ENSG Géomatique

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

Master : Data Science (ENSG-CNAM)
Auteur : Guillaume Morat
Date : Décembre 2022

Etude de la répartition spatiale et de la distribution des prix des locations AirBNB dans Paris pour définir des communautés touristiques

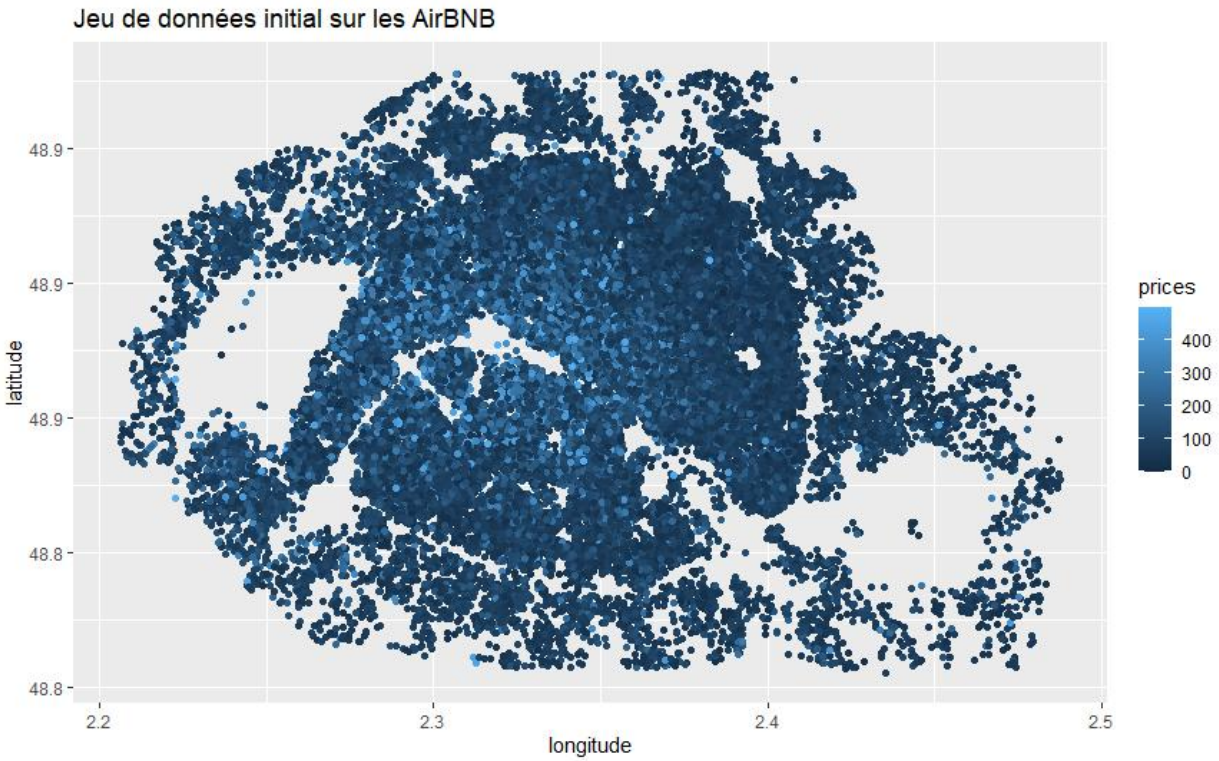

Fig.1 : *Nuage de points représentant les locations AirBNB en région Parisienne en fonction des prix de location en euros par nuit*

Sommaire :

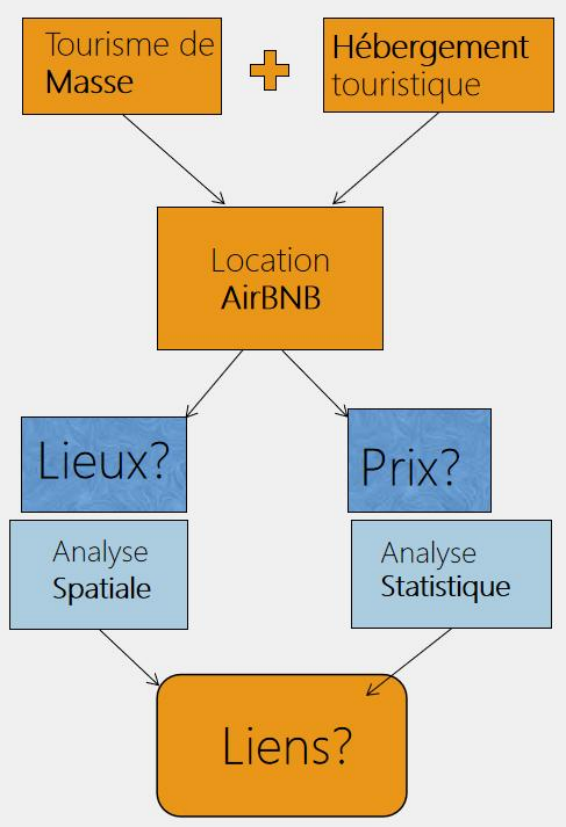

Fig.2 : *Diagramme récapitulatif du contexte, des méthodes et des objectifs du projet*

Introduction et contexte:

L'étude de la répartition spatiale de phénomènes socio-économiques permet de mieux comprendre les interactions et définir les lieux d'intérêts associés aux activités étudiées. Le tourisme permet le développement de l'activité économique
des territoires concernés et la création de nombreux emplois ; notamment par le biais de l'hébergement des touristes dans les villes attractives.

Depuis plus de 10 ans, cet aspect de l'activité touristique a été bousculé avec l'apparition de plateformes de locations communautaires tel AirBNB. Ce réseau de particuliers omniprésents à travers le monde fait de l'ombre aux classiques chambres d'hôtel tant pour le tourisme d'agrément que pour le tourisme d'affaires. En effet, c'est le cas de Paris où plusieurs milliers d'appartements et chambres AirBNB sont disponibles à la location courte ou longue durée !

Ainsi, on peut se demander si l'étude de ce phénomène relativement récent pourrait nous apporter des informations sur le rayonnement touristique des quartiers impactés. Existe-t-il des liens entre le prix et la géo-localisation de ces locations ? Est-ce que l'analyse des prix des AirBNB et leur distribution spatiale à Paris permettrait de mettre en évidence l'existence de quartiers accessibles et attractifs d'un point de vue touristique. Ce projet à pour objectif de quantifier et de qualifier ces zones à fort potentiel attractif dans Paris via différentes techniques d'analyse statistique et spatiale.

I] Analyse des prix des locations AirBNB

Pour cette première partie de notre étude, nous allons nous intéresser aux prix de locations par nuit en euros des appartements AirBNB à Paris. En effet, on fait les hypothèses suivantes :

1. Les prix de location reflètent une certaine attractivité touristique de la zone géographique.
2. Les distribution des prix varient selon les quartiers et arrondissements et peuvent être utilisées comme critère de classification de zones touristiques attractives.

1) Présentation et nettoyage du jeu de données

A) Données

◆ Locations AirBNB

Les données sur les locations AirBNB utilisées dans cette partie du projet proviennent d'**Inside AirBNB**, un projet communautaire visant à mieux comprendre les impacts d'AirBNB sur les communautés résidentielles.

Type : *points*
Nombre d'entité : *61 365*
Source : *«http://insideairbnb.com/get-the-data»*

Ces données sont disponibles au format *csv* avec la latitude et longitude de chaque location en WGS84, on utilisera le package R **«sf»** pour les transformer en données spatialisées.

◆ Les quartiers et arrondissements

On utilisera également les Arrondissements et Quartiers parisiens disponibles au format *geojson* sur le site open data de la ville de Paris. On utilisera le package «geojson sf» pour lire les données.

Type : *polygones*
Nombre d'entité : *80*
Source : *«https://opendata.paris.fr/explore/dataset/quartiers»*

Type : *polygones*
Nombre d'entité : *20*
Source : *«https://opendata.paris.fr/explore/dataset/arrondissements»*

B) Exploration des données

Outre la position géographique et le prix, le jeu de données des AirBNB peut être divisé selon le type de chambre proposée à la location.

◆ *Type de Location*

Il existe 4 type de locations AirBNB :

- les logements entiers (*Entire Home/Appartement*)
- les chambres privées chez l'habitant *(Private rooms*)
- les chambres d'hôtel (*Hotel rooms*)
- les chambres partagées (*Shared rooms*)

Dans la capitale française, elles sont réparties comme dans la *figure 3* avec une majorité de **logements entiers (~84%)**, une petite part de **chambres privées (~13%)** et une **minorité de chambres d'hotel et de chambres partagées (<3%)**

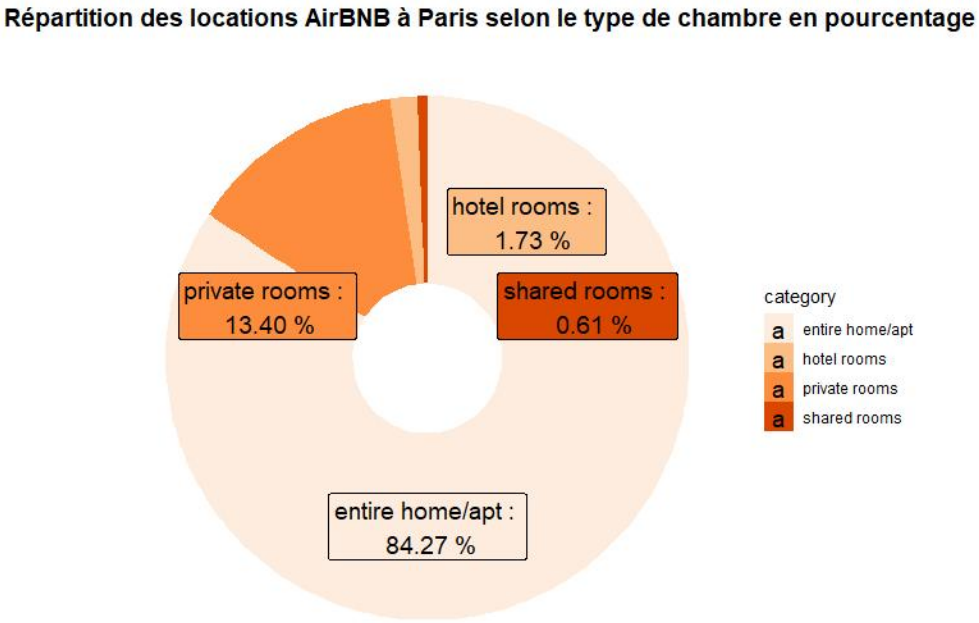

Fig.3 : *Diagramme en disque de la répartition des types de locations AirBNB à Paris (en %)*

◆ *Outliers*

Aussi en voulant afficher les premiers histogrammes des prix, on voit l'existence de certains **outliers**, c'est-à-dire des individus avec des **valeurs aberrantes** en termes de **prix** par rapport à la majorité de la population étudiée. Ces outliers sont représentés par des points noirs dans la *figure 4 ;* avec des prix de location pouvant atteindre plusieurs milliers d'euros par nuit. L'échelle de prix actuelle ne permet pas la lisibilité, il faut filtrer nos données !

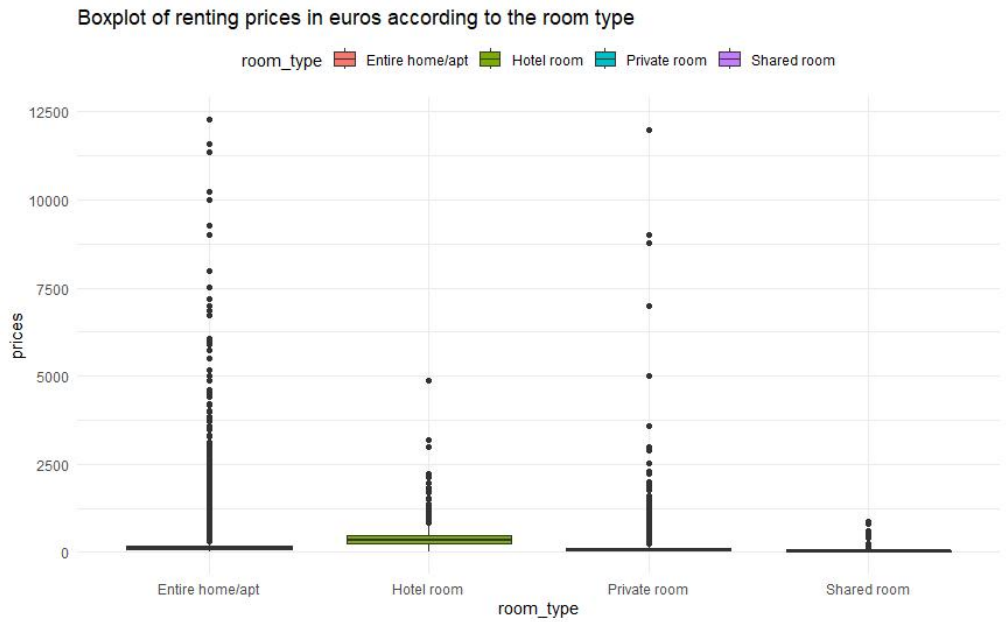

Fig.4 : *Histogramme des prix en fonction des types de locations AirBNB à Paris (en euros)*

En ne sélectionnant que les appartements inférieurs à un certain seuil de prix fixé à 500 euros ***[Fig.5]*** et en affichant le même histogramme on obtient une bien meilleure idée de la répartition des prix de la grande majorité des logements étudiés. On voit directement que les prix médians des chambres d'hôtel est 3 fois plus élevé que celui des logements entiers.

Aussi la médiane des prix pour les chambres partagées est seulement de 40 euros. Les deux types d'appartement minoritaires dont les médianes sont extrêmes par rapport à la majorité de logements entiers et de chambres privées ne seront pas considérés pour la suite de l'analyse des prix. En effet, ils sont peu significatifs en termes de nombres par rapport à la taille totale de l'échantillon d'AirBNB et risqueraient de biaiser certains résultats.

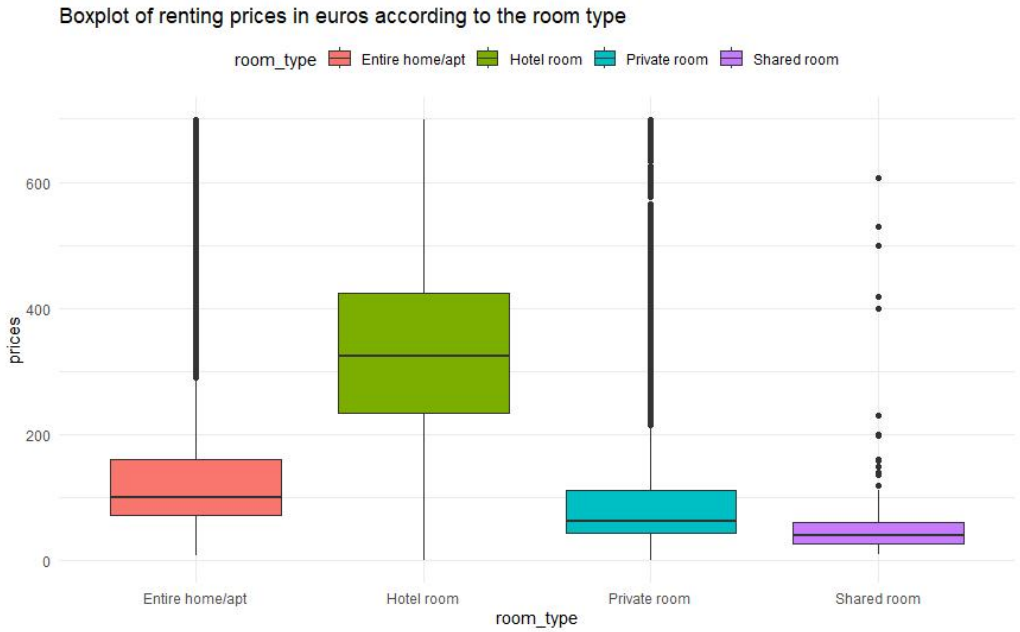

Fig.5 *: Boxplot des prix en fonction des types de locations AirBNB à Paris (dont les prix sont inférieurs à 500 euros)*

2) Analyse des prix

A) Ville : Paris

A l'issue de la partie précédente, nous avons sélectionné uniquement les logements entiers et les chambres privées dont les prix par nuit sont inférieurs à 500 euros pour uniformiser l'analyse des prix ; puisque les types de chambres sélectionnées ont des distributions de prix resserrées autour de leur médianes, on aura un échantillon relativement homogène en terme de prix. Notons que les distributions sont légèrement asymétriques vers les valeurs supérieures et donc que la moyenne est tirée vers le haut.

C'est également ce qu'on pourra observer dans l'histogramme et la fonction de distribution des prix de l'échantillon ***[fig. 6].***

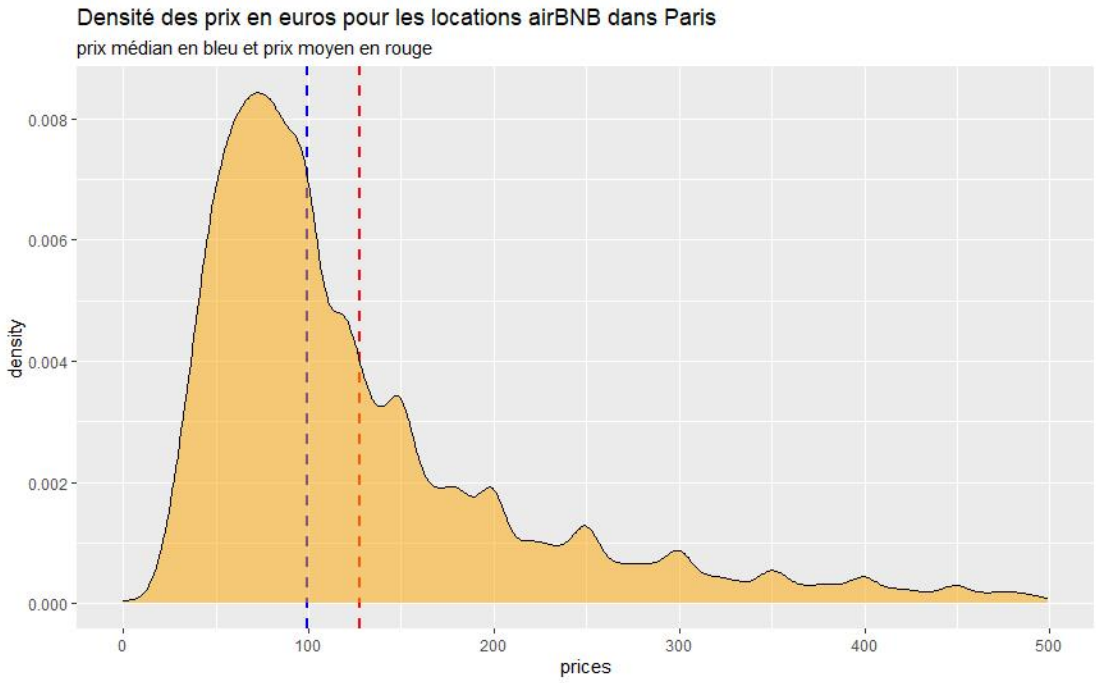

***Fig.6** : Densité des prix en euros des locations AirBNB à Paris (dont les prix sont inférieurs à 500 euros)*

On peut repérer le trait du **prix médian** (99 euros/nuit) en bleu et celui du **prix moyen** (125 euros/nuit) sur les 59 027 individus restants. On a 50% des prix inférieur à 99 euros/nuit et 50% supérieurs et on remarque un **mode principale** pour des prix avoisinant les 70 euros/nuit. Pour des prix supérieurs au 1er mode, on note des diminutions continues du nombre d'appartement concernés avec des **maximums locaux** tous les **intervalles de 50 euros**. On peut supposer que c'est autres modes ont principalement des explications économiques comme le fait que les prix de locations ont tendances à se regrouper autour de «valeurs rondes»

B) Arrondissements

a) Profils de distribution des prix

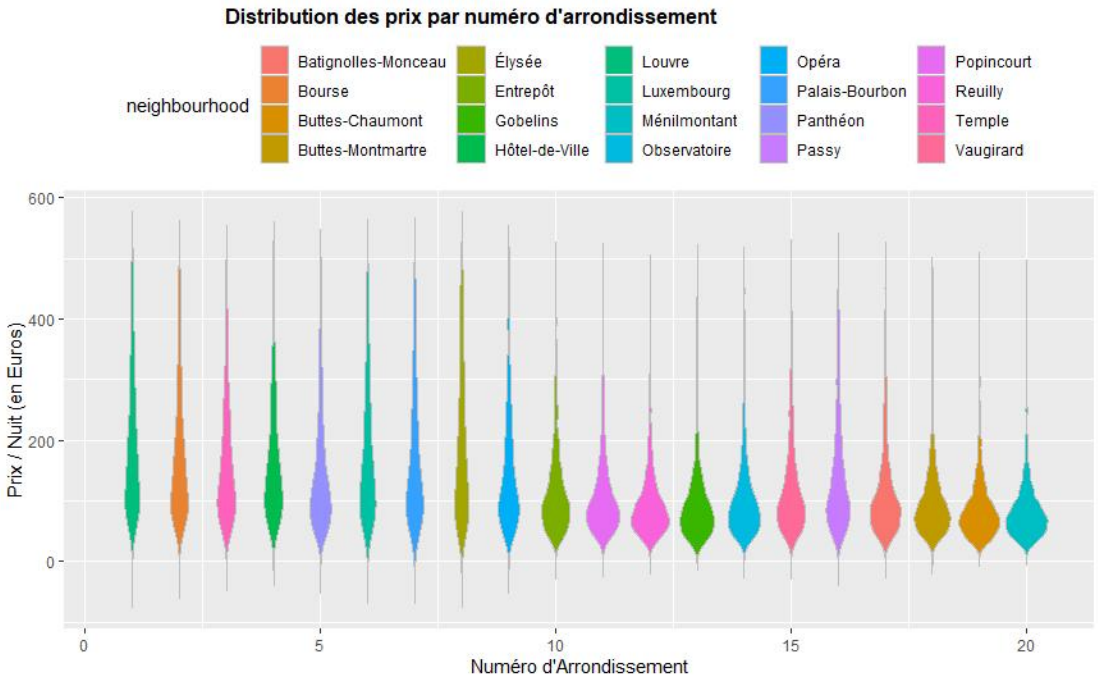

***Fig.7** : Violin plot des prix/nuit en euros des locations AirBNB à Paris (dont les prix sont inférieurs à 500 euros)*

On veut comparer les distributions des prix de locations AirBNB pour les 20 arrondissements parisiens. Pour cela, on réalise un violin plot en fonction du numéro d'arrondissement qui nous permet d'avoir sur une même figure les profils de distribution de prix pour chaque arrondissement. Avec une première analyse visuelle du graphique *figure 7* on peut discerner l'existence de ***3 types de distribution de prix*** :

- Les **profils minces et étirés** qui correspondent aux **arrondissements centraux** *(n° 1,2,3,4,6,7 et 8).*
- Les **profils fins et étirés** qui correspondent à des **arrondissement intermédiaires** *(n° 5,9 et 16)*
- Les **profils larges et concentrés** autour de la valeur médiane*(n° 10,11,12,13,14,15,17,18,19 et 20)* arrondissements plus **périphériques et moins attractifs**

On pourra observer les boxplots des prix *figure 8* dans chaque arrondissement pour avoir une meilleure idée d'indicateurs pour classifier l'attractivité des arrondissements selon le prix de location AirBNB. On peut notifier que les différents profils de distribution de prix par arrondissement peuvent être bien définis par les valeurs médianes, les valeurs moyennes et les écarts-interquartiles.

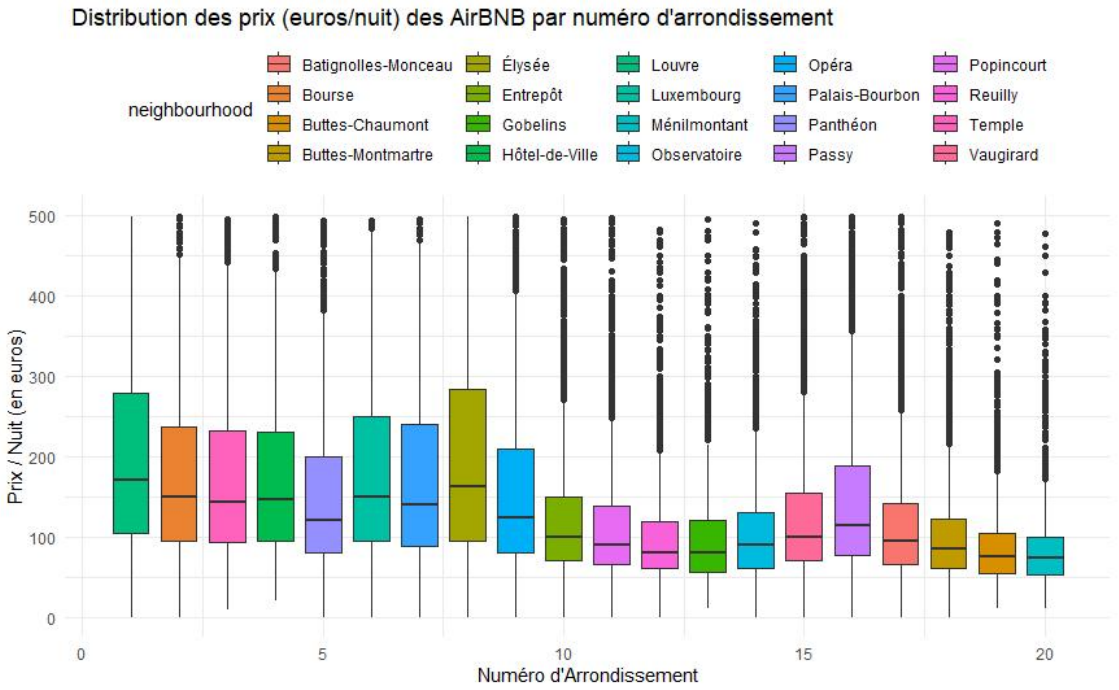

***Fig.8**: Boxplot des prix/nuit en euros des locations AirBNB à Paris par arrondissement*

b) Première classification

En affichant les prix médians des arrondissements en fonction des écarts-interquartiles des prix de location on observe une fonction presque *linéaire (R^2 = 0.90)*. On peut y appliquer un algorithme de **classification K-Means** (en fixant le nombre de clusters à sortir K = 3); qui va regrouper les arrondissements en **minimisant la distance intra-clusters** suivant les deux paramètres choisis.Le partitionement s'effectue en 3 classes *figure 9* qui vont correspondre à celle définie précédemment. Cette classification va donc rassembler les arrondissements les plus **similaires en termes de distribution de prix**.

On définit donc 3 classes d'arrondissement selon le profil de prix, qui reflète une certaine attractivité de la zone géographique d'un point de vue touristique (proximité de sites attractifs, locations bien desservis par les transports en commun, vue sur un monument).

En visualisant ces 3 classes d'attractivité des arrondissements sur la figure 10, on observe que les arrondissements les plus attractifs sont ceux situés dans le centre et l'Ouest de Paris et les arrondissements les moins attractifs quant à eux sont situés en périphérie Nord, Sud et Est de la capitale.

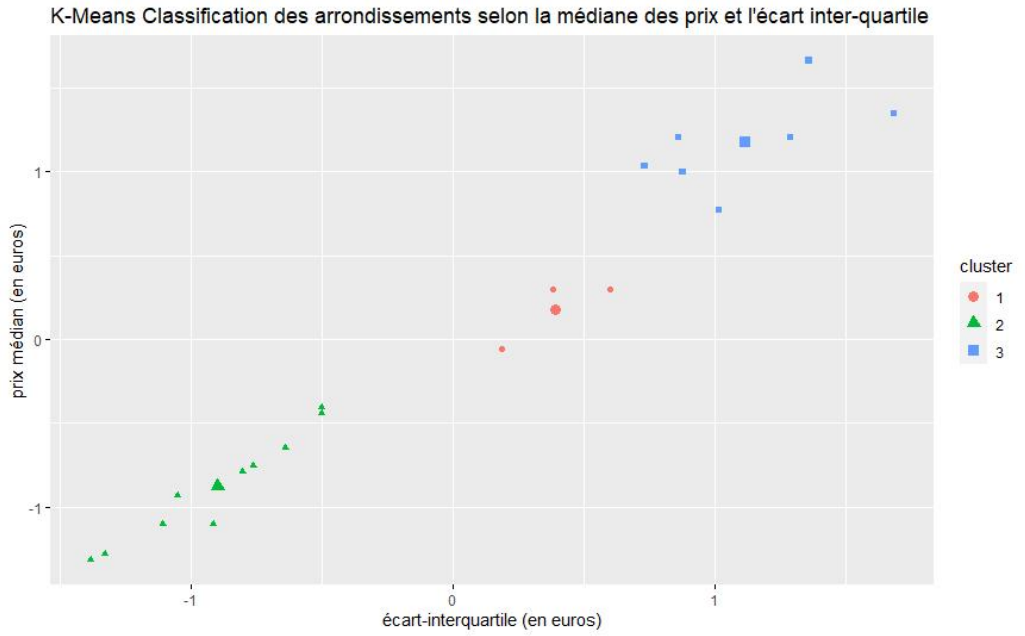

Fig.9: Classification par K-Means des arrondissements à Paris en fonction des prix médians et des écart inter-quartile (normalisés)

Classe : 3 ; Arr. attractif = {1,2,3,4,6,7,8} prix médian fort et écart interquartile élevé
Classe : 2 ; Arr. intermédiaire = {5,9,16} prix médian moyen et écart inter-quartile moyen
Classe : 1 ; Arr. abordable = {11,12,13,14,15,17,18,19,20} prix médian faible et écart interquartile faible

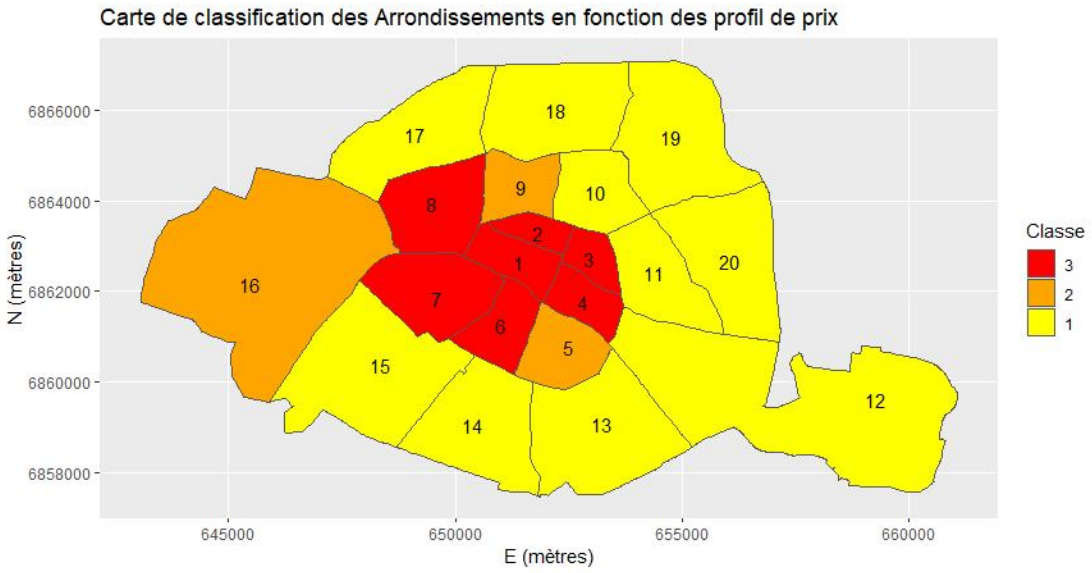

Fig.10: Classification des arrondissements à Paris en fonction des profil de prix

On pourrait désormais être légèrement plus précis et s'intéresser à la distribution des prix des quartiers pour voir les spécificités intra-arrondissement.

C) Quartiers

a) Profil de distribution de prix

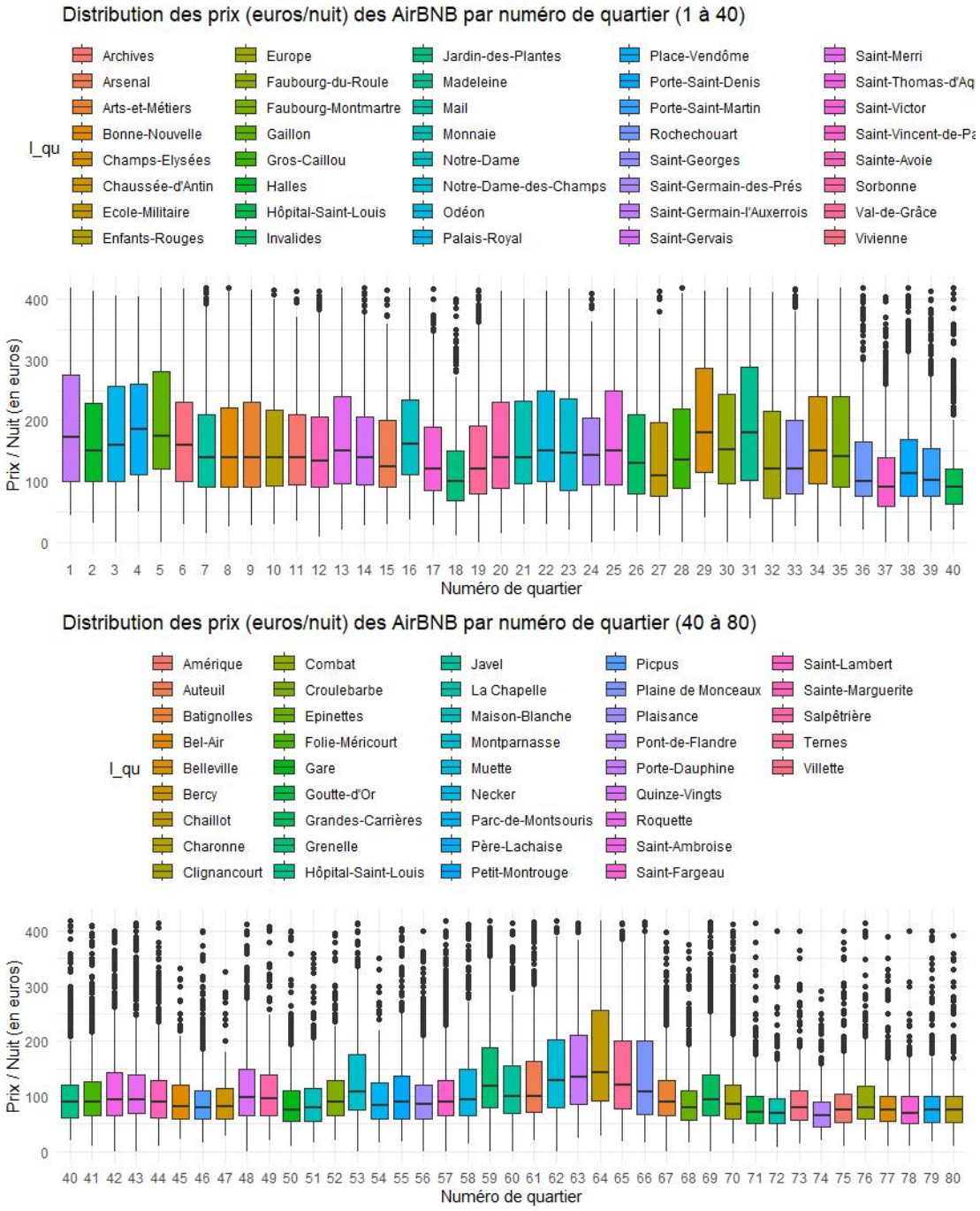

Fig.11: Distribution des prix de location AirBNB par quartiers à Paris en fonction des numéros de quartiers

On peut remarquer dans la *figure 11 ,* les mêmes tendances et les mêmes types de distribution que pour les arrondissements (comme classifié dans la partie précédente). Cependant, on note la présence de quartiers qui vont être plus attractifs en termes de prix que d'autres du même arrondissement. En appliquant une classification des quartiers basée

sur le prix médian et l'écart interquartile et en affichant ce résultat sous forme de carte on peut observer des effets de voisinages des prix comme sur la **figure 12.**

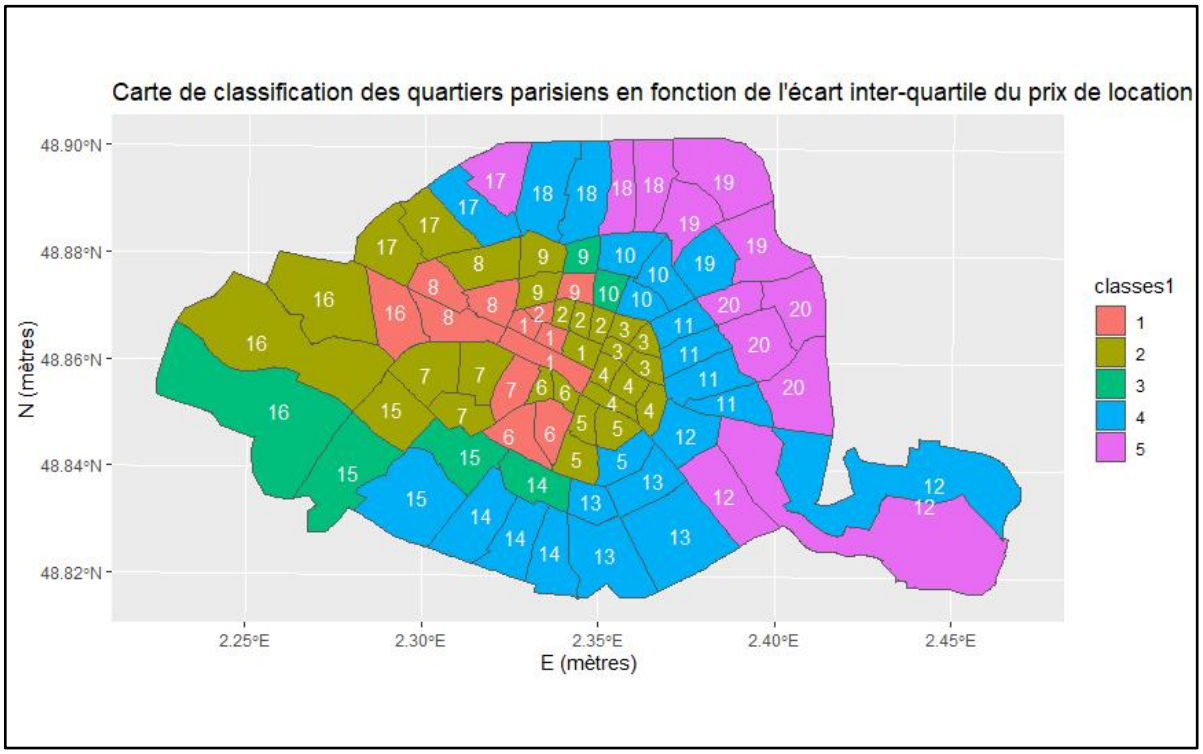

Fig.12: Classification des quartiers à Paris en fonction des prix médian des locations AirBNB

Plus le numéro de classe est petit, plus l'attractivité des prix est importante. On remarque cependant que la classification n'est plus aussi évidente que pour les arrondissements avec seulement ces critères de profil de prix. Les figures ci-dessus mettent en évidence des **effets de voisinage** entre **quartiers des arrondissements limitrophes** pour les valeurs de prix médians et d'écarts inter-quartiles. Plus spécifiquement, pour les quartiers les plus proches du centre, on note des valeurs de prix médians et l'écart inter-quartiles supérieures aux autres quartiers de l'arrondissement.
De même, pour les quartiers des arrondissements périphériques,on remarque que le voisinage de quartiers plus attractifs augmente les valeurs des indicateurs de profil de prix.

On pourra à l'avenir exploiter ces effets de voisinages via une ***régression géographique pondérée (GWR)*** et améliorer notre classification des quartiers en réalisant au préalable une ***analyse par composante principale (ACP)***, afin de définir les meilleurs descripteurs de l'attractivité des prix et de pouvoir mieux identifier les ressemblances entre quartiers.

II] Analyse de la répartition spatiale des AirBNB

1. Jeu de données et méthode

A. Données

◆ Airbnb

Même jeu de données que dans la partie précédente cependant, comme nos traitements d'**analyse spatiale** utilisent des calculs de **densité basés sur la distance**. On va projeter nos données de **WGS84** (*EPSG:4326 en °*) en **Lambert 93** *(EPSG:2154 en métrique)* pour être en **projection conforme**.

◆ Gares de transport

Nom : *metro*
Type : *points*
Nombre d'entité : *390*
Emprise : *région parisienne*
Source : *«https://data.iledefrance-mobilites.fr»*

Nom : *rer*
Type : *points*
Nombre d'entité : *258*
Emprise : *Ile-de-France*
Source : *«https://data.iledefrance-mobilites.fr»*

◆ Sites touristiques

Nom : *monu*
Type : *points*
Nombre d'entité : *692*
Emprise : *Ile-de-France*
Source : *«https://data.iledefrance.fr/explore/dataset/principaux-sites-touristiques-en-ile-de-france»*

On réalise des **intersections spatiales** des **gares de transport** et des **monuments** avec les limites des **arrondissements** pour n'avoir que les sites dans Paris.

B. Hypothèses et méthode

Pour cette section sur **l'analyse de la répartition spatiale** des **AirBNB**, on va essayer de vérifier et de mettre en oeuvre les hypothèses suivantes :

- Il y a un **lien spatial** entre la **densité de gares de transport** et de **sites touristiques** avec la **densité de AirBNB** dans **Paris**.
- Ces **potentielles corrélations** nous permettront de définir des **indices d'accessibilité (transports) et d'attractivité (monuments) par quartiers ou arrondissements**.

Pour la méthode utilisée ici, on va d'abord **estimer la densité de probabilités** de chaque nuage de points à l'aide d'une **estimation par noyau (KDE)**. Puis on calculera une corrélation spatiale entre la densité de AirBNB et celle de chaque site d'intérêts. Enfin pour avoir une idée de l'incertitude de notre mesure de corrélation on va faire un **bootstrap** (tirage avec remise pour modifier légèrement et aléatoirement nos données) 100 fois sur nos échantillons de points et calculer une corrélation spatiale à chaque itération. Puis l'écart-type de cette série de *«corrélations simulées»* nous donne une **estimation de l'incertitude** de notre **corrélation spatiale réelle**. Enfin pour valider la robustesse de ces corrélations spatiales de densité on comparera les corrélations obtenues , avec celle des AirBNB et une répartition aléatoire (simulée 100 fois). Pour conclure, on validera ou non l'hypothèse.

2. Densité

On va donc estimer la densité de probabilité en considérant chaque nuage de points comme une **agrégation de noyaux de distance caractéristique h**. Ici h, dépendras du semi de point considéré et sera dans un premier temps optimisé à la main pour éviter toute sur/sous-estimation de la corrélation.
On utilise la librairie ***R «MASS»*** pour réaliser la ***KDE (Kernel Density Estimation)*** *on peut y fixer deux paramètres* ***h*** *qui va correspondre à la* ***distance d'agrégation des noyaux*** et dépendra donc du jeu de données traitée. On fixera aussi **n = 500** qui correspond aux dimensions de l'image en sortie de l'estimation de densité.
En sortie du calcul, on obtient des images de densités d'AirBNB, de métro, de RER et de monuments géo-localisées **[*fig.13,14*].**

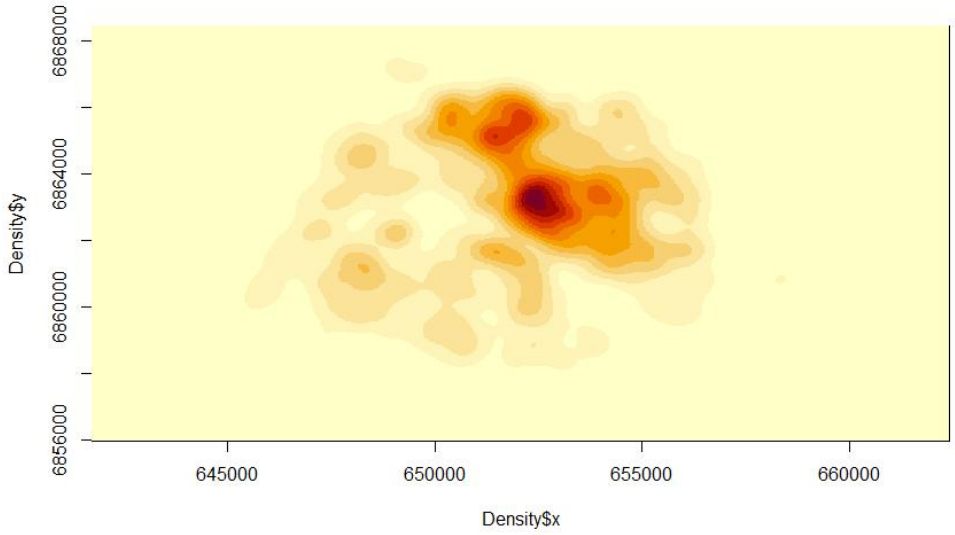

Fig.13: Kernel Density Estimation (h = 1050, n =500) des locations AirBNB à Paris

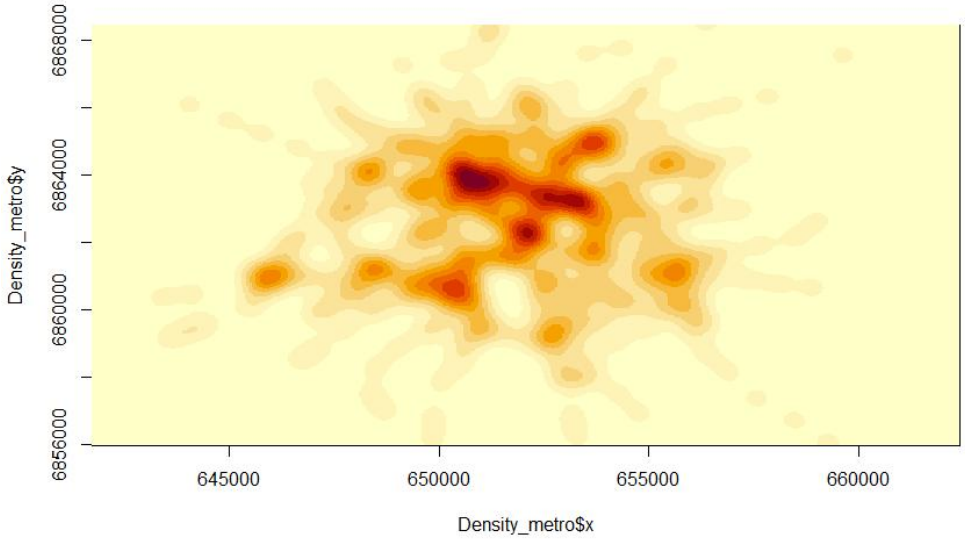

Fig.14: Kernel Density Estimation (h = 1300, n = 500) des stations de métro à Paris

Visuellement, on remarque des **fortes similitudes** entre la **densité de métro** et la **densité de AirBNB** estimées dans la partie Nord et Est de Paris. On pourrait penser pour la suite du projet, effectuer ces estimations de densité par arrondissement ou quartier pour donner des scores d'influence touristique (**accessibilité** et **attractivité** *[Annexe 1]*). Aussi, on pourrait prendre en compte dans ce calcul la **proximité directe** de **sites d'intérêts** ou le ***nombre de sites d'intérêts*** dans un certain **rayon autour des locations**.

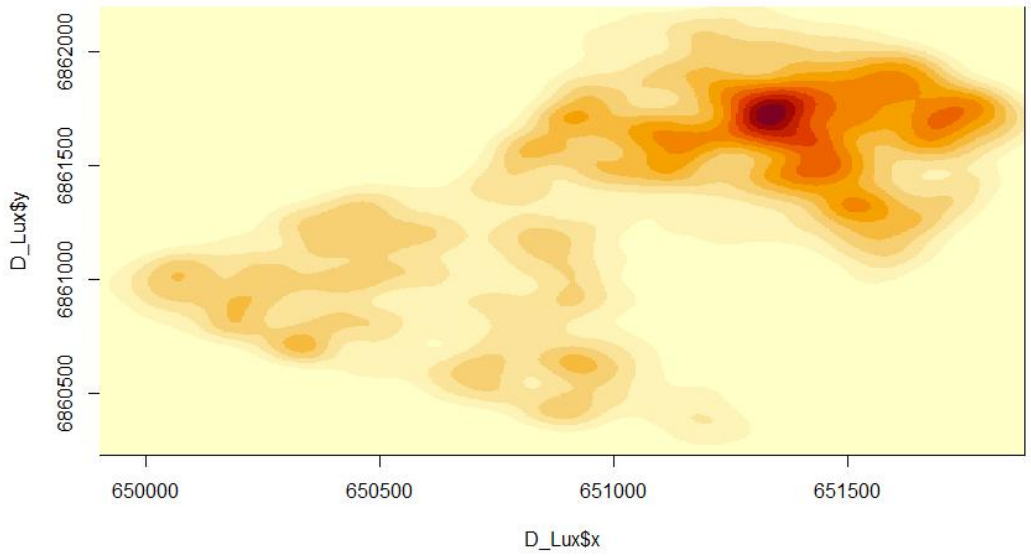

Fig.15: Kernel Density Estimation(h = 200,n =500) des locations AirBNB dans l'arrondissement du Luxembourg

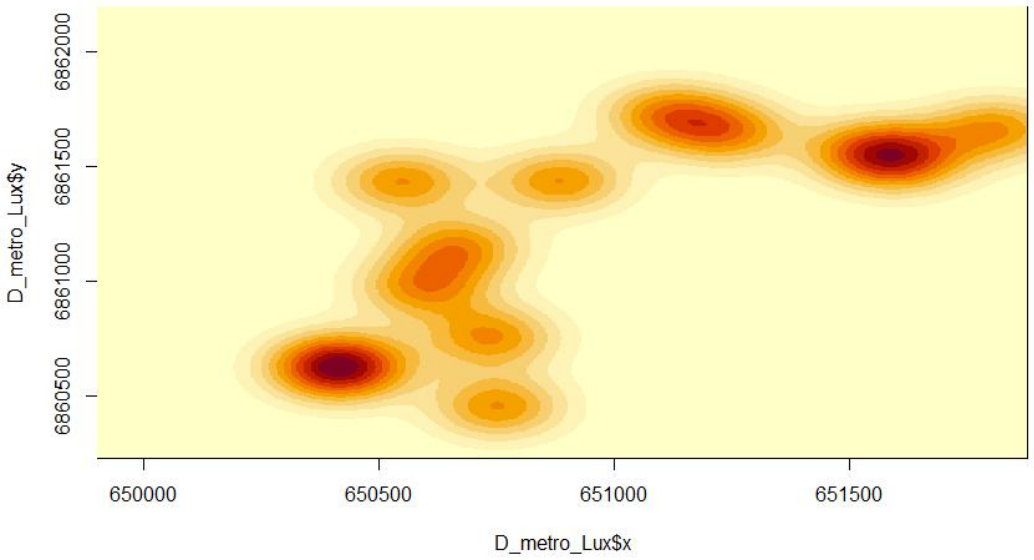

Fig.16: Kernel Density Estimation(h = 200,n =500) des stations de métro dans l'arrondissement du Luxembourg

Pourrait-t'on quantifier ces similitudes?

3. Corrélations spatiales

On va comparer les pixels de densité des sites d'intérêts avec ceux de la densité de AirBNB et calculer une corrélation spatiale un à un ; comme les images ont les mêmes dimensions. Puis, pour chaque corrélation on détermine son incertitude par Bootstrap en s'assurant que l'incertitude reste relativement faible (< 5%) et on compare sa valeur final à la corrélation entre les AirBNB et une répartition de points aléatoire ; afin de s'assurer que cette corrélation n'est pas due au hasard.

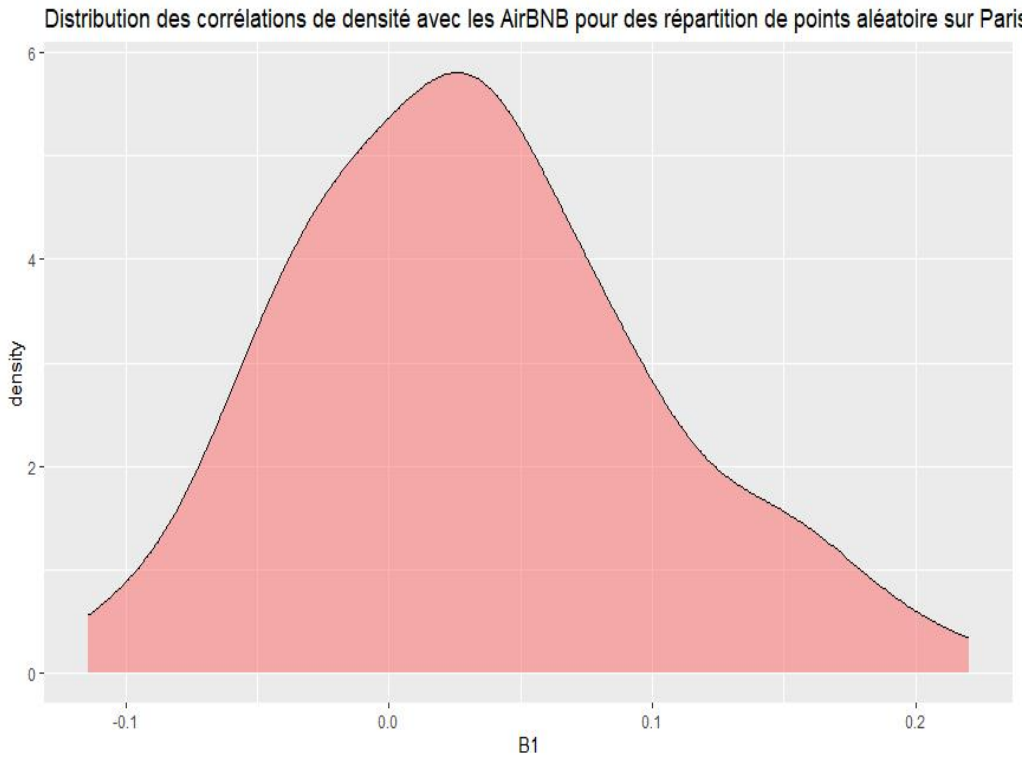

***Fig.16**: Distribution des corrélations spatiales de densité entre AirBNB et des répartitions aléatoires*

	Corrélation spatiale	Incertitude Bootstrap	Sup à 95% aléatoire
AirBNB - Métro	(75.5 +/- 4)%	0,04	Yes
AirBNB - Monu	(41.7 +/- 3)%	0,03	Yes
AirBNB - RER	(37.1 +/- 3)%	0,03	Yes

***Tab.1:** Résultats de la corrélation spatiale entre les densité de points*

Les résultats montrent que la répartition spatiale des gares de transport et des sites touristiques à Paris est corrélée avec la densité de AirBNB ! Nous avons donc pu confirmer notre hypothèse et quantifier la bonne **corrélation spatiale** entre **AirBNB** et les **stations de métro (75,5 +/- 4)%**, **monuments (41.7 +/- 3)%** ou les **RER *(37.1 +/-3)%*** *avec des incertitudes plutôt faibles.*

Conclusion et Perspectives

Ce projet nous a permis de souligner des liens entre les prix et le quartier ou arrondissement de location lors de l'analyse des distribution de prix de locations qui pourra être amélioré à l'avenir... Notamment en déterminant les meilleures variables quantitatives pour décrire les distributions de prix des quartiers de Paris. En prenant en compte le **nombre** de locations et la **surface** du quartier. Aussi, une *régression géographique pondérée* par les **prix de location** pourrait rendre compte des **effets de voisinage** observés pour les quartiers lors de la **première classification** *par analyse de la distribution des prix*. Des calculs d' indices d'auto-corrélation spatiale **(Moran, LISA)** entre les prix de locations pourront être envisagés.

En outre, la mise en évidence de l'**existence d'une corrélation spatiale** entre la *densité de sites d'intérêts touristiques* **(monuments et gares)** et la *densité de **AirBNB*** laisse entrevoir la possibilité de pouvoir affecter des **scores d'attractivité et d'accessibilité** aux quartiers et arrondissements parisiens comme explicité dans l'*annexe 1* . De plus, le calcul **d'indices de densité et de concentration** dans les arrondissements permettrait de définir de nouvelles zones touristiques qu'on pourrait comparer avec les quartiers traditionnels.

Pour conclure, la **distribution des prix des AirBNB** et leur *corrélation spatiale* avec les **monuments** dans Paris reflètent une certaine réalité en termes *d'attractivité du quartier/arrondissements*. En addition, on peut définir des *quartiers accessibles* en termes de **corrélation spatiale** avec le **réseau de transport local.** Cette classification des quartiers pourrait être utilisée pour mettre en place un **modèle d'interactions spatiales** pour estimer et modéliser des **flux liés au tourisme**.

ANNEXES

Annexe 1 : Définition de quartiers d'intérêts

Attractivité : Prix des locations élevés, locations bien corrélées avec les sites d'intérêts

*On pourra par la suite définir des **quartiers attractifs** en :*

- affinant et améliorant les **classes** définies par **l'analyse de la distribution des prix**.

- attachant aux quartiers un **score d'attractivité** fonction de leurs **corrélations spatiales** respectives avec les **sites d'intérêts *(monuments, sites classés, commerces)***

- Appliquant du **clustering spatiale** *[fig.18]* en prenant en compte le **prix** et la **localisation des AirBNB *(DBSCAN)***

- Compter le **nombre de points d'intérêts** dans un certain **rayon de dilatation autour des clusters** correspondant à une distance courte pouvant être parcouru à pied

Accessibilité : Prix des locations modérés, locations bien corrélées avec les gares de transport

*On pourra par la suite caractériser des **quartiers accessibles** en :*

- attachant aux quartiers et arrondissements un **score d'accessibilité** fonction de leurs **corrélations spatiales** respectives avec les ***gares de transport (métro, RER)***

- Faire du **clustering** spatiale en prenant en compte la **distance** et des **indicateurs de centralité *(closeness, betweeness)*** du clusters dans le réseau de transport

- **Classifier les gares** en fonction de leur **importance** dans le **réseau local** et ensuite utiliser cette classe d'importance pour **pondérer le score d'accessibilité** d'un quartier.

- Compter le **nombre de Gares** dans un certain **rayon de dilatation** autour des clusters

Ces deux critères pourront potentiellement être utilisés pour définir un certain **modèle d'interactions spatiales touristiques** entre les **quartiers et/ou les arrondissements** de Paris.

En ajoutant les prix comme paramètres en plus des coordonnées géographiques dans le **DBSCAN** *figure 18* on pourra rassembler les **locations AirBNB** par proximité en termes de prix et de localisation... Les clusters issues de cette classification spatiale seront-ils géographiquement et géométriquement proches des quartiers administratifs?

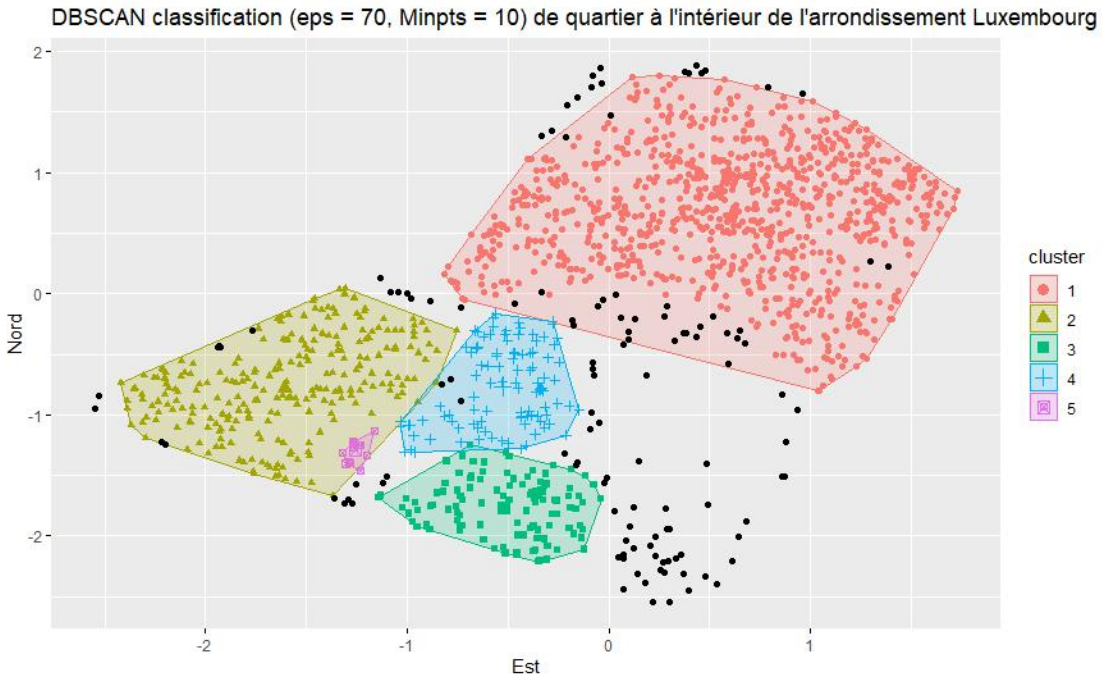

Fig.18: *Classification DBSCAN (eps= 70, Minpts= 10) dans l'arrondissement du Luxembourg*