout d'abord, notons que les cartes réalisées s'intéressent majoritairement à la présence de la fibre optique sur le territoire. Nous avons effectué ce choix volontairement car celui-ci nous semblait pertinent afin de mettre en lumière les zones disposant de la meilleure qualité de connexion possible. Cependant, ce choix présente des limites car il est possible que certaines zones soient occultées, bien que la connexion y reste convenable, via le câble coaxial ou le Très Haut Débit Radio (THDR) par exemple.

D'autre part, notons que nous n'avons pas été en mesure de réaliser des cartes sur les données de revenus, car le jeu de données contenait des informations manquantes. En effet, nous disposions de trop peu de données pour réaliser une étude sur l'ensemble du territoire. Nous voulions essayer de mettre en relation le niveau de richesse d'un IRIS ou d'une commune avec la qualité de sa connexion internet, mais cela ne nous a donc pas été possible.

Une autre limite notable de notre étude concerne les échelles analysées. En effet, nous avons choisi, pour des questions de qualité des données, de nous limiter à l'étude des échelles départementales, communales et des IRIS. Cependant, nous ne dégageons ainsi pas réellement les spécificités et les différences au sein-même des IRIS, bien qu'il puisse exister des variations significatives. Notre analyse permet donc d'avoir un aperçu général des grandes tendances, mais ne nous permet pas de nous pencher sur des niveaux de détail importants.

Enfin, soulignons que notre étude est partiellement biaisée, du fait des valeurs manquantes dans certains de nos jeux de données, à commencer par les informations concernant les revenus. Si certaines données sont manquantes, alors la pertinence et la représentativité de nos analyses s'en trouvent nécessairement impactées.

CONCLUSION :

Pour conclure, il est vrai que notre méthodologie est perfectible et certaines optimisations auraient pu nous permettre un gain de temps considérable. En effet, la partie la plus chronophage de ce projet fut sans nul doute le nettoyage et la fusion de nos jeux de données sur Rstudio. La question du poids trop important de certains fichiers, notamment, ne s'est posée que tardivement, au moment de l'exportation. Cela nous a contraint à modifier de nombreuses lignes de code déjà écrites, afin de gérer cet obstacle. Si nous avions pris en compte cette éventualité au préalable, nous aurions ainsi pu éviter une charge de travail inutile.

De même, nous avons pris un certain nombre de décisions, concernant les données utilisées et notre analyse ne se veut pas exhaustive dans son traitement. Une vision plus claire, dès de départ, des variables que nous allions exploiter nous aurais également permis un nettoyage plus approfondi et un allègement de la taille de nos fichiers.

Toutefois, si nous avons rencontré, au cours de notre analyse, un certain nombre de problèmes que nous n'avions pas anticipés, nous sommes toujours parvenus à les résoudre. Une répartition efficace des taches au sein de notre équipe, ainsi qu'une bonne communication, nous ont permises d'exploiter au mieux les compétences de chacun et de surmonter les obstacles rencontrés. Aussi, lors de futurs projets de la sorte, nous pourrons plus aisément optimiser notre approche méthodologique grâce à l'expérience acquise.