Projet Data Quality
Emploi-Manche

Cursus DPM
Session du 24 janvier-2023

Apprenant : Cédric Couche

Table des matières

Contexte

Ce projet est réalisé dans le cadre de la formation Data Product Manager, dispensée par DataScientest.

Problématique

L'offre d'emploi des entreprises auprès des bénéficiaires du RSA dans la Manche est-elle compatible avec la recherche & demande de travail des BRSA. Sous quelles conditions peut-il y avoir adéquation?

Source des données

Pour réaliser cette étude, plusieurs sources de données ont étés utilisées

Information de référence

Ces informations agrégées (Nombre d'habitant, nombre de communes, …) ont principal intérêt de valider la justesse des données

- Article sur le département de la Manche | wikipedia :
https://fr.wikipedia.org/wiki/Manche_(d%C3%A9partement)
- Dossier sur le departement de la Manche | INSEE :
https://www.insee.fr/fr/statistiques/2011101?geo=DEP-50
- Infographie RSA | Departement de la Manche :
https://www.manche.fr/espace-presse/rsa-les-donnees-cles-dans-la-manche-au-30-09-2021/

Données Géographiques

- [s06] : INSEE : jeu de données sur communes en France avec géolocalisation
https://www.data.gouv.fr/fr/datasets/contours-des-communes-de-france-simplifie-avec-regions-et-departement-doutre-mer-rapproches/

Sources consultées mais non retenues :
- [s01] Base officielle des codes postaux | DataNOVA (La Poste) :
https://datanova.laposte.fr/explore/dataset/laposte_hexasmal/table/?disjunctive.code_commune_insee&disjunctive.nom_de_la_commune&disjunctive.code_postal&disjunctive.ligne_5
- [s07] codes communes officiels | INSEE: https://www.insee.fr/fr/information/6051727
- [s08] "Intercommunalités au 1er janvier 2021 - compatibles avec le fond communal" :https://www.data.gouv.fr/fr/datasets/contours-des-communes-de-france-simplifie-avec-regions-et-departement-doutre-mer-rapproches/

Données Démographiques

- [s04] Population en 2019 - IRIS - France hors Mayotte | INSEE :
https://www.insee.fr/fr/statistiques/6543200

Bénéficiaires du RSA (BRSA)

- Information générales : https://www.service-public.fr/particuliers/vosdroits/N19775
- [s02] RSA par commune | CAF : https://www.data.gouv.fr/fr/datasets/type-de-revenu-de-solidarite-active-rsa-par-commune/

Sources consultées mais non retenues :

- https://www.data.gouv.fr/fr/datasets/revenu-de-solidarite-active-rsa-par-departement/
- Rapport de la cour des compte (janvier 2022) : https://www.ccomptes.fr/fr/publications/le-revenu-de-solidarite-active-rsa
- Source CAF : http://data.caf.fr/dataset/foyers-allocataires-percevant-le-revenu-de-solidarite-active-rsa-par-commune

Commentaire sur le jeu de données BRSA :

- Malgré des recherches extensives, je n'ai pas pu trouver de données plus récentes que 2017 sur les sites de la CAF, INSEE et du département de la Manche et autres, alors que le dispositif est toujours en vigueur.
- La variable Nombre de personnes par foyer est inexploitable, le lien vers les méta-données est cassé (donc aucune information méthodologique n'est disponible). Cette variable reflète des valeurs élevées (> 1000 individus) non cohérentes une possible moyenne de membre de foyers par commune. Également, lorsque l'on divise ce nombre de membres par foyer par le nombre de BRSA, on obtient toujours des valeurs incohérentes pour des membres d'un foyer (exemple pour 50002 Agneaux : 1636 / 127 = 12.88 pers. par foyer en moyenne)
- La seule variable intéressante et exploitable est le nombre de bénéficiaires du RSA par commune.

Emplois

- [s11] Estimation des emplois par commune en 2021 / données communes 2022 | Observatoire des territoires : https://www.observatoire-des-territoires.gouv.fr/nombre-demplois-au-lieu-de-travail

Sources consultées mais non retenues :

- [s14] : API Urssaff : https://open.urssaf.fr/explore/dataset/etablissements-et-effectifs-salaries-au-niveau-commune-x-ape-last/api/

Activités / Equipements

- [s12] Bassin de vie 2022 | INSEE : https://www.insee.fr/fr/information/6676988
- [s13] Zone d'attractivité 2020 | INSEE : https://www.insee.fr/fr/information/4803954

Sources consultées mais non retenues :

- [s09] https://www.insee.fr/fr/statistiques/3606476?sommaire=3568656
- [s03] Insee - Données harmonisées des recensements de la population 1968-2019 Recensement de la population - Fichier détail: https://www.insee.fr/fr/statistiques/6671801?sommaire=2414232

Pré-traitement

L’ensemble des pré-traitements ont été réalisé dans un notebook spécifique :
Emploi-Manche_Preprocess.ipynb

Filtrage amont

Toutes les sources de données sont filtrées pour ne contenir que le département de la Manche pour éviter tout pre-process non nécessaire.

Corrections géographiques

Les jeux de données utilisés ne reflètent pas la même année, les données RSA les plus récentes sont de 2017, tandis que les données démographique de 2019 et celles géographiques de 2022.

Cette différence à un impact sur le jeu de données RSA, car il contient 30 communes de plus que le jeu de données géographiques (476 contre 446), dans le département de la Manche.
Pour cette étude, les données seront fusionnées sur la base du jeu de données géographique, nous nous sommes donc concentrés sur les communes ayant un nombre de bénéficiaires du RSA, les autres communes n'ayant aucun impact sur la fusion.

Après recherches sur Wikipédia, il s’avère que toutes les communes concernées ont changé de statut au 1er janvier 2019 pour devenir des communes déléguées, et donc rattachées à une autre commune.
Les codes communes INSEE ont donc été remplacées par ceux de leur commune de tutelle.
Une agrégation Group-by à ensuite été réalisée pour retrouver la granularité 1 entrée = 1 commune, en prévision de la fusion.

Gestion des NANs

Les données issues de l'INSEE et de la CAF respectent le principe du secret statistique. Cela signifie qu'une entrée ne peut afficher un effectif de 5 personnes ou moins, pour éviter de faire des recoupements et ainsi identifier les personnes concernées. Dans ce cas, la valeur est simplement remplacée par *** ou un NANs.

Dans notre étude, seul le jeu de données concernant le nombre de bénéficiaires de RSA présente des entrées avec valeurs nulles.

Cas 1 : les valeurs nulles sont remplacées par 1.
Justification : si il y à un NANs, cela signifie qu'il y à au moins 1 personne dans le foyer

- NB_Pers_par_Foyer_Alloc 4

Cas 2 : les valeurs nulles sont remplacées par 0.
Le jeu de données est exhaustif, il contient toutes les communes, donc il est très probable qu'il y ait plusieurs communes avec aucun Bénéficiaire du RSA, donc il n'est pas possible de préjugé si ce NANs représente un effectif de 1 à minima. Les NANs sont remplacées par zéro, ce qui implique que le nombre de Bénéficiaire du RSA sera légèrement sous-représenté par rapport à la réalité.

- NB_Pers_couv_RSA
- RSA_SOCLE_non_Majore_Pers_couv
- RSA_SOCLE_Majore_Pers_couv

Fusion des jeux de données

Les jeux de données ont été fusionnés à partir du DataFrame contenant les données géographique, sur la variable code commune INSEE (['codgeo'])

Exploration

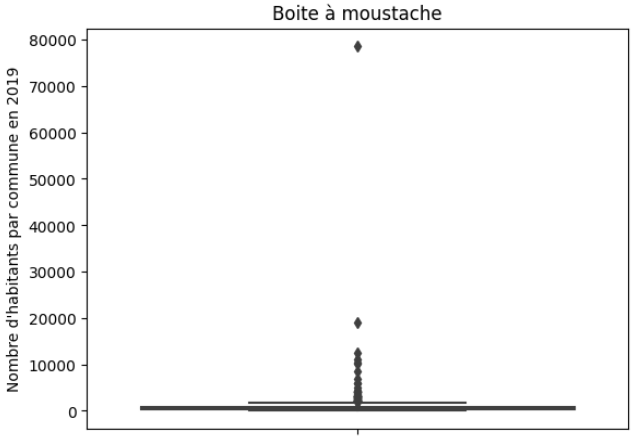
L'ensembles des analyses qui suivent ont été réalisées dans le notebook
Emploi-Manche_Exploration.ipynb

Le département de la Manche fait environ 137 km de long du Nord au Sud, et environ 73 km d'Est en Ouest.
Les distances à parcourir sur l'axe Nord-Sud sont donc à priori plus déterminantes que celles de l'axe Est-Ouest pour définir une zone d'emplois à partir d'un lieu d'habitation.

Sur la démographie du département

Le département compte 446 communes depuis 2019 pour 495 045 habitants.
Certaines communes ont une population nettement importante. On peut distinguer 6 communes de plus de 10 000 habitants, représentant 141 601 habitants, soit près de 29% de la population du département.

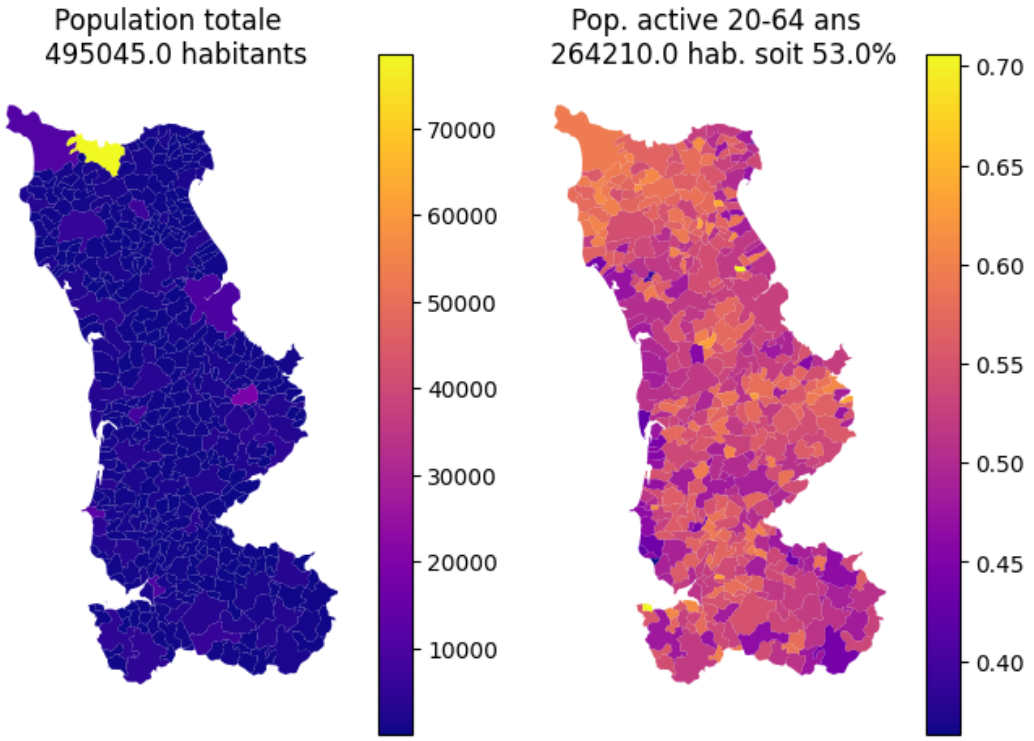
Distribution population / commune

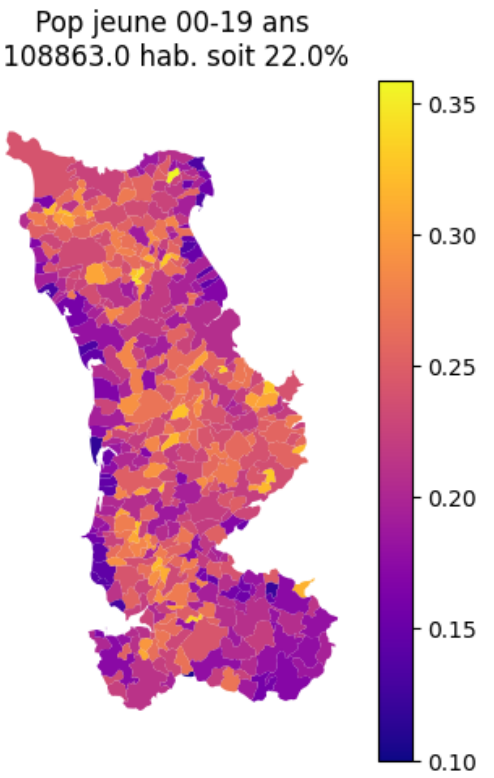


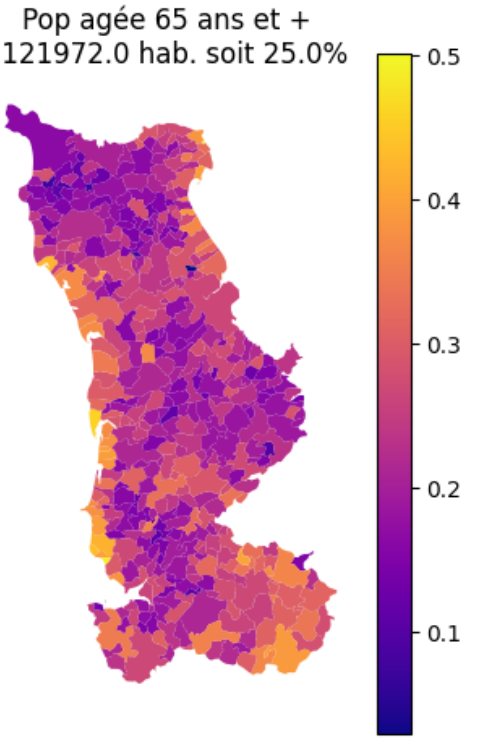
Top 6 communes selon la population

codgeo	libgeo	P19_POP
50129	Cherbourg-en-Cotentin	78549.0
50502	Saint-Lô	19050.0
50218	Granville	12513.0
50041	La Hague	11173.0
50025	Avranches	10264.0
50099	Carentan-les-Marais	10052.0

La répartition de population active [20 à 64 ans] est assez homogène comparativement à la population jeune [0-19 ans] ou plus âgée [65 ans et plus].

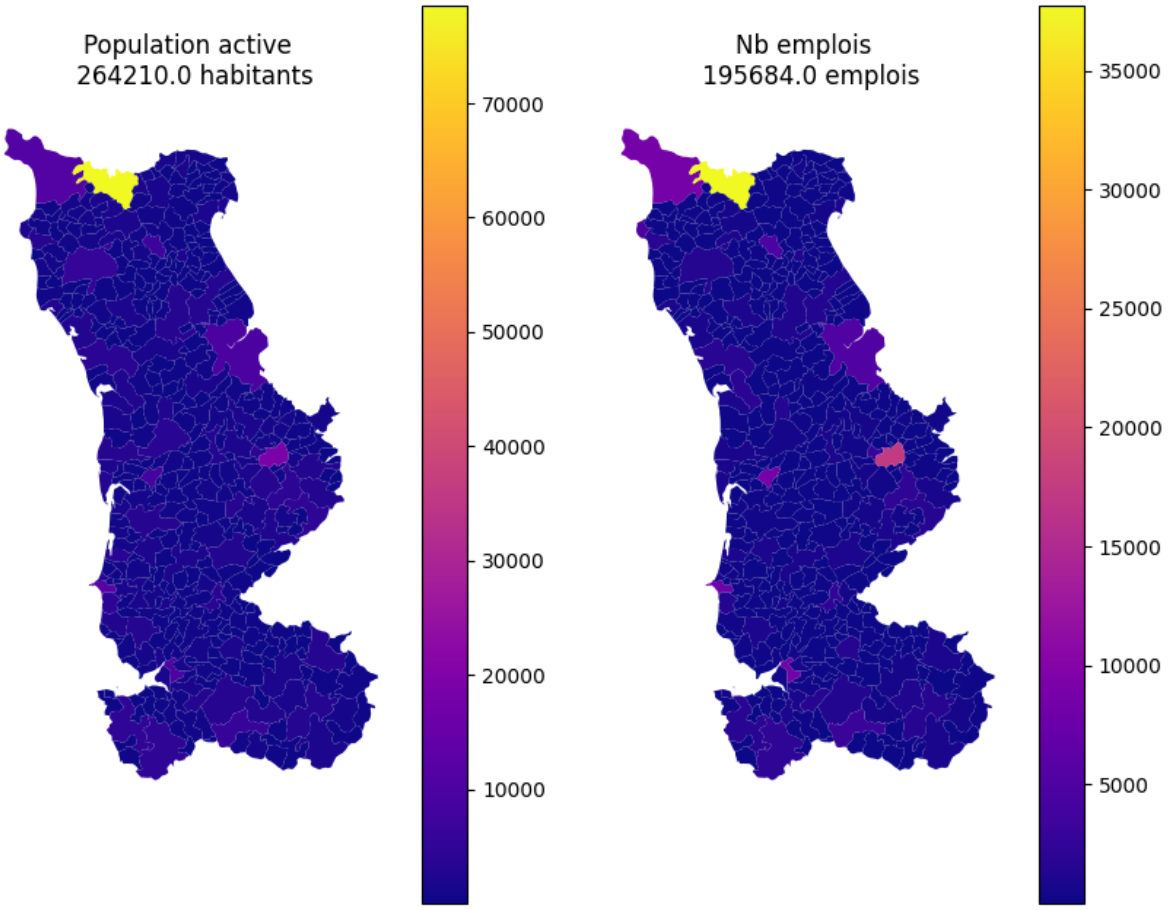






Sur l’emploi dans le département

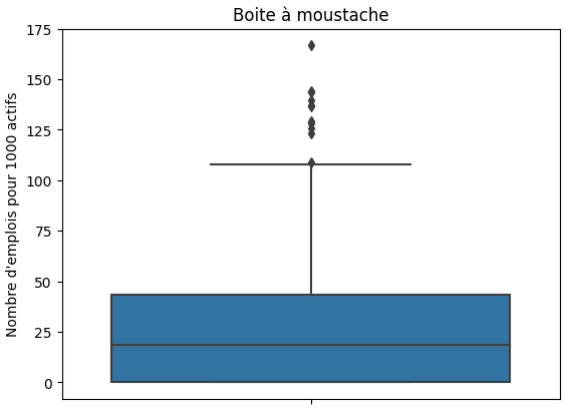
De manière analogue à la répartition de la population, les emplois ont une répartition présentant des forts outliers, qui sont proches de ceux constatés pour la population.



La population active s’entend ici par la tranche [20-64] ans.

Les 10 premières communes en nombre d'emplois représentent 106 548 emplois, soit 54% des emplois du département.

Distribution emplois / commune



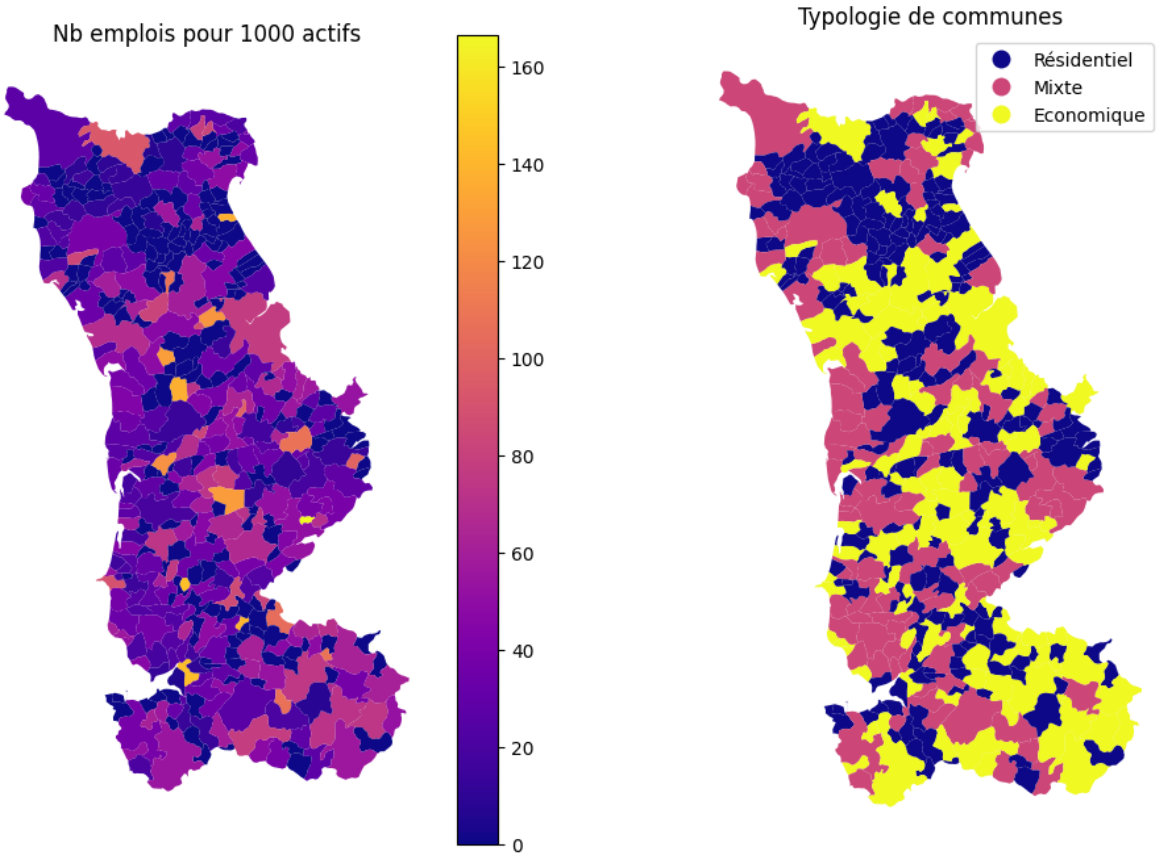
Top 10 commune selon l'emplois

codgeo	libgeo	nb_emplois
50129	Cherbourg-en-Cotentin	37728.0
50502	Saint-Lô	17258.0
50147	Coutances	8818.0
50025	Avranches	8651.0
50041	La Hague	8555.0
50218	Granville	7916.0
50099	Carentan-les-Marais	5117.0
50184	Flamanville	4741.0
50615	Valognes	4591.0
50484	Saint-Hilaire-du-Harcouët	3173.0

Il est important de noter que population et emplois sont deux variables qui ne peuvent se comparer sur un même plan, pour la simple raison qu'une part très significative (non mesurable ici) des habitants ne travaillent pas dans la commune où ils habitent.

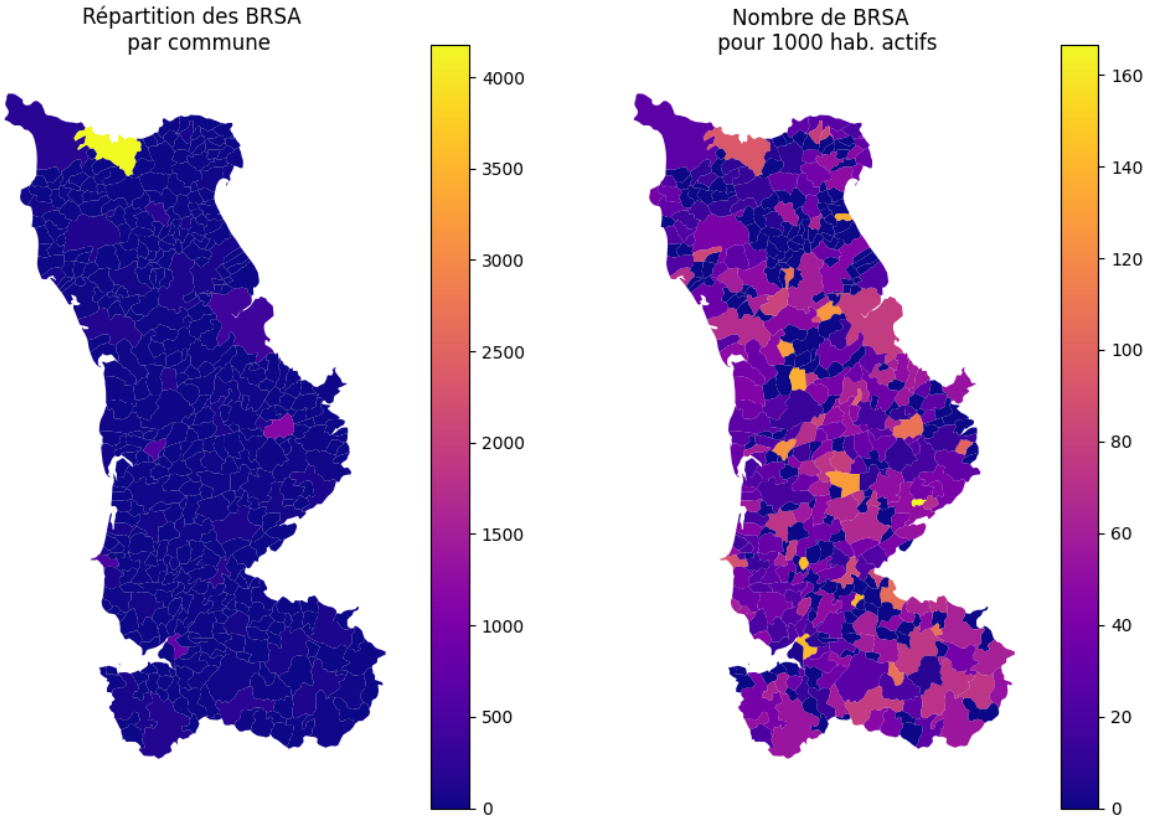
Néanmoins, ce rapport emplois / population permet de créer une typologie de commune, en distinguant celles ayant plus tendance plus résidentielle de celle ayant une activité plus marquée.

Pour cela, en utilisant le ratio nombre d'emplois pour 1000 actifs, il est possible de créer des catégories suivant les quartiles avec la fonction pandas.qcut().

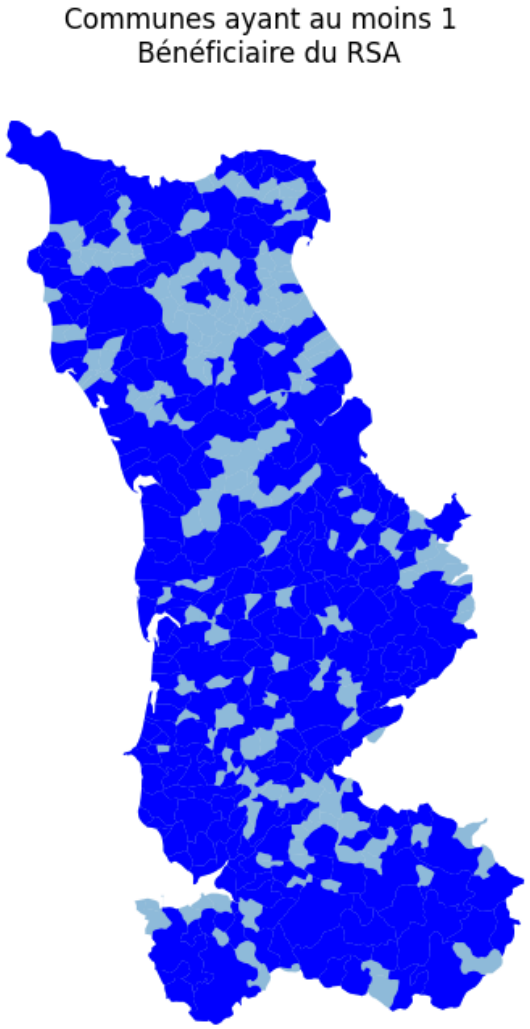


Répartition des Bénéficiaires du RSA

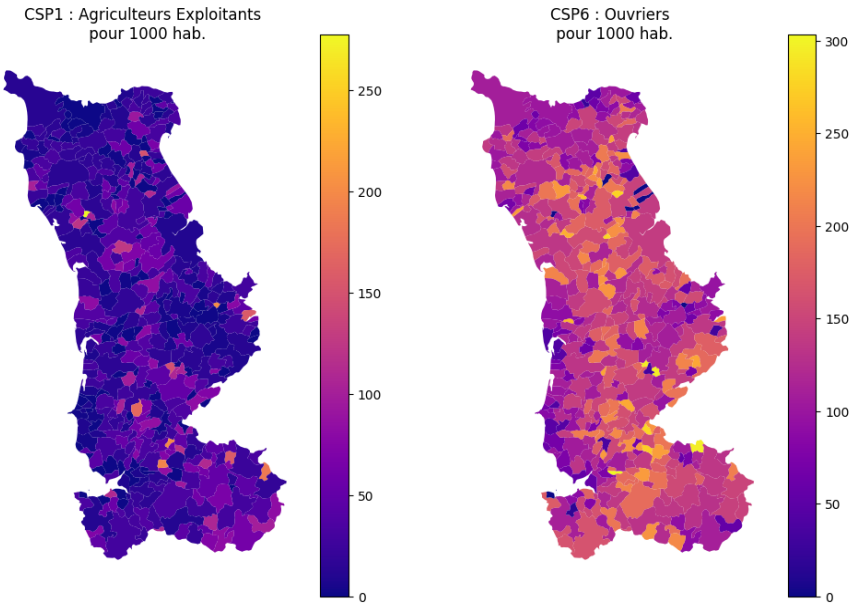
En valeur absolue les bénéficiaires du RSA (BRSA) suivent la répartition de la population active, il n'y pas de spécificités significatives.
En termes relatif, les BRSA sont légèrement plus présents sur un axe Nord-Sud, avec un tendance plus marquée sur l'Est de cette verticale.



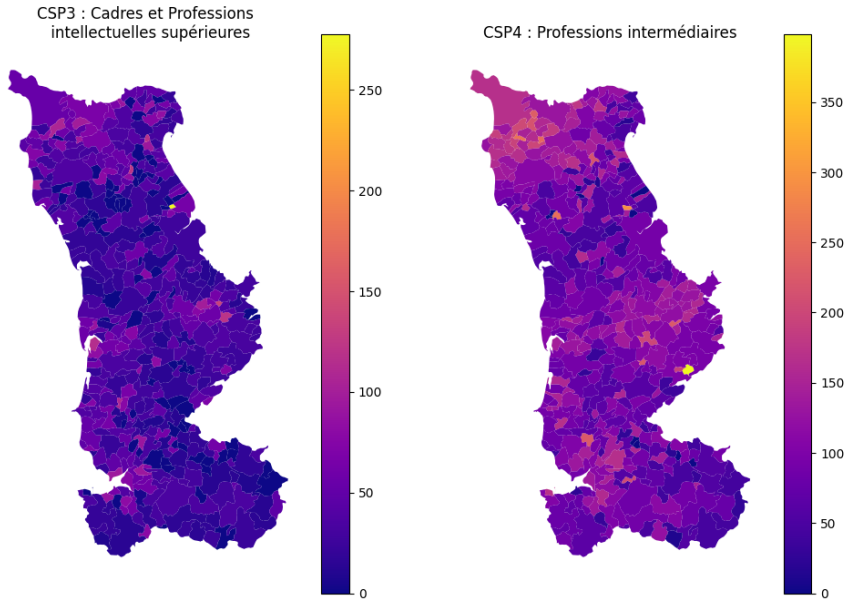
Les bénéficiaires du RSA sont très largement répartis sur l'ensemble du département :



Si l’on compare cette répartition, celle-ci est plus proche de certaines catégories CSP que d’autres.



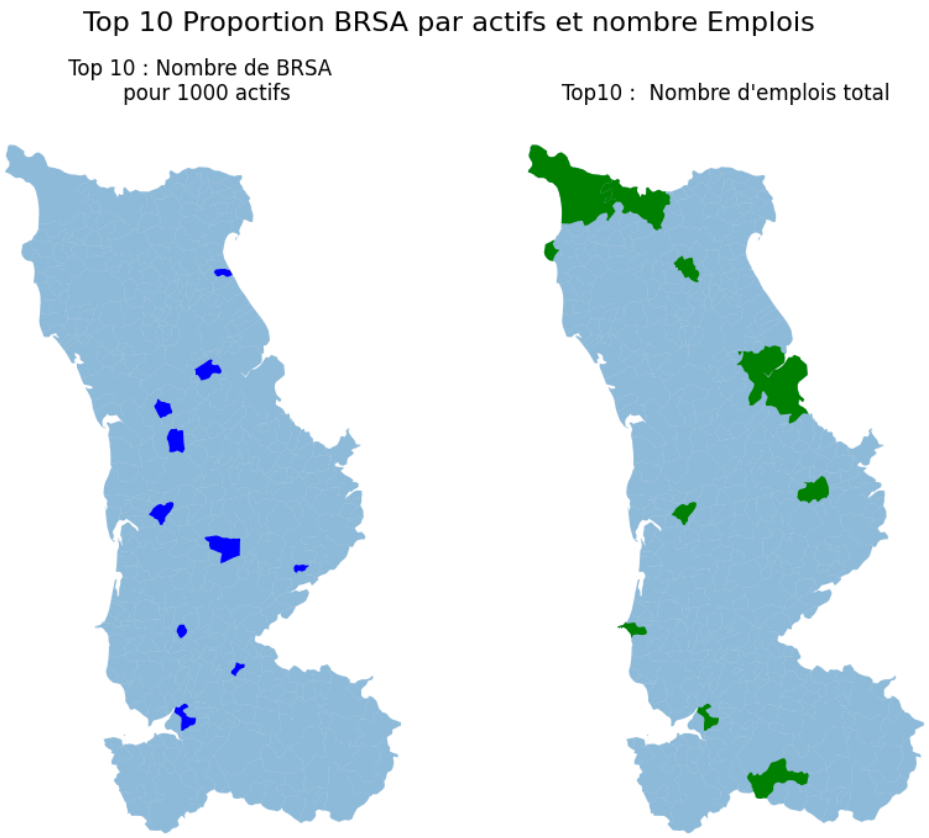
et est à contrario très différentes des CSP3 et CSP4



Le Top 10 des communes ayant le ratio nombre de BRSA sur nombre d'actif le plus élevé.

codgeo	libgeo	P19_POP	ratio_brsa_1000actifs	typo_activite_communes
50192	Fourneaux	130.0	166.666667	Economique
50038	Beauchamps	426.0	144.080784	Economique
50025	Avranches	10264.0	143.639351	Economique
50495	Saint-Jean-du-Corail-des-Bois	75.0	139.534884	Economique
50394	Périers	2256.0	137.087002	Economique
50421	Quinéville	267.0	136.762455	Economique
50265	Laulne	196.0	129.162191	Economique
50378	Notre-Dame-de-Cenilly	634.0	128.316040	Economique
50016	Appeville	182.0	125.555296	Economique
50147	Coutances	8408.0	123.414855	Economique

On remarque que toutes ces communes ont une tendance économique.



On peut observer que les communes où la proportion de BRSA est la plus élevée sont peu souvent celles où il y à le plus d'emplois, à l'exception de:

- Avranches (50025)
- Coutances (50147)

Il y à donc **peut être** une tendance pour les BRSA à être éloignés des principaux "centres d'emplois", mais à résider près de zones économiques secondaires.

L’accessibilité des équipements

Analyse sur les zones d’attractivité

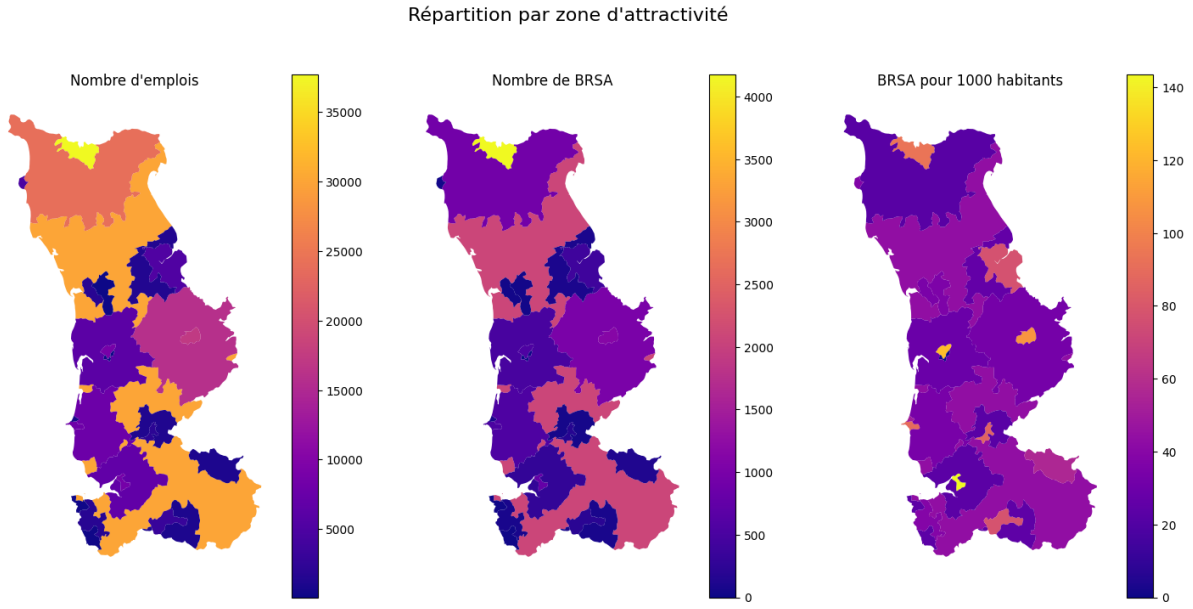
Définition d’une aire d’attractivité selon l’INSEE :

“L’aire d’attraction d’une ville est un ensemble de communes, d’un seul tenant et sans enclave, qui définit l’étendue de l’influence d’un pôle de population et d’emploi sur les communes environnantes, cette influence étant mesurée par l’intensité des déplacements domicile-travail.

Le zonage en aires d’attraction des villes succède au zonage en aires urbaines de 2010.

Une aire est constituée d’un pôle et d’une couronne.

- Les pôles sont déterminés principalement à partir de critères de densité et de population totale, suivant une méthodologie cohérente avec celle de la grille communale de densité. Un seuil d’emplois est ajouté de façon à éviter que des communes essentiellement résidentielles, comportant peu d’emplois, soient considérées comme des pôles. Au sein du pôle, la commune la plus peuplée est appelée commune-centre. Si un pôle envoie au moins 15 % de ses actifs travailler dans un autre pôle de même niveau, les deux pôles sont associés et forment ensemble le cœur d’une aire d’attraction.
- Les communes qui envoient au moins 15 % de leurs actifs travailler dans le pôle constituent la couronne de l’aire.”

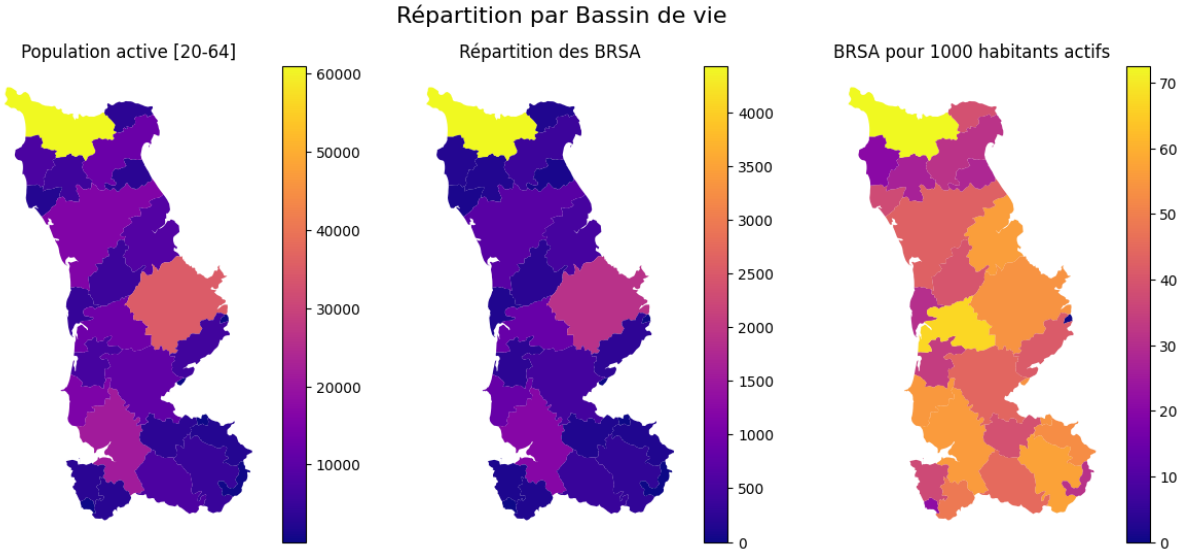


Malgré la pertinence a priori de ce découpage géographique, nous n’avons pas pu tirer d'observation utile pour comprendre un éloignement des BRSA des zones d’emplois.

Analyse sur les bassins de vie

Définition d'un bassin de vie, selon l'INSEE :

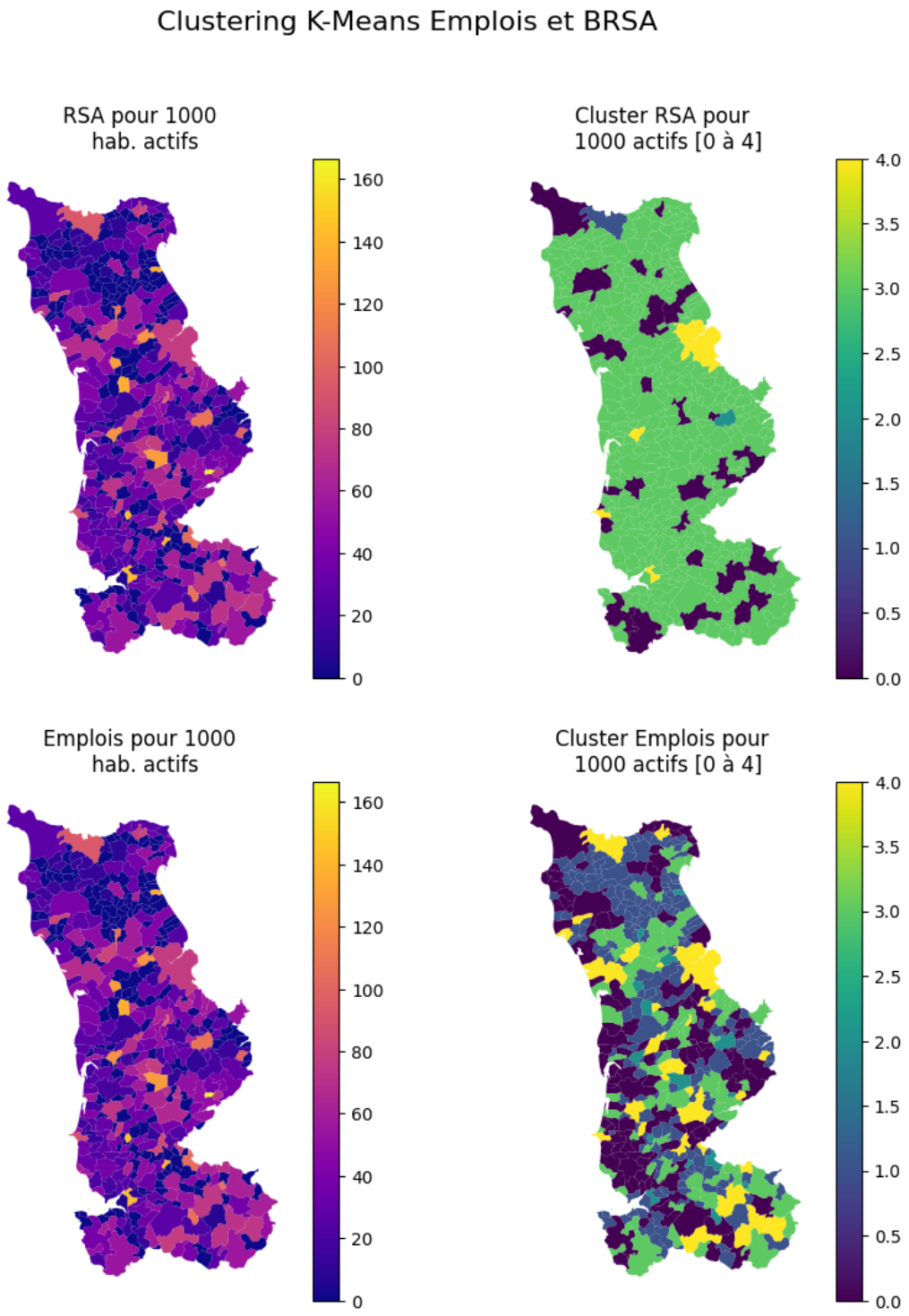
"Le bassin de vie constitue le plus petit territoire sur lequel les habitants ont accès aux équipements et services les plus courants. On délimite son contour en plusieurs étapes.
On définit tout d'abord un pôle de services comme une commune disposant d'au moins 18 équipements sur les 35 de la gamme intermédiaire, avec au moins un équipement par sous-domaine.
La présence d'équipements de cette gamme a été retenue car, moins fréquents sur le territoire, ils sont plus susceptibles de témoigner de la présence, dans ces communes, de services effectivement structurants.
Des zones d'influence de chaque pôle de services sont ensuite délimitées en regroupant les communes les plus proches, la proximité se mesurant en temps de trajet par la route.
Ainsi, pour chaque commune et pour chaque équipement non présent sur la commune, on détermine la commune possédant l'équipement le plus proche de la population.
Les équipements intermédiaires mais aussi les équipements de proximité sont pris en compte. Des itérations successives permettent de dessiner le périmètre des bassins de vie.
"



Malgré la pertinence à priori de ce découpage géographique, nous n'avons pas pu tirer d'observation utile pour comprendre une dynamique des BRSA liée à leur bassin de vie.

Clustering K-Means

Une tentative de clustering à été réalisée, avec pour objectif de distinguer plusieurs clusters d'emplois et plusieurs clusters de communes ayant une plus forte proportion de BRSA.
nous avons choisi de fixer arbitrairement le nombre de cluster afin d'avoir un nombre de cluster identiques qui reste synthétiques, sans être trop simpliste.



Les résultats de ce clustering ne sont pas très exploitables, notamment parce que les BRSA sont largement répartis sur le département et dans un cas de classe minoritaire.

De même, la comparaison entre localisation d'emplois et habitation à des limites, puisque une part très significative des actifs travaillent dans une commune différente de celle où ils résident.

Conclusions

Sur la méthode

La recherche de jeu de données à été une étape difficile et chronophage, car un même jeu de données peut exister sur plusieurs plateformes, être disponible en plusieurs versions géographiques (par commune, IFRI, département, région, …) ou temporelles (2016, 2017, …).
De même, la qualification de la clé de fusion (en l'occurrence le code commune INSEE) n'est pas évidente, car une même variable peut prendre différents noms et définitions pour chacun des jeux de données. Fort heureusement, les jeux de données utilisés ont, à l'exception des données RSA, une documentation détaillée qui permet de faire les liens.

En revanche, j'ai été agréablement surpris par la qualité des jeux de données, qui pris séparément sont très propres, comparativement à d'autres jeux de données que j'ai pu utiliser dans d'autres projets.
Les différences temporelles entre les différents jeux (2017 pour le RSA à 2022 pour l'aspect géographique), et le principe du secret statistique ont requis des traitements.

Sur le sujet

En débutant ce projet, nous avions pour intuition que l'éloignement géographique et des conditions de service publique (par exemple : les écoles) allaient être des axes d'analyses importants, et à donc orientés les premières recherches.

Cependant ce sujet à eu deux principales difficultés, à savoir la pauvreté des informations présente dans le jeu de données RSA (évoquée plus en détail dans la partie source de données), et la situation de classe minoritaire associée à une répartition relative large sur l'ensemble du département (258 communes sur les 446).

Ces facteurs limitants nous ont conduit à envisager d'élargir l'analyse vers un axe plus sociologique (notamment au travers des CSP), cependant n'ayant aucune information de cette nature dans le jeu de données BRSA, nous n'avons pu tirer d'enseignements plus forts que de simples intuitions.

D'après les analyses réalisées, nous avons pu observer qu'il existait une localisation différenciée selon les catégories d'âge et socio-professionnelle.
Mis en parallèle avec la localisation des emplois, nous avons pu en déduire quelques intuitions sur l'importance de la localisation par rapport au lieu de travail.
Nous avons également pu observer une similarité entre la localisation des BRSA et des CSP agriculteurs exploitants et la CSP Ouvriers.

Cependant, en l'absence de données sociologiques sur les personnes BRSA, il n'est pas possible de poursuivre l'analyse.

Le second axe d'analyse envisagée, relatif équipements dont peut bénéficier ou non une personne BRSA, a dans un premier temps été considérée via les découpages géographiques réalisés par l'INSEE sur les bassins de vie et zone d'attractivité.
Ces bassins de vie et zones d'attractivités ont le grands avantages d'être construits sur une méthodologie solide (puisque émanant de l'INSEE), et d'être une référence comparable à d'autres travaux.

Cependant l'exploitation de ces deux découpages n'ont pas permis de mettre en avant une disparité entre les personnes BRSA et les actifs de 20-64 ans.
Nous revenons à une difficultées déjà rencontrée, à savoir la répartition large et diffuse des bénéficiaires du BRSA sur le département.

Axes d'améliorations

Il est également important de tenir compte du fait que les personnes ont une forte tendance à travailler dans une commune où ils ne résident pas. Il conviendrait alors de définir des zones géographiques atteignable en un temps maximal (moins d'1 heure par exemple), pour chaque personne BRSA.
Avec une information sur cette aire géographique d'emplois potentiels, il aurait été possible de faire un lien avec les zones d'emplois les plus importants.
Cependant, la répartition très large des bénéficiaires du RSA (258 communes sur les 446), les conclusions auraient été faibles.

Aussi, la réalisation technique de ces cluster de zones d'emplois potentiels auraient nécessité des techniques lourdes, difficiles à mettre en œuvre dans le temps impartis pour ce projet.
Il était envisagé de récupérer le réseau routier du département (via la librairie OSMnx), calculer une aire d'emplois potentielle pour chaque BRSA, puis réaliser un clustering en utilisant réseau de graphe (avec NetworkX).