Analysis of the composition of big cities around Points of Interest

ENSAE Stat'App Report

Github

Group n°32

Students

Simon GENET
Léopold MAURICE
Marie-Olive THAURY

Supervisors

Paola TUBARO
Sarah J.BERKEMER

ENSAE supervisor

Emilien MACAULT

May 2023

Acknowledgements

We would like to express our sincere appreciation to Paola Tubaro, Sarah J. Berkemer, and Emilien Macault for their invaluable support, guidance, and feedback throughout our project.

Paola Tubaro and Sarah J. Berkemer's expertise and insightful feedbacks have been instrumental in shaping our research and setting realistic goals. Their encouragements and mentorship have been a constant source of inspiration. The participation to a submitted paper is a great opportunity that we are grateful for.

We are also deeply grateful to Emilien Macault for his advice and feedback on the mid-semester note. His expertise and constructive criticism have been helpful to improve the project.

We are truly fortunate to have had the privilege of working on the BATO-MOUCHE project, Thank you all for everything.

Contents

Introduction

This project aims to study the accessibility of important amenities (i.e. essential shops and services, schools and cultural institutions) through the concept of X-minute city in Paris and in its periphery. The X-minute city concept was coined by Carlos Moreno in 2015 [1] and reflects the urban planning objective of giving inhabitants access to a certain number of essential amenities within a distance achievable on foot or by bicycle in X minutes. This objective, which was popularised by the Mayor of Paris, Anne Hidalgo, during the 2020 municipal campaign, is to reduce carbon emissions from mobility while promoting healthy activity. The city was a finalist for the *World Resources Institute Ross Center* Prize for Cities in 2021-2022 [1].

Paris is the second most unequal city in France (2018 figures) according to the *Observatoire des inégalités* in terms of income, behind Neuilly-sur-Seine[2]. As matter of fact, the ratio between the minimum income of the wealthiest 10% and the maximum income of the poorest 10% is 6.4 compared to 3.4 for metropolitan France. Furthermore, Paris and the region remain highly polarised areas in terms of wealth, with poles of poverty in the north-east of Paris (St-Denis, Val d'Oise, Seine et Marne) and Ivry, Vitry, Evry or Corbeil; and poles of wealth such as the 16th arrondissement, the "Golden Triangle" or Neuilly-sur-Seine [2, 3]. The districts of the capital also show great disparities in terms of urban functions, with large hubs and intermodal areas such as the Halles district, commercial centres such as the Opéra district and areas considered less dynamic such as the 16th arrondissement. As a result, studying Paris and its region means studying a very diverse and unequal space.

Therefore, this project proposes to study the possible inequalities of accessibility between the different districts of Paris and the *Petite Couronne*[3] through the concept of X-minute city.

To do so, we used the points of interest (POIs) data from *OpenStreetMap* and implemented the two-step floating catchment area method (2SFCA) to measure accessibility. POIs are notable urban locations like schools or shops. We first performed descriptive statistics to get a first idea of the composition of the neighbourhoods. To measure possible economic or socio-demographic inequalities, we then performed spatial regressions of our accessibility indicator on different variables provided by the INSEE *Filosofi* database[4]. We also performed clustering to obtain different groupings of Paris neighbourhoods according to accessibility measures for each amenity. Lastly, we are implementing a Principal component analysis (PCA) to study which types of POI best explain the accessibility of a given neighbourhood.

1 Accessibility to local services : literature review

As accessibility to shops and essential local services is a major issue for sustainable development in the territories, many recent studies sought to measure it in order to assess possible inequalities between different areas, particularly between the city and the suburbs.

Thus, in 2020, INSEE published a study [4] in which researchers implemented the 2SFCA, whose method will be explained later (Section 2.4) in order to evaluate individuals' access to shops throughout the country. They also used the distance between individuals and shops as well as a Family Budget enquiry to measure the preference of households for local shops. This study reveals a strong inequality of access between households living in the centres of large urban areas and those living on the outskirts, since 20% of those living on the outskirts have poor access to the essential retail offer, compared to 0.4% for those living in the centres. While these results are interesting for measuring inequalities between different types of spaces, the distance implemented in the model is 20km, i.e. a distance that can be achieved daily by car and not by foot or bicycle. Thus, this study does not allow us to account for possible inequalities within large cities.

On top of showing that the countryside has fewer shops per capita than other cities, another INSEE study from 2017 [5] highlighted differences within large cities. As a matter of fact, by calculating the distance as the crow flies between households and shops, they showed that the population density of the neighbourhood is a

[1] "The 15-Minute City" WRI Ross Center Prize for Cities

[2] "Le palmarès des villes françaises les plus inégalitaires"(2021), *Observatoire des inégalités*

[3] The *Petite Couronne* is the immediate suburbs of Paris, made up of the three departments bordering the capital: Hauts-de-Seine (92), Seine-Saint-Denis (93) and Val-de-Marne (94). It includes more than a hundred towns.

[4] INSEE is the French national institute of statistics and economic studies.

major factor in the proximity of shops and that butcher's shops are more accessible in modest districts, and fishmongers in well-to-do districts. We will seek to verify some of these results in our study using the 2SFCA and regression method.

As the concept of the X-minute city has grown in popularity, many researchers have looked into the issue. For example, an article by Knap and al [6] seeks to evaluate the 10 and 15-minute city concept in the city of Utrecht. To do so, they build an accessibility score per amenity type based on the infrastructure (roads, bike network), the residents' behaviours (cycling speed, age), a distance decay function (the further away the amenity is, the less likely it is to be visited), and the demand for this amenity (the higher the demand for the service, the less accessible it is). They also create a X-minute global accessibility score : they make a service-need weighted sum of the accessibility scores of each type of amenity (e.g. food supply is more important than restaurants, and will therefore be taken more into account). Accessibility scores used in this article are an application of the **2SFCA method**, which will be discussed in more detail in Section 2.4.

The authors namely shed light on the fact that people living in the city centre of Utrecht have a higher 10-minutes score than those living in the peripheries. To understand such differences, they implemented spatial regression models on socio-economic variables such as the percentage of people receiving the minimum income in the neighbourhood. For instance, they found a negative relationship between the 10-minutes score and the percentage of the population below 15 years old, which means that households with children have a poorer access to amenities and shops compared to households without children. This article was very interesting for us to understand the X-minute city concept and gave us a precise and clear method to implement in our project.

2 Methodology

2.1 Data : OSM and Filosofi

2.1.1 OSM : question of quality

We used data from *OpenStreetMap* (OSM). OSM is a free data set which covers spatial information such as buildings, transportation networks and point-of-interest data. Just like *Wikipedia*, the development of this OSM is collaborative, which means that it is the users who fill in the information.

Points-of-interest are divided into several categories (public, health, leisure, catering, accommodation, shopping, money, tourism...), which are themselves subdivided into tags. For example, OSM distinguishes bars from restaurants and fast-food outlets. These data thus make it possible to see the difference between a touristic, residential, commercial or wealthy district.

Nevertheless, OSM has some shortcomings. As a matter of fact, as mentioned before, the database is filled in by anonymous contributors, which can lead to errors and heterogeneous information. Moreover, there is no qualitative information on the POIs. Thus, we can not know how expensive a restaurant is and whether it is well rated unlike other databases such as *Foursquare*. However, if one is only interested in the geographical quality of the data, a review by IGN officers evaluated OSM data in France as relatively good [7].

2.1.2 Combination with socio-economic data from Filosofi

To carry out our various analyses, we used the data from the *Filosofi* system (localized fiscal and social income). This INSEE database divides the territory into 200-metre squares, thus overcoming administrative boundaries. It provides, among others, variables such as the age pyramid of the inhabitants, their income and the year of construction of the buildings. One of the problems with this database is that it dates from 2015, so it may have observations that are a bit outdated.

2.2 Choice of categories

In order to analyse the city composition around POIs, we define aggregated categories of POIs, based on *OpenStreetMap* tags :

- **Restaurants** : all types of restaurants including cafes, bars, fast-foods, pubs, ice-cream shops.
- **Culture and art** : shops and amenities related to literature, music, cinema, plastic arts, performances, video games, games.
- **Education** : primary schools, middle schools, high schools, colleges, universities.
- **Food shops** : including supermarkets as well as specialist food shops (e.g. bakeries, butchers, dairy shops, seafood shops, wine shops...).
- **Fashion and beauty :** all shops related to clothes, fashion accessories (e.g. jewellery, watches), beauty care (e.g. cosmetics, hairdresser, massage, hair removal, perfumery).
- **Supply shops :** everyday life shops apart from food shops (e.g. insurance, sport shops, furniture shops, household appliance shops ...).

Details of the *OpenStreetMap* tags in each category are listed in Table 4.

2.3 Basic descriptive statistics

So as to verify the relevance of our analysis, our first intention was to produce statistics in order to analyse the composition of a handful of districts in Paris. For this purpose, we arbitrarily chose a dozen of places, such as monuments or metro stations, around which the social life of the neighbourhood is organised. The neighbourhood we analyse corresponds to a perimeter of one km of the pedestrian network around the chosen location. In each neighbourhood, we count the POIs for each of the categories defined above.

Basic descriptive statistics allows us to analyse and compare the composition of Paris at the scale of neighbourhoods organised around a central location. In a second step, we use accessibility scores to obtain a more global analysis of the city.

2.4 2SFCA and accessibility measures

2.4.1 Accessibility scores for each type of amenity

Accessibility scores are an application of the **two-step floating catchment area method (2SFCA)** [8]. It has been first used to measure health services accessibility, but it can be applied to any sort of extensive variables like the number of amenities for instance. The idea behind 2SFCA is to measure for each service provider the surrounding demand. Then for each person (or place) asking for this service, we calculate the surrounding offer by considering that each service provider divides itself up on the previously calculated demand. Figure 2.1 - originating from an IRDES paper [9] - illustrates it.

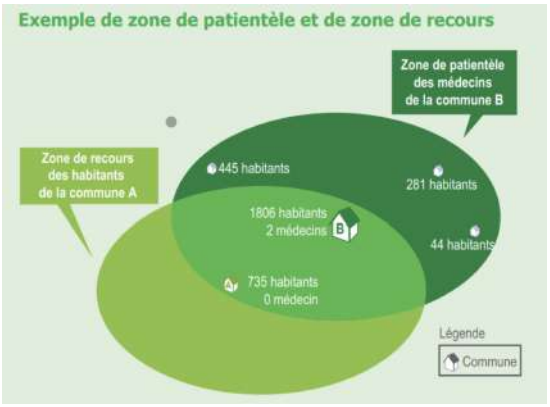


Figure 2.1: 2SFCA rationale illustration.

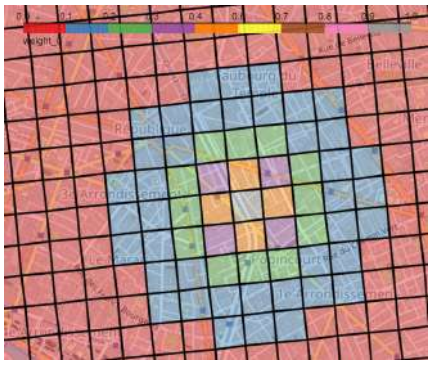


Figure 2.2: Weights associated to each square for the grey square in the center

As explained earlier, Paris Mayor insisted on a 15 minutes walking city, which corresponds to a 1-kilometer radius. Inspired by Knap and al. [6] and Paris Mayor policy, we focus on the composition of cities driven by services accessibility in a one-kilometer radius (corresponding to roughly a 15-min walk at a 5 km/h speed). So

we decided to limit the accessibility to this radius to measure the 15-min accessibility. We obtain weights as illustrated in Figure 2.2.

To go a bit more into details, we will note by j a square representing a supplier, while we will note by i a square consuming services (each square are both supplier and consumer). We also note k the index for the squares in the one-kilometer zone around j. Then, if we call D_k the demand (measured in inhabitants) in the square k, S_j the supply in the square j (measured in number of POI), W_{kj} the permeability coefficient of demand from k to go to square j (e.g. $W_{kj} = \frac{1}{d_{kj}}$), and $\mathbb{P}_{kj}$ the probability that inhabitants in square k visit square j (e.g. $\mathbb{P}_{kj} \propto W_{kj}$), we can define the demand received by the square j (counted in inhabitants) as :

$$\mathbb{D}_j = \sum_k \mathbb{P}_{kj} D_k$$

From that aggregated demand taking into account the zone around square j, we can compute the supply per inhabitants (the "patientèle" zone in Figure 2.1) as :

$$R_j = \frac{S_j}{\mathbb{D}_j} = \frac{S_j}{\sum_k \mathbb{P}_{kj} D_k}$$

And lastly, the accessibility indicator (or 2SFCA score) for the square i counted in number of POI accessible in a one-kilometer radius per inhabitants :

$$2SFCA_i = \sum_j \mathbb{P}_{ij} R_j$$

Here, we choose the following weight function, with d_{kl} the distance between k and l :

$$W_{kl} = \frac{1}{d_{kl}^2} \mathbb{1}(d_{kl} \leq 1000m)$$

For each category (established on Section 2.2) or for each sub-type of points of interest (*OSM* tags), we count the total number item in each square in a grid (square of 200m*200m), which represents the supply S_j as defined earlier. This grid - named INSPIRE - comes from the Filosofi data set previously presented. We obtain for instance maps like in Figure 2.3a for restaurants in Paris. Then, for each category, we compute the 2SFCA score as shown in Figure 2.3b, with here for instance the number of restaurants per square as the supply (and like for all computation, the number of inhabitants per square as the demand). On top of the accessibility score for each category of amenities, we also calculate the accessibility score for housing and social housing thanks to the data provided by the Filosofi dataset.

2.4.2 A basic measure of aggregated accessibility score

Unlike Knap and al. [6], we are not given weights for each type of amenity corresponding to its use. One solution we found was to calculate, for each type of amenity (i.e. for each category), the ratio between the number of items for the category N_p on the total number of amenities N and give a weight opposite to this frequency :

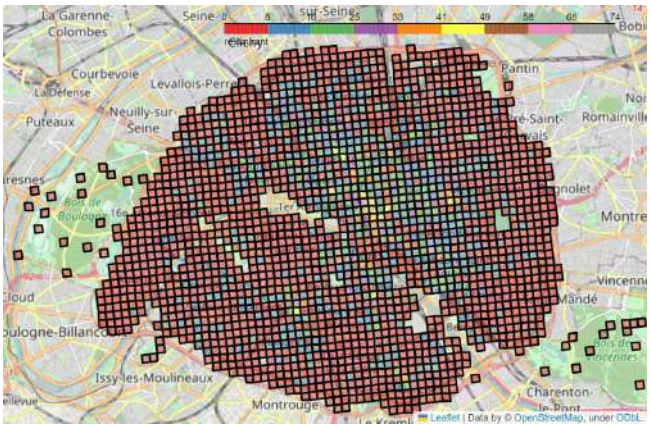
$$w_p = \frac{N_p}{N} \qquad \text{where } p \text{ describes the categories established in Section 2.2.}$$

The idea is that the fewer amenities there are, the more important the amenity in question is : there is only one school when there are 20 restaurants. But this leaves many biases by overvaluing cultural and educational places.

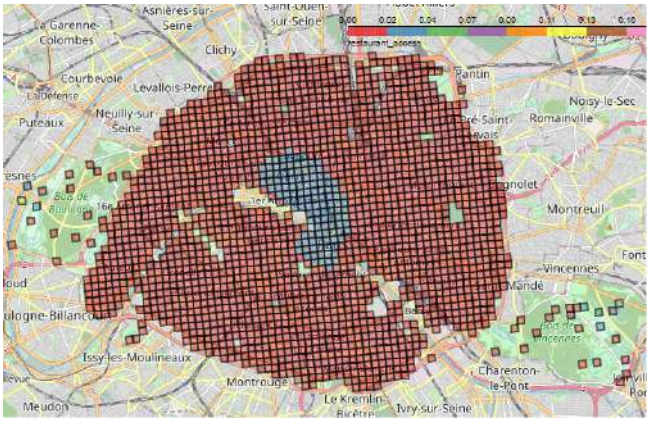
This would give the aggregated 2SFCA (CS_i^{15}) for cell i:

$$CS_i = \sum_{p=1}^{P} (1 - w_p) \times X_{i,p}$$

where $X_{i,p} = \frac{2SFCA_{i,p} - \min_j 2SFCA_{j,p}}{\max_j 2SFCA_{j,p} - \min_j 2SFCA_{j,p}}$ is the min-max normalisation of the accessibility score of cell i for amenity of type $p \in P$ (the total number of amenity types), $2SFCA_{i,p}$.



(a) Number of restaurants in Paris aggregated on the INSPIRE grid



(b) 2FSCA score of restaurants in Paris on the INSPIRE grid

Figure 2.3: INSPIRE grid (200mX200m squares) : number by square vs. accessibility score

2.5 Regression

The purpose here is to obtain econometric results for Paris that we can compare with Utrecht. To do so, we set up a methodology as close as possible to the one introduced by Knap and al for Utrecht [6].

2.5.1 Spatial Weight Matrix

The spatial weight matrix makes it possible to account for the geographical relationships and influences that exist between the different units in the database. There are two types of weights[5]: contiguity weight and distance-based. Given the characteristic of Paris and our database (each unit is a square) we will only be interested in the first type. Figure 2.4 summarises the three different types of contiguity weights. One way of checking whether a particular type of weight is relevant to a particular phenomenon is to calculate the Local Moran Index (LISA). This indicator of spatial auto-correlation makes it possible to check whether a phenomenon is distributed randomly or, on the contrary, according to the spatial interactions between each unit. If it is close to 1 (resp. -1), there is a perfect spatial auto-correlation (resp. dispersion)[6].

[5]For more information on the spatial weight matrix, please refer to this article of *geographicdata.science*
[6]For more information on the Moran Index, please refer to this article of *geographicdata.science*

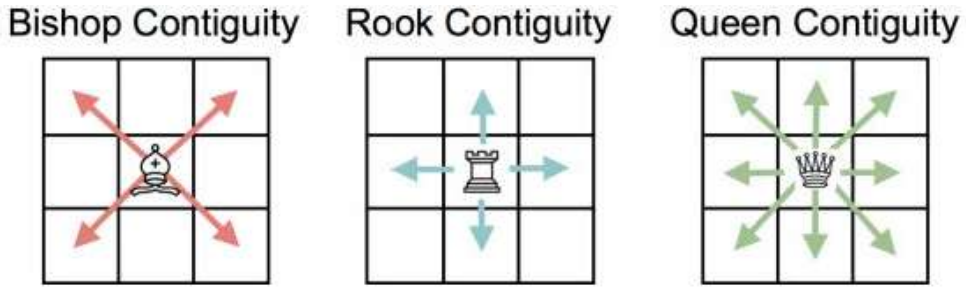


Figure 2.4: Contiguity weights

2.5.2 The different spatial regression models

For econometric reasons, the Ordinary Least Square (OLS) model is not suitable for spatial data. Indeed, the spatial auto-correlation of the residuals (i.e. the dependence between nearby observations) violates the assumptions leading to a loss of OLS efficiency or biased estimators. There may be a spatial lag and/or an auto-correlation of errors (spatial auto-correlation or spatial error in the models) introducing an auto-regressive effect into the model. Therefore, we must rely on econometric models thought for spatial data.

Spatial error Model (SEM) : This model defined by Anselin [10] and Lesage and Pace [11], introduces a spatial lag in the error term.

$$Y_i = \alpha + X_i\beta + u_i \tag{1}$$
$$u_i = \lambda u_{lag-i} + \epsilon_i$$

With X_i the independent variables, α the constant, $u_{lag-i} = \sum_{j \neq i} w_{i,j} u_j$, ϵ_i the error term and $w_{i,j}$ the spatial weight. The parameter λ represents the intensity of the interdependence between the u error terms.

Spatial lag Model (SAR or SLM) : This model developed by Anselin et Bera[12], introduces a spatial lag ρ in the dependent variables.

$$Y_i = \alpha + \rho W Y_{lag-i} + X\beta + \epsilon_i \tag{2}$$

With $Y_{lag-i} = \sum_{j \neq i} w_{i,j} Y_j$, W designs the spatial weight matrix. This model looks at the spatial interdependence with the other variables. The endogenous lag is calculated by running a TWO-SLS regression with the spatial lag of all explanatory variables as the instrument for the endogenous lag.

Spatial AutoRegressive with additional AutoRegressive error structure (SARAR)[7]: This model defined by Kelejian and Prucha [13], combines the two previous approaches.

$$Y = X\beta + \rho WY + \lambda u + \epsilon \tag{3}$$

To implement our models we use the *spreg* library which computes this model (spreg.GM_Combo_Het) using the generalized method of moments (GMM).

Model selection : To select our model we perform a Lagrange-Multiplier (LM) test which successively tests for the presence of spatial lag (robust and non-robust) ($H_0 : p = 0$), the presence of spatial error (robust and non-robust) ($H_0 : \lambda = 0$) and the joint presence of spatial error and spatial lag ($H_0 : p = \lambda = 0$).

[7]It should be noted that there is also the Spatial Durbin Model (SDM) which also combines the two previous approaches but whose implementation is not available on *Python*. The Durbin model is the most widely used and is generally recommended.

2.5.3 Choice of independent variables

As in the article on the city of Utrecht [6] we implemented two models (Model A and B). We run regressions of our aggregate 2SFCA on different socio-economic variables based on *Filosofi* data. The variables we work with are summarized in Table 1 below. Unlike the article, we do not have the percentage of people receiving unemployment benefits or RSA, nor the percentage of migrants, nor the distance to the nearest transport.

Variables	Description	Units	Model
%_soc.minimum	Percentage of households living below the social minimum threshold	%	A,B
%_ $\geq$ _65	Percentage of individuals over 65 years of age	%	A,B
%_ $\leq$ _17	Percentage of individuals under 17 years of age	%	A,B
%_ $\leq$ _*bat*_45	Percentage of dwellings built before 1945	%	B
%_ $\geq$ _*bat*_90	Percentage of dwellings built after 1990	%	B
%_residences	Percentage of collective residences	%	B
mean_income	Mean yearly income per person (sum of living standards windorised/nb of inhabitants)	€/person	B
density	Residents per km^2 (number of inhabitants in the square/(0.2km*0.2km)	pop/km^2	B

Table 1: Variables used in the regression models

2.6 Clustering and analysis with big dimensional space

2.6.1 Clustering

A more sociological point of view is to try to create a typology of neighbourhoods based on the accessibility measurements. To do so, first, we test several clustering methods : MiniBatchKMeans (because the results are similar to KMeans but more interpretable and because *Petite Couronne* analysis depends on many more data than Paris analysis[8] and MeanShift in several set-ups. The idea is to clusterize on the different categories' accessibility we defined earlier, added to the accessibility of housing and social housing. From that we can define the main functions of different neighbourhoods. To conclude, we can try to correlate those function with the socio-demographic variable of the population that inhabits the found clusters.

KMeans algorithm works by randomly selecting k initial centroids (center points of clusters) and assigning each data point to its closest centroid. The centroid is then updated based on the mean of the data points assigned to it, and the process is repeated until the centroids no longer change significantly or a maximum number of iterations is reached. MeanShift algorithm works by iteratively shifting the centroid of a group of data points to the mean of the points within a certain radius, hence the name "mean shift". This process continues until the centroid converges to a point where no further shifts are possible, at which point a cluster is formed.

This first clustering analysis points out the difficulty of choosing the number of clusters. To better understand the composition of the city of Paris and the construction of the clusters, we can use a hierarchical clustering. Here we will use the AgglomerativeClustering method from sklearn [14]. Hierarchical clustering is a type of clustering analysis used to group similar objects or data points together based on their similarity/dissimilarity in a hierarchical manner. The basic idea is to start with each data point as a separate cluster, and then iteratively group the most similar clusters together until all data points belong to a single cluster. This algorithm works like *Russian dolls*, so we can explore the "sub"-clusters of each cluster. The algorithm requires a connectivity matrix, we use the Queen one as defined previously. The result is often represented on a a tree-like structure called a dendrogram, where the branches represent the nested clusters at different levels of similarity, as on Figure 2.5. We will explore the spatial evolution of the clusters. A default of its method is that the obtained clusters are continuous.

[8] Around 2 000 squares for Paris, around 14 000 squares for *Petite Couronne*

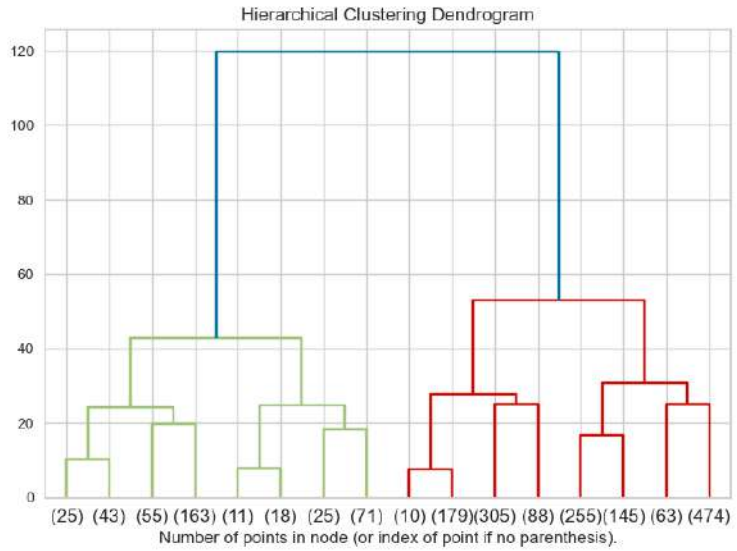


Figure 2.5: Partial dendrogram obtained with sklearn's AgglomerativeClustering on the categories' accessibility, illustrating that AgglomerativeClusstering algorithm : for each iteration, the two closest clusters are merged together until obtaining one cluster

2.6.2 Analysis with big dimension

Another and last point of view is to try to find back the amenity categories that we defined earlier. To do so, we calculate the accessibility for each sub-type of POI (*OSM* tags) that are encompassed in our defined categories (Table 4). From that we have space with 191 dimensions reflecting each type of OSM shops and amenities accessibility. From that big dimensional space, we seek to come back to a smaller one by doing a PCA.

2.6.3 Removal of islands

For the clustering and reduction of dimensions, we exclude the isolated INSPIRE squares that are in the *Bois de Boulogne* and the *Bois de Vincennes*. Even though those space are really relevant to understand Paris city (in particular because they are the biggest green spaces in the whole *Petite Couronne* and because their amenities are dominated by the Parisian bourgeoisie as shown by the Pinçon-Charlot [15]), their amenities are isolated, and so, they are not relevant in a 1-kilometer radius analysis.

3 Results

3.1 Basic descriptive statistics

Figure 3.1 highlights the obvious disparities between neighbourhoods in terms of the concentration of activities. In particular, we can see that the most outlying neighbourhoods, such as Porte de la Chapelle or Eglise du Saint Esprit (Daumesnil district, 12th arrondissement), which are also relatively poor neighbourhoods (particularly Porte de la Chapelle), are those containing the fewest amenities. On the contrary, the districts around Place des Vosges and Garnier Opera, which are more central and more affluent, have a much higher number of amenities.

However, it is also noticeable that other outlying districts (Javel district) and popular ones (Olympiades district) also have a high number of amenities.

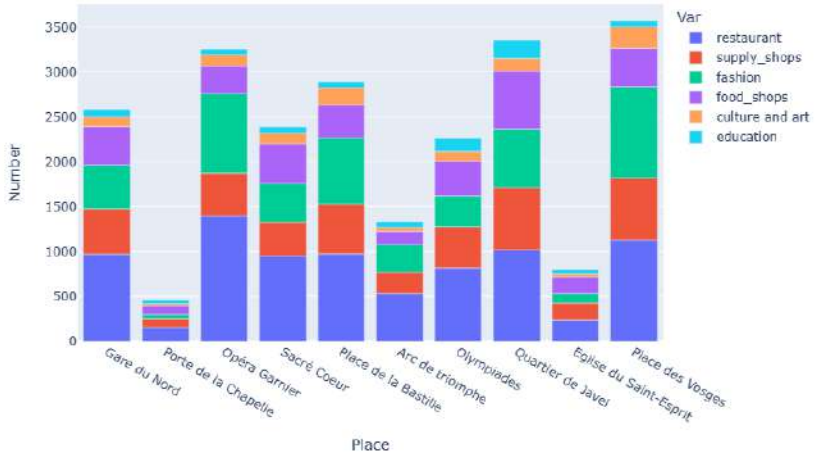
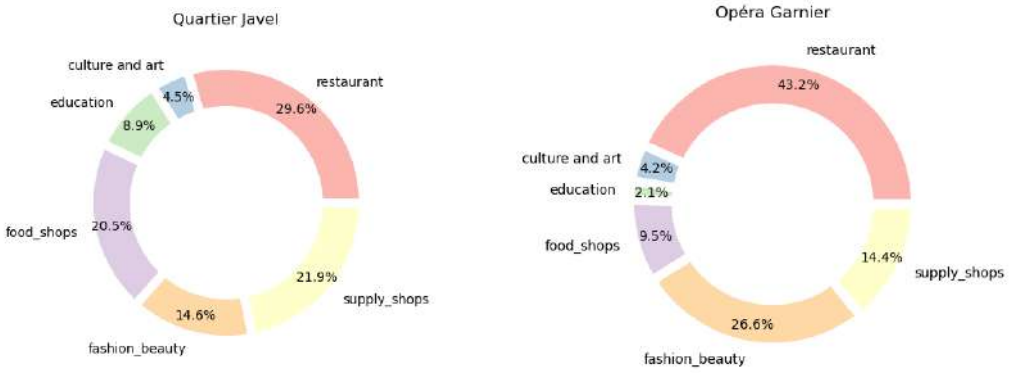


Figure 3.1: Comparison of the composition of 10 Parisian districts



(a) Composition of Javel district

(b) Composition of Opera district

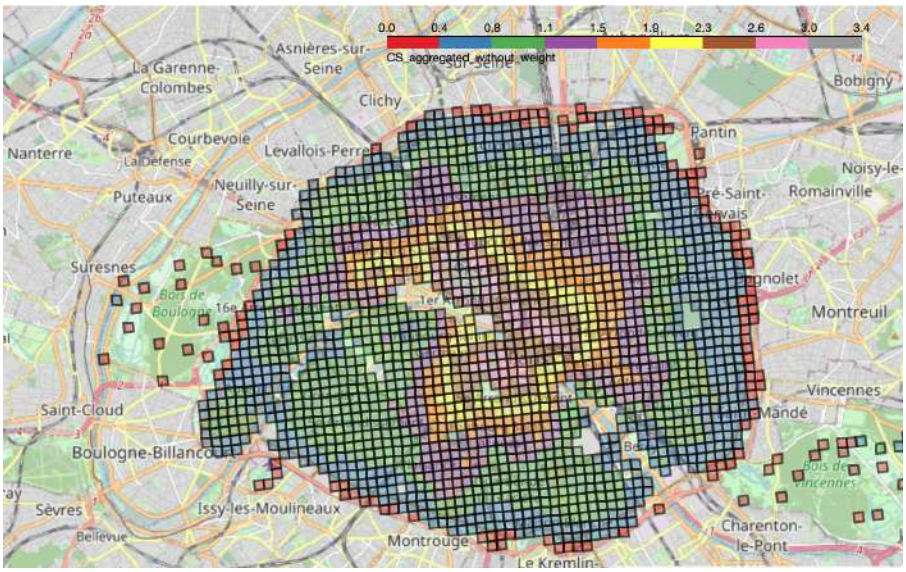
Figure 3.2: Composition of two Parisian districts.

For an equal number of services, Figure 3.2 enables us to distinguish, a touristic district (Garnier Opera district) from a more residential district (Javel district). It is easy to see that more than 50% of the Javel district is composed of daily services (food shops, supply shops, education), compared to 26% in the Garnier Opera district. On the contrary, the Garnier Opera district has almost 70% of luxury services (restaurants, fashion), compared to 44% in the Javel district.

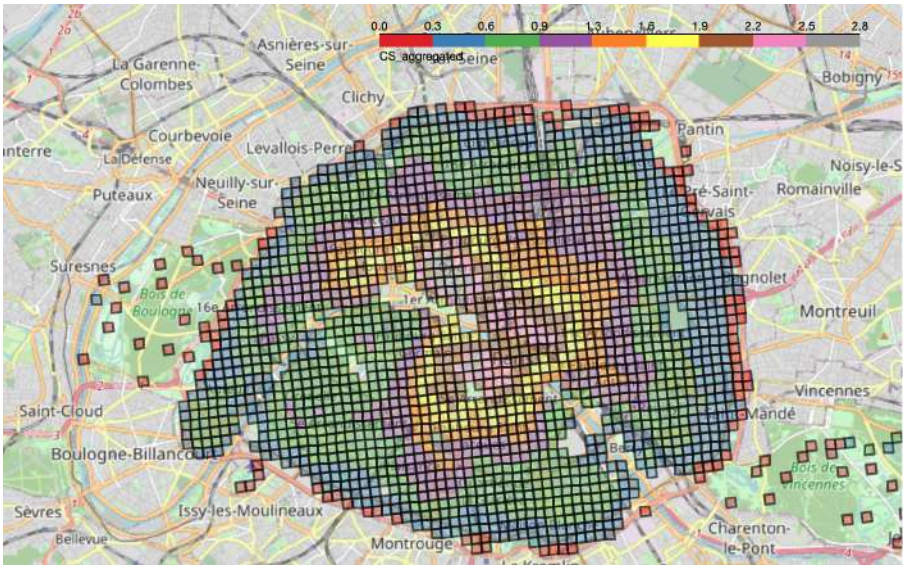
3.2 Aggregated 2SFCA

We created the aggregated accessibility score with and without the weights w_p (Figure 3.3). As can be seen on these two maps, there is little difference between the two methods, but it can be noted that, as expected, the accessibility score with weights tends to emphasize the presence of schools. This is particularly true in the Passy district located in the South-West of Paris in the 16th arrondissement, which has a stronger score with weights because of the presence of many high schools (Saint-Jean de Passy, Saint-Louis de Gonzague, etc.). In the following, we chose to keep the weighted score as it seemed more appropriate, although there are some biases.

In both cases, a concentric circle gradient in the aggregated access to amenities can be observed, starting from three main centres : the Opéra district, Les Halles and the Latin Quarter. This confirms the descriptive statistics for the Opéra district conducted earlier (Figure 3.2b). For the Latin Quarter, this can be explained by the strong presence of both museums and universities/secondary institutions. As for the Halles district, there is a large shopping center as well as a strong economic activity in the surroundings.



(a) without weights



(b) with weights

Figure 3.3: Aggregated 2SFCA with and without amenity weights for Paris

In Figure 3.4, that represents the weighted aggregate score of accessibility to amenities for the entire *Petite Couronne*, we can also notice a concentric circle that starts in Paris and extends over the entire *Petite Couronne* : services are highly concentrated within Paris intra-muros and the nearby suburbs are less endowed with amenities. This is consistent with the 2020 INSEE study [5] that shed light on a strong inequality of access between households living in the centres of large urban areas and those living on the outskirts.

However, it should be noted, that even within the inner suburbs, there is once again intense activity in the center of the towns at the expense of the edges of the towns : this is the case, for example, for Saint-Maur and Rueil-Malmaison, presented in Figure 3.5.

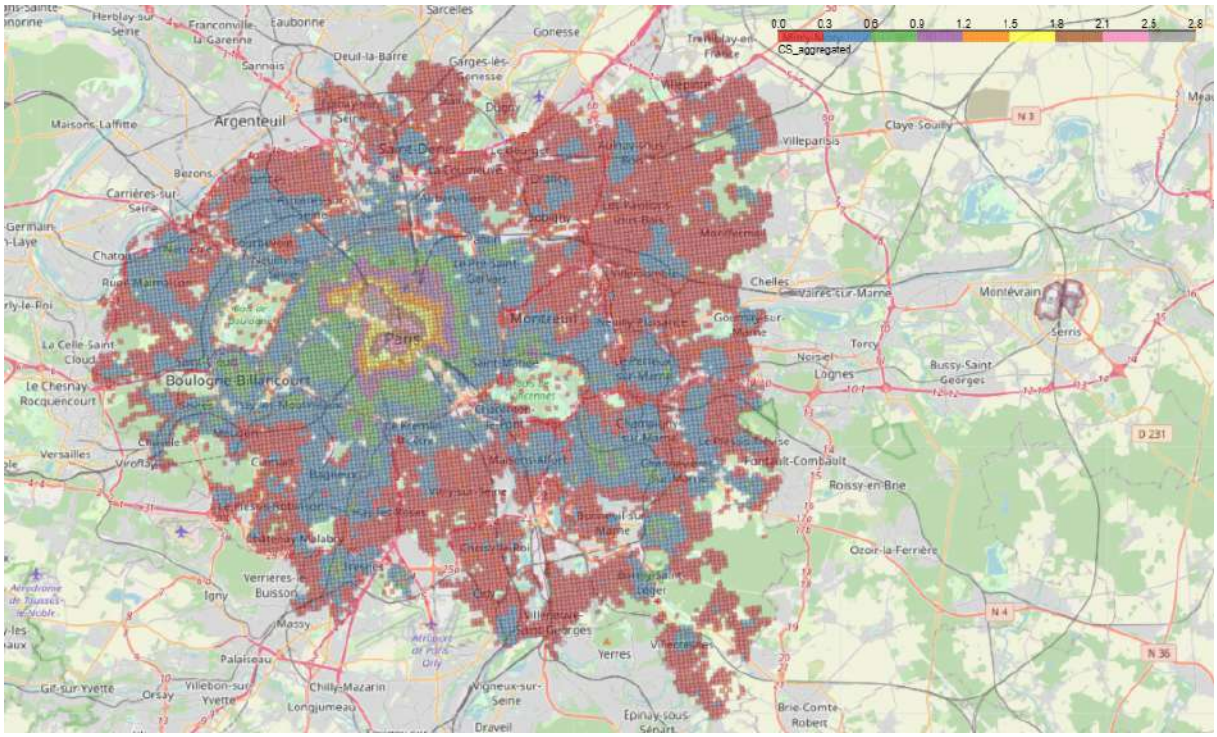


Figure 3.4: Aggregated 2SFCA for the *Petite Couronne*

(a) Rueil Malmaison (b) Saint Maur

Figure 3.5: Zoom of the aggregated 2SFCA for the *Petite Couronne* on two towns

3.3 Regression

Results for Paris

Both the A and B models were built with a queen weight. As a matter of fact, after calculating the local Moran index for the queen weight on the accessibility score, we find an index equal to 0.95 and a p-value equal to 0.01. This means that there is a strong spatial auto-correlation and that the hypothesis of a random distribution of the aggregate accessibility indicator (H_0 : *random distribution*) can be rejected at the 95% threshold. Therefore, we can consider that the queen weight is to be taken into account for our regressions.

After implementing the Lagrange-Multiplier test, we find that the adequate model for our two regressions is the SARAR model (3). We have chosen to take into account heteroskedasticity[9] to avoid errors in the significance of the coefficients. The results are presented in Table 2 (the β values correspond to those of model (3)).

[9]In statistics, heteroskedasticity occurs when the variance of the residuals depends on the value of the variable of interest. For example, the variance of the residuals decreases with the variable of interest.

	Model A			Model B		
Variables	β	St. Error	Sig	β	St. Error	Sig
Constant	0.0757173	0.0349554	**	0.2433732	0.0993165	**
%_soc.minimum	-0.0003339	0.0008628		-0.0015136	0.0011965	
$\%\_ \geq 65$	-0.0008846	0.0004741	*	-0.0007074	0.0005022	
$\%\_ \leq 17$	-0.0024618	0.0006601	***	-0.0027939	0.0006454	***
$\%\_ \leq \_bat\_45$					0.0001288	0.0000798
$\%\_ \geq \_bat\_90$				0.0001431	0.0000856	*
%_residences				-0.0017876	0.0007521	**
mean_income				-0.0000006	0.0000005	
density				0.0000000	0.0000000	***
Model fit:						
Pseudo R^2	0.9859			0.9869		
Spatial Pseudo R^2	0.3715			0.0008		

*** = P $\leq$ 0.01
** = P $\leq$ 0.05
* = P $\leq$ 0.1

Table 2: Spatial regression results for Paris

The results in Model A indicate that the higher the percentage of the population below 17 years old the lower the CS_{15} score. This outcome may seem a little surprising but is consistent with the results found in Utrecht (Table 5). One explanation could be that families prefer to live in less "dynamic" areas like the 15th arrondissement. The negative relationship between the percentage of people over 65 and the CS_{15} score can be explained in the same way but the coefficient is not very significant. As for the percentage of households living below the social minimum, the coefficient is not significant. This indicates that, in Paris, accessibility to essential amenities cannot be explained by the poverty rate of a neighborhood.

The Model B results are similar to Model A for the percentage of the population below 17 years old but the percentage of people over 65, the percentage of buildings built before 1945 and the average income are not significant. For the latter variable, this confirms that economic criteria are not useful to predict the level of accessibility to essential amenities. With regard to the buildings built before 1945, this can be explained by the presence of almost all buildings dating from the 19th century in Paris. This also explains the low significance of the percentage of buildings built after 1990, whose positive relationship with the CS metric could be due to the presence of this type of building in freshly renovated neighbourhoods and thus a bit more dynamic. There is no relationship between our metric and the density of the neighbourhood in contrast to what the INSEE study of 2017 seemed to indicate [5], particularly because of the homogeneity of density in Paris intra-muros.

The different results show that Paris seems to be a rather egalitarian city in terms of accessibility since the economic criteria are not significant. Nevertheless, these results must be taken with a lot of nuance since we do not have the quality of the amenities and do not take into account their importance (through their number of visits).

Results for the *Petite Couronne*

As for the tests performed on Paris, we obtain for the *Petite Couronne* a Moran index that is very close to 1 (0.97) and a very low p-value (0.001) which leads us to choose a queen weight for these regressions. Moreover, the p-values obtained during the Lagrange-Multiplier tests are very close to 0, so the appropriate econometric model is also the SARAR model. We also decide to take into account a possible heteroskedasticity.

The results of the regressions are given in Table 3. As with the regressions for Paris intra-muros, young people under 17 have relatively less access to services than adults under 65 ceteris paribus. Moreover, while in Paris there was no inequality in accessibility to services correlated to socio-economic criteria, here we observe that the coefficient of *%_soc.minimum* is significant and negative : the more households living below the minimum

	Model A			Model B		
Variables	β	St.Error	Sig.	β	St. Error	Sig.
Constant	-0.0157894	0.0045639	***	-0.0042972	0.0047322	
%_soc.minimum	0.0003369	0.0001184	***	-0.0004332	0.0001574	***
$\%\_ \geq 65$	-0.0001949	0.0000805	**	-0.0000086	0.0000846	
$\%\_ \leq 17$	0.0002000	0.0001081	*	-0.0003271	0.0000933	***
$\%\_ \leq \_bat\_45$				0.0002047	0.0000245	***
$\%\_ \geq \_bat\_90$				0.0000511	0.0000168	***
%_residences				0.0002356	0.0000254	***
mean_income				-0.0000000	0.0000001	
density				0.0000005	0.0000000	***
Model fit:						
Pseudo R^2	0.9838			0.9834		
Spatial Pseudo R^2	omitted			0.4363		

*** = P $\leq$ 0.01
** = P $\leq$ 0.05
* = P $\leq$ 0.1

Table 3: Spatial regression results for the *Petite Couronne*

social threshold, the lower the accessibility score. As a matter of fact, we can assume that Paris is a relatively homogeneous city in terms of socio-economic level because the poorest households will tend to move to the outskirts of Paris as the cost of living is cheaper there. Performing the analysis on the *Petite Couronne* and not only on Paris therefore allows us to take into account certain socioeconomic inequalities that are not noticeable within Paris.

In addition, the coefficients on the percentage of buildings built before 1945 and after 1990 are significant : geographic units containing more buildings built before 1945 and after 1990 have a better accessibility score to services than units containing buildings built between 1945 and 1990. This could be a consequence of the urban policies implemented during the *Trente Glorieuses* period[10], aiming to respond to the post-war housing crisis : as with the Courant plan (1953) these policies, based on the principle of zoning[11], resulted in the construction of large residential areas on the outskirts of Paris far from shops and services. Moreover, we observe that geographic units containing more buildings built after 1990 have a relatively poorer accessibility score than units containing pre-war buildings : this could reflect the disparities between Paris and its suburbs observed in Figure 3.4, as the majority of intra-muros buildings were built under Haussmann.

It can also be noted that the coefficients of *%_residences* and *density* are significant : the more houses rather than apartments are located in an area, the lower the accessibility score. In addition, the more densely populated an area is, the higher the accessibility score.

3.4 Clustering

First clustering : MeanShift, centre vs border

First, as we do not know the number of expected clusters we try the clustering method MeanShift. We obtain two clusters as shown on Figure 3.6.

Those two clusters oppose the narrow centre of Paris to a very broad border. We can see on Figure 4.1 that the central cluster has a better accessibility for every type of amenity apart from the housing where they do not differ and for social housing where the external cluster is better. It seems to corroborate the fact that the centre of Paris is really determined by its touristic and commercial functions, which dominate the rest.

[10]Les *Trente Glorieuses* was a thirty-year period of economic growth in France between 1945 and 1975, following the end of the Second World War.

[11]Zoning is a method of urban planning in which a municipality or other level of government divides the land into zones, each of which serves a specific need.

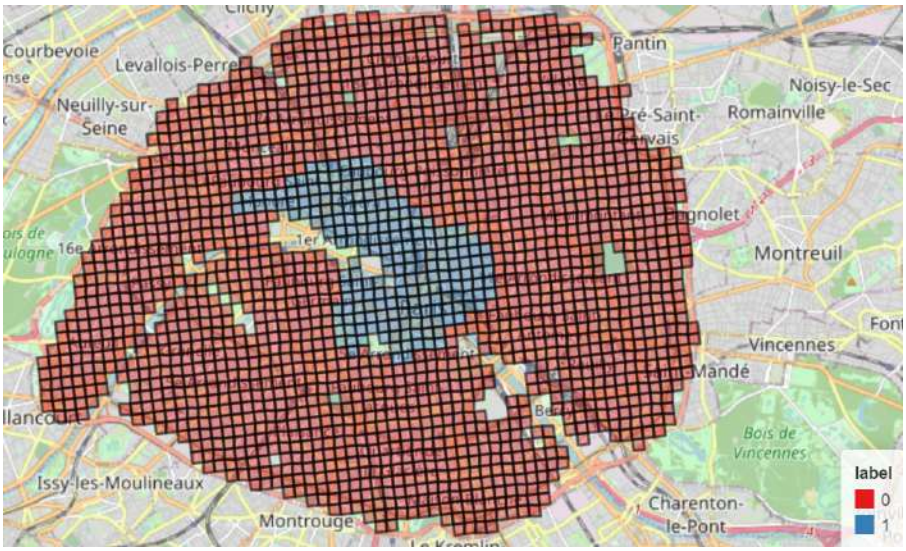
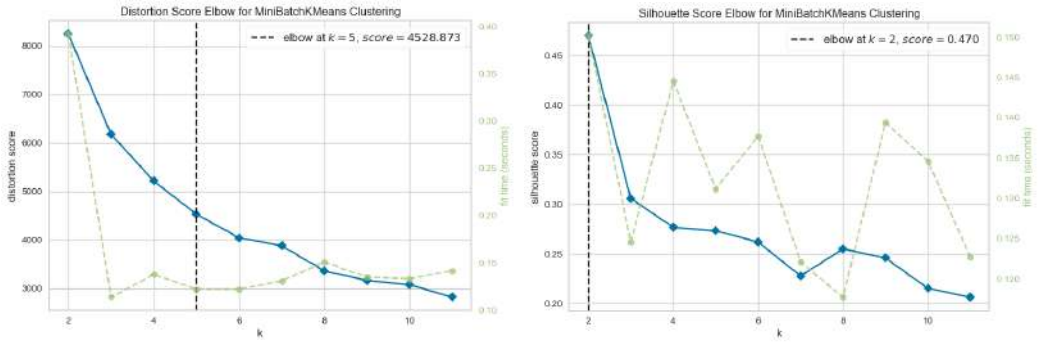


Figure 3.6: Map of the clusters made using the MeanShift methods on the accessibility measurement of each category

Second clustering : MiniBatchKMeans, border, inhabitants neighbourhoods, commercial and cultural neighbourhoods

However, to better understand the composition of Paris neighbourhoods, it is not satisfactory. To complete the analysis, we decide to do a MiniBatchKMeans clustering with more than two clusters. To choose the number of clusters, we carry out an elbow method with the distortion score as shown on Figure 3.7a : it suggests taking 5 clusters. To confirm that number of clusters, we calculate the silhouette score as shown on Figure 3.7b. We encounter a slight problem : the silhouette score suggests taking two clusters. At first, it confirms what we find with the MeanShift method : Paris is driven by an opposition between a touristic, commercial centre and the others residential neighbourhoods. It is also possible that the silhouette score is being wronged by the fact that one of the cluster is an *"edge"* cluster, as we will see.



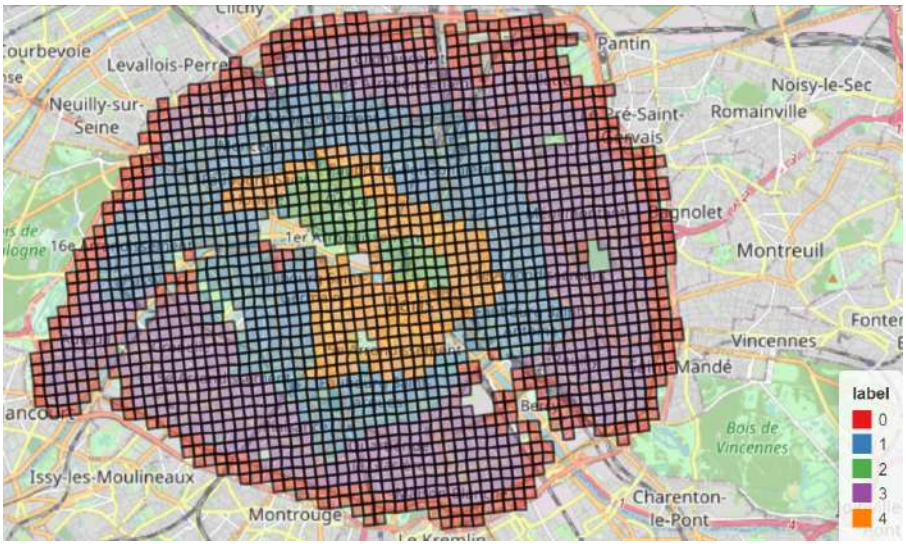
(a) Distortion score

(b) Silhouette score

Figure 3.7: Elbow method for two scores (blue) for the clustering on the categories accessibility. Green is the computation time.

We still decide to work with 5 clusters and the MiniBatchKMeans method, because the results were reliable and easier to interpret. Those findings are shown on Map 3.8a. Globally we see the previous centre cluster has been divided into two clusters whereas the other previous cluster has been divided into 3 clusters.

The green cluster (2) corresponds to the commercial function previously identified. Indeed, with a closer look (Figure 3.8b), we see that it matches *les Halles*, well known for its shopping centre, and the *Opéra* neighbourhood, well known for its shops. The orange cluster (4) seems more centred on cultural amenities



(a) Map of the five MiniBatchKMeans clusters

(b) Zoomed Map

Figure 3.8: Map of the five KMean clusters, based on the different categories accessibility

compared to the green one, that's why it includes *Le Marais* and *Saint Germain-des-Prés*, the neighbourhoods with plenty of art galleries, and the *Quartier Latin*, which is not only an educational centre but also a cultural one, as shown by Mariotte, Maytraud and Wolf [16]. Those impressions are quantitatively confirmed as shown on Figure 4.2. It shows the clusters (2) and (4) are standing out for the overall amenities' accessibility. But cluster (2) is more inclined towards fashion, food shops, restaurants and supply shops whereas the cluster (4) has more accessibility to culture and arts POI and to education.

The external cluster is being divided into 3 clusters : the red (0) one which corresponds to the border, the blue (1) one and the violet (3) one.

The red one is clearly standing out. Its geographical location is on the border of Paris, and it is the poorest in terms of accessibility apart for social housing. Multiple explanations can be given to that cluster. At first, it could be an artefact due to the arbitrary limit of the data to Paris itself. However, historically the first real development of social housing in Paris - the so-called *Habitat Bon Marché*, literally cheap habitat - have been build on those areas at the edge of Paris, as shown in a study by the *Atelier Parisien d'Urbanisme (APUR)* [17]. This historical fact - also described by Stébé in his last book [18] - could explain the importance of social housing on the border of Paris (excluding the edges of the *Bois de Boulogne*). Also, Paris and its suburb have been separated by the *périphérique*, an important ring road around Paris. So the lack of amenities can also be the physical effect of that ring road. However, this cluster can also be found on some edges of the Seine and on some edges of the *cimetière du Père Lachaise* in the North. To conclude on that particular matter, the red (0) cluster is probably the mixed consequences of being on the edges (of Paris, of the Seine, of parks) and the history of Paris planning where the first social housing have been in place on those edges. This history has probably socio-economic consequences, as the poorer communities have been isolated from the rest of the amenities in comparison to the blue and violet clusters.

The comparison between the blue (1) and the violet one (3) is easier. Apart from social housing and housing, the blue one has better accessibility in every aspect, even though the blue cluster (1) is still behind the green (2) and orange (4) clusters.

Last remark, all the commented differences are statistically significant as we ran for each one of the commented difference, a t-test, and the differences were significant at the 95% level.

Now, let's take a look at the socio-demographic description of each cluster, as shown in the different violin plots of Figure 4.3 (a violin plot is comparable to a box plot, it just adds a representation of the distribution).

The red (0) cluster is clearly the poorest cluster by looking at the percentage of poor household and the mean of standard of living, followed by the violet (3) one. This poverty is probably partially due to the fact that the red cluster includes the most vulnerable families, like the single-parent households and large families. We can also see that this cluster is home to mostly renters. So there are probably different causal effects between having a lot of social housing, having a lot of renters, having more socially difficult situations and having more poverty.

The violet one seems to host a similar amount of poor people than the other cluster, but the mean revenue is lower. This cluster is probably home to a small middle class. We can also look at the fact that it is hosting the denser squares, which is corroborated by the geography of the cluster : the 13th, the 15th, the 19th, the 20th are the most populated *arrondissements*.

By contrast, the green and orange clusters are the less dense ones, and the richest ones. In particular, the orange one is hosting the richest square in all Paris.

Third clustering : AgglomerativeClustering, "sub-clusters"

As explained in the method section, we apply the AgglomerativeClustering. On Figure 4.4, we represented the *Russian dolls* like maps we obtained by increasing the number of clusters for the Agglomerative Clustering method. It is a spatial representation of the dendrogram figure (Figure 2.5) : for each number of clusters, we represent the clusters on the map of Paris, so we see the same *Russian dolls* decomposition but spatially which is easier to read than the dendrogram.

This series of clustering confirms several previous facts. The main opposition is between the touristic centre of Paris, and the larger border. We can see that on the map with 2 clusters. Secondly, looking at the maps with four, five and six clusters the progressive apparition of the different type of centres.

The MiniBatchKMeans and the Agglmerative clustering methods opposed themselves on certain specific features. First, the border of Paris is separated into two and one of the two obtained clusters seems closer to the rest of the 16th *arrondissement* than of Paris. It is probably due to the fact that the edges of the *Bois de Boulogne* has been a privileged space for the Parisian bourgeoisie as explained in [15] which in particular led to the absence of *Habitat Bon Marché* and more generally of social housing as seen in the APUR's study on HLM [17]. More importantly, there is never an arbitral edge effect, contrary to what we saw with KMeans and MiniBatchKMeans where some squares on the edges of the *Seine* or the *Cimetière du Père Lachaise* were included in the border cluster. The lack of edge effect is probably due to how AgglomerativeClustering works, since it merges smaller clusters into bigger one using the Queen continuity matrix that we defined. Secondly, the most equipped clusters outside the centre have been separated into 2 with the 16th *Arrondissement* showing its specificity. Without detailing the full results, shown on Figure 4.7 and on map 4.8, we find similar results as with the MiniBatchKMeans method : clusters (0) and (5) are the centres. (0) is the commercial one, (5) the education/culture/arts one. (2) is a middle ground clearly better equipped than the rest of the clusters outside the centre, it places the same role as the (1) in the MiniBatchKMeans method. For the rest of Paris, the cluster (1) and (4) represent the residential clusters but as shown on figure 4.9, (1) is clearly more popular. The clusters (3) and (6) represent the border of Paris, but there is now a difference in the social housing accessibility : (6) holds the less equipped in social housing of the border of Paris neighbourhoods, in coherence with the Pinçon-Charlot analysis [15] and APUR's map [17]. It is possible that the (6) cluster is just an edge effect.

Overall results of the analysis through clusterization

To conclude, the overall cluster analysis allows us to better understand the composition of Paris at different levels of details.

At first, Paris is driven by a dichotomy between a really well-equipped centre and the rest of Paris. Secondly, this centre itself is composed of a commercial centre on *Rive Droite*, an educational, cultural, artistic centre on *Rive Gauche*. Those two centres are surrounded by really well-equipped neighbourhoods, but less specialized.

The rest of Paris, can be decomposed into 3 big spaces : a border (itself decomposed in 2), a residential and popular spaces (for instance the 15th, 13th, 19th, 20th) which are just a bit better equipped than the border and have globally the same number of social housing, and last but not least a rich space which includes the 7th, 8th, 9th, and 16th *Arrondissements* and consists of the most well-equipped neighbourhoods outside the very centre.

This decomposition in the accessibility space has no trivial explanation into our socio-demographic space. One of the most determinant variable seems to be the accessibility of social housing, which drives half of specificity of the border and the specificity of the popular neighbourhoods. The less equipped clusters ((1), (3), (4), (6) in the last specification) are mostly popular apart from the cluster (4). The demography seems also an important factor : the central clusters are characterized by their low percentage of minors, which corroborates what we found with the regression. People inhabit and have children mostly outside the really well-equipped centres, it is particularly pronounced for the border clusters.

Lastly, as shown on the last map of Figure 4.4, we can continue to subdivide to understand at a smaller scale Paris city, in a *Russian dolls* idea.

3.5 Clustering on *Petite Couronne*

We carry out a clustering on accessibility measures on the aggregated categories we defined first (and not on the whole big dimensional space, for computation time reasons). We use a MiniBatchKMeans with 5 clusters (suggested by the elbow method). MiniBatchKMeans is preferable because it is faster and also because it does not aggregate continuous squares like our specified AgglomerativeClustering method. This particular fact allows to identify similar neighbourhoods in different cities.

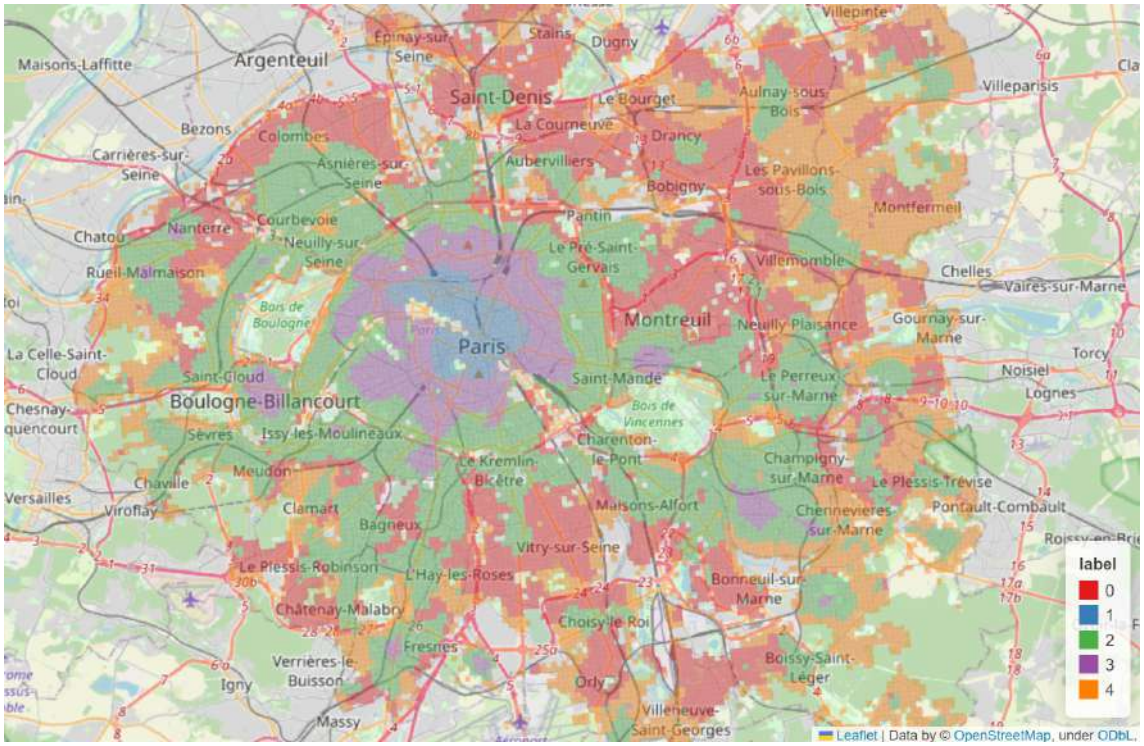


Figure 3.9: Map of the clusters made using MiniBatchKMeans method on the accessibility measurements at the scale of the whole *Petite Couronne*

We find the clusters as shown on Map 3.9. Quickly, we find back the same centre (in blue, cluster n°1) as in the cluster analysis of Paris, showing that it does stand out, no other centre seems to match it. Some other city centres can be identified in the violet (n°3) cluster, which is the cluster of Paris "middle ground" neighbourhoods. Globally, the same city centres are standing out as in the map obtained with the aggregated accessibility method 3.4 and highlighted in Figures 3.5a,3.5b. They are for instance Reuil-Malmaison, Saint-Maur, Montrouge, Boulogne-Billancourt. The *Seine-Saint-Denis* at the north of the *Petite Couronne* seems to

be the only territory without any well-equipped city centre. In this part of the region, there is the larger part of less equipped clusters (orange and red). The orange one is the less equipped of all clusters (even in terms of housing and social housing !). This cluster is probably representative of a suburbia (*banlieue pavillonaire*) with low density of housing and amenities accessible by walking. On the contrary, the green and red clusters are denser, with more housing and social housing, and are more compatible with the 15-minutes city because more amenities are accessible by walking.

3.6 Big dimensional space analysis : PCA results

As said in the method sections, we do a PCA on the big dimensional space containing the accessibility measurements for every OSM shops and amenities points of interests. We obtain the explained variance as presented on 3.10. We see that the first PC is popping out : it explains almost 30% of the variance on its own, while the other principal components are explaining less than 10% each.

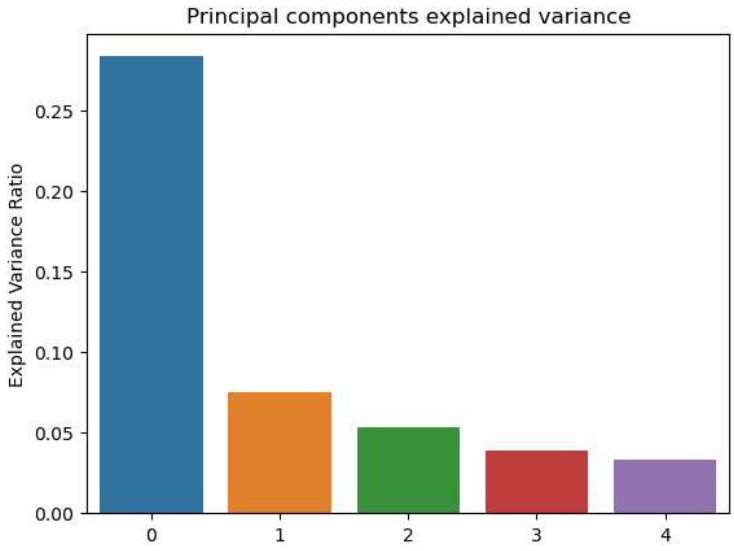


Figure 3.10: Percentage of explained variances of the first five principal components

In the first principal component, 80% of the original space features have a positive contribution and only 10% have a negative contribution, which are linked mostly to car reparation as shown on Figure 4.10b. We can understand that the first dimension as an overall accessibility measurement as shown on Map 4.10a, equivalent to what we found in the overall accessibility measurement, as shown on Figure 3.3. This result shows once again that Paris has a confounding variable which is the distance to its centre, and more probably to its centres that are *Opéra* and *Les Halles*. It also suggests that car reparation amenities are oriented towards the *Périphérique* and thus towards the border of Paris.

The second principal component seems to be more linked to fashion, clothing and luxury as shown on Figure 4.11b, we can note in particular the influence of the department stores, jewellery, watches. We can also note that green groceries store and frozen foods have both really negative coefficients (not shown in the figure), which tends to convince ourselves that the 2nd component is a luxury shop vs common groceries store measurement. We can clearly identify that the maximum of this measurement is now located between *Opéra* and *Les Champs Elysée* as shown on Map 4.11a. It is quite remarkable that the second driven effect in Paris is luxury.

Lastly, the third component is linked to education (university), culture (book stores, library, cinema, games) and arts as shown on Figure 4.12b, its maximum is reached around the *Quartier Latin* as shown on map 4.12a. It is also important to note that the minimum is reached around *les Champs Elysée*, where there are a lot of luxury shops.

To conclude, first, with the PCA, we find back a measurement of overall accessibility, driven by the proximity to the centre of Paris (in reality 2 centres) which confirms the results found with the regressions and the clustering. Secondly, according to the second and third principal components, luxury and education/culture/arts are two important and separate drivers of Paris neighbourhoods. Those two categories correspond to our findings

with the different really well-equipped central clusters, but we should not over interpret them, as the second principal component is explaining only 7.5% of the variance and the third, only 5.3%.

4 Discussion

Although the analysis carried out in this project leads to relevant and satisfactory results, it has some shortcuts and flaws that we must point out.

First, the use of OSM data may have constrained our analyses. Because we lack some information, the list of categories established in section 2.2 may be incomplete : we omitted, for example, access to medical care, access to green spaces and sports infrastructures or the possibility of finding a job near home. Moreover, OSM data do not provide us with qualitative information on the amenities : we would have refined our analysis by taking into account the quality of the services. Furthermore, more socio-economic data would be great to improve our regressions and the description of our clusters. It would be possible with new iterations of the Filosofi data set but it has not been released to the public. Lastly, we could have taken advantage of the network road data to estimate if each square is walkable or not.

Then, the accessibility measurement set in Section 2.4 can be improved. The 2FCA method can be made more complex on three aspects. First, we can incorporate S_j (resp. R_j) in the probability that i visits j. It is called 3SFCA (resp. fixed point SFCA method). In each case, the idea is that people take into account the offer in j when they go there and not only the distance. This reflects that the offer creates partially the demand. The second improvement is to adjust the demand function in regard to the socio-demographic composition of each square. For instance, we can expect that the youth "consume" more schools than old ones. Those adjustments can be made, for instance, using the mean consumption of each demographic group at the national scale as used in INSEE studies ([4], [5]). As we have the demographic pyramid of each square, we could construct for each category a consumption indicator based on national accountability data. Lastly, the distance could be computed using the distance on the road network and not the distance as crow flies. But for a computational purpose, it is not really possible with the quantity of data here (due to the complexity of the graph exploration times the complexity of the number of square). To really understand how Paris is working as an accessible city, it would probably also be better to compute a 15-minutes area with public transport taken into account and not only a 1-km radius. It would probably change the accessibility of many neighbourhoods along the metro lines. Lastly, we do not take into account the demand by non-inhabitants like tourists. It is quite a problem as it neglects the demand in touristic neighbourhoods like the centre of Paris, so we are probably overestimating the availability of for instance the central restaurants. However, our measurement, if imperfect, is representative of the spatial and demographic composition of amenities in Paris. The same analysis with a second data source (like the SIRENE database, or the APUR collection of shops) could be a way to check this bias.

On the analysis side, the PCA first dimensions are only representative of 40% of the variance : there are still a lot of Paris Morphology that cannot be explained by the chosen OSM POIs. Nevertheless, we checked the robustness of the methods by confronting multiple ones (comparing regressions between each others, clustering methods between each other) and the several approaches give us a few numbers of consistent results across methods and approaches.

Conclusion

The purpose of our project was to develop a statistical model based on different types of points-of-interest (POIs) in order to analyse the composition of big cities. After defining amenity categories, we relied on the 2SFCA aggregate accessibility measure, which takes into account supply and demand for each type of service.

The analysis shed light on a strong dichotomy of access within Paris : on the one hand, there is a very commercial center in the heart of Paris and, on the other hand, more residential neighborhoods on the outskirts of the city, which have less access to services. Composed of the surroundings of Les Halles, the Latin Quarter, the Opera district and the Champs Elysées, the dynamic center fulfills two distinct functions : while the activity located in the Champs Elysées, the Opera district and les Halles is more inclined towards luxury shops, food shops, restaurants and supply shops, the Latin Quarter is focused on cultural and education-related activities. The analysis also reveals inequalities in access to services correlated to socio-economic criteria : the poorest neighborhoods, which are located on the edge off the city, are the least endowed with amenities. We can also

note disparities of access according to the age of the individuals as the youngest live in the neighborhoods with fewer services.

Extending the analysis to the *Petite Couronne* allowed us to better understand the inequalities in access to services within the *Ile-de-France* region. As a matter of fact, it quickly becomes apparent that services are highly concentrated within Paris, to the detriment of its nearby peripheries. In addition, we observe in the same way as before, that even in the cities of the inner suburbs, accessibility to services is once again high in the heart of these towns to the detriment of their edges. Moreover, the extension to the *Petite Couronne* highlights socio-economic inequalities in access to facilities that were less visible for Paris intra-muros : households living below the minimum social threshold and those living in areas built between 1945 and 1990 have poor access to amenities.

References

[1] Carlos Moreno et al. "Introducing the "15-Minute City": Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities". In: *Smart Cities* 4.1 (2021), pp. 93–111. ISSN: 2624-6511. DOI: 10.3390/smartcities4010006. URL: https://www.mdpi.com/2624-6511/4/1/6.

[2] Michel Pinçon and Monique Pinçon-Charlot. *Sociologie de Paris*. "Repères". La Découverte, 2014. ISBN: 9782707156105.

[3] Anne Clerval. *Paris sans le peuple. La gentrification de la capitale*. La Découverte, 2016. ISBN: 9782707191021. URL: https://www-cairn-info.proxy.rubens.ens.fr/paris-sans-le-peuple--9782707191021.htm.

[4] Arthur Cazaubiel and Clément Cohen. "Des commerces moins accessibles dans les espaces périurbains". In: *Les Entreprises en France* ,coll. "Insee Références" (2020).

[5] Corentin Trevien. "Commerces et inégalités territoriales". In: *Les Entreprises en France* ,coll. "Insee Références" (2017).

[6] Elizabeth Knap et al. "A composite X-minute city cycling accessibility metric and its role in assessing spatial and socioeconomic inequalities – A case study in Utrecht, the Netherlands". In: *Journal of Urban Mobility* 3 (Dec. 1, 2023). ISSN: 2667-0917. DOI: 10.1016/j.urbmob.2022.100043.

[7] Jean-François Girres and Guillaume Touya. "Quality Assessment of the French OpenStreetMap Dataset". In: *Transactions in GIS* 14.4 (2010), pp. 435–459. ISSN: 1467-9671. DOI: 10.1111/j.1467-9671.2010.01203.x.

[8] Wei Luo and Fahui Wang. "Measures of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region". In: *Environment and Planning B: Planning and Design* 30.6 (2003), pp. 865–884. DOI: 10.1068/b29120.

[9] Muriel BARLET et al. "L'accessibilité potentielle localisée (APL) : une nouvelle mesure de l'accessibilité aux médecins généralistes libéraux". In: *ÉTUDES et RÉSULTATS* 795 (Mar. 2012). URL: https://drees.solidarites-sante.gouv.fr/sites/default/files/2020-10/er795.pdf.

[10] L Anselin. *Spatial Econometrics: Methods and Models*. Studies in Operational Regional Science. Springer Netherlands, 1988. ISBN: 978-90-247-3735-2. URL: https://books.google.fr/books?id=3dPIXClv4YYC.

[11] Jame LeSage and R.Kelly Pace. *Introduction to Spatial Econometrics Statistics*. Series:Statistics:TextbooksandMonographs. BocaRaton,Florida:CRCPress, 2009. ISBN: 978-1-4200- 6424-.

[12] Luc Anselin and Anil K Bera. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics". English (US). In: *Handbook of Applied Economic Statistics*. Ed. by Aman Ullah. CRC Press, 1998, pp. 237–290. ISBN: 9780824701291. DOI: 10.1201/9781482269901-36.

[13] Harry H. Kelejian and Ingmar R. Prucha. "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances". In: *Journal of Real Estate Finance and Economics* 17:1 (1988).

[14] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[15] Michel Pinçon and Monique Pinçon-Charlot. *Les ghettos du Gotha: Comment la bourgeoisie défend ses espaces*. Editions Zones, 2013.

[16] Apur. *Atlas des lieux culturels du Grand Paris*. Directeur et directrice de la publication: Alexandre Labasse and Patricia Pelloux. Étude réalisée par: Clément Mariotte, Flora Maytraud, Martin Wolf. Sous la direction de: Patricia Pelloux. Cartographie et traitement statistique: Morad Khaloua. Datavisualisation réalisée par: Anaïs Moreau. Photos et illustrations: Apur sauf mention contraire. Mise en page: Florent Bruneau. Retrieved from http://www.apur.org. 2023.

[17] Jean-François Arènes et al. *Cartographie du logement social à Paris - Situation au 1er janvier 2019*. Sous la direction de Stéphanie Jankel. Photos et illustrations : Apur sauf mention contraire. Mise en page : Apur. 2019. URL: http://www.apur.org/fr/nos-travaux/cartographie-logement-social-paris-situation-au-1er-janvier-2019.

[18] Jean-Marc Stébé. *Le logement social en France. (1789 à nos jours)*. Que sais-je ? Presses Universitaires de France, 2022. ISBN: 9782715413320. URL: https://www.cairn.info/le-logement-social-en-france--9782715413320.htm.

Appendix

Category	OSM key	OSM tags
Restaurant	amenities	bar, biergarten, cafe, fast_food, food_court, ice_cream, pub, restaurant
Culture and art	amenities	arts_centre, cinema, conference_centre, events_venue, library, music_school, planetarium, public_bookcase, studio, theatre, toy_library
	shops	anime, antiques, art, books, camera, collector, craft, frame, games, model, musical_instrument, music, photo, ticket, trophy, video, video_games
Education	amenities	college, kindergarten, school, university
Food shops	shops	alcohol, bakery, beverages, brewing_supplies, butcher, cheese, chocolate, coffee, confectionery, convenience, dairy, deli, farm, frozen_food, greengrocer, ice_cream, pasta, pastry, seafood, spices, tea, wine, water, supermarket
Fashion and beauty	shops	bag, boutique, beauty, clothes, cosmetics, erotic, fabric, fashion_accessories, hairdresser, hairdresser_supply, jewelry, leather, massage,perfumery, sewing, shoes, tailor, tattoo, watches, wool
Supply shops	shops	agrarian, appliance, atv, baby_goods, bathroom_furnishing, bed, bicycle, boat, bookmaker, candles, cannabis, car, caravan, car_parts, carpet, car_repair, charity, chemist, computer, copy-shop, curtain, department_store, do-it-yourself, doors, dry_cleaning, e-cigarette, electrical, electronics, energy, fireplace, fuel, fishing, flooring, florist, fuel, funeral_directors, furniture, garden_centre, garden_furniture, gas, general, gift, glaziery, golf, groundskeeping, hardware, health_food, hearing_aids, herbalist, hifi, household_linen, housewar, hunting, insurance, interior_decoration, jetski, kiosk, kitchen, laundry, lighting, locksmith, lottery, mall, medical_supply, military_surplus, mobile_phone, money_lender, motorcycle, newsagent, nutrition_supplements, optician, outdoor, outpost, paint, party, pawnbroker, pest_control, pet, pet_grooming, pyrotechnics, radiotechnics, religion, scuba_diving, security, ski, snowmobile, sports, stationery, storage_rental, swimming_pool, telecommunication, tiles, tobacco, toys, trade, trailer, travel_agency, tyres, vaccum_cleaner, weapons, window_blind

Table 4: OSM tags selected in each category

For further details on OSM tags, please refer to the OSM amenity and shop wikis.

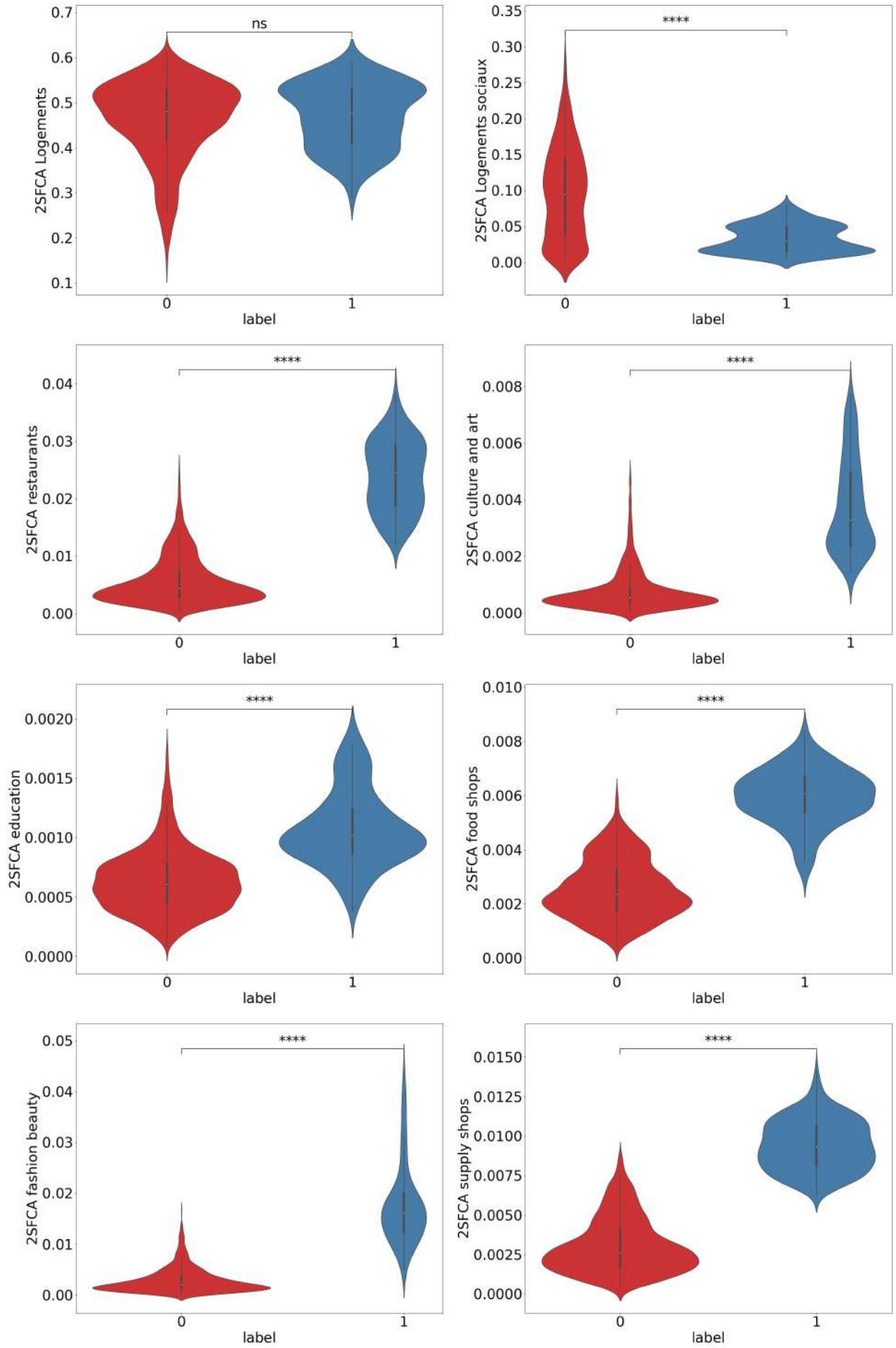


Figure 4.1: Violin plots of the MeanShift clusters for the different accessibility scores, with t-tests. $* : p \leq 0.05, ** : p \leq 0.01, *** : p \leq 0.001, **** : p \leq 10^{-4}$

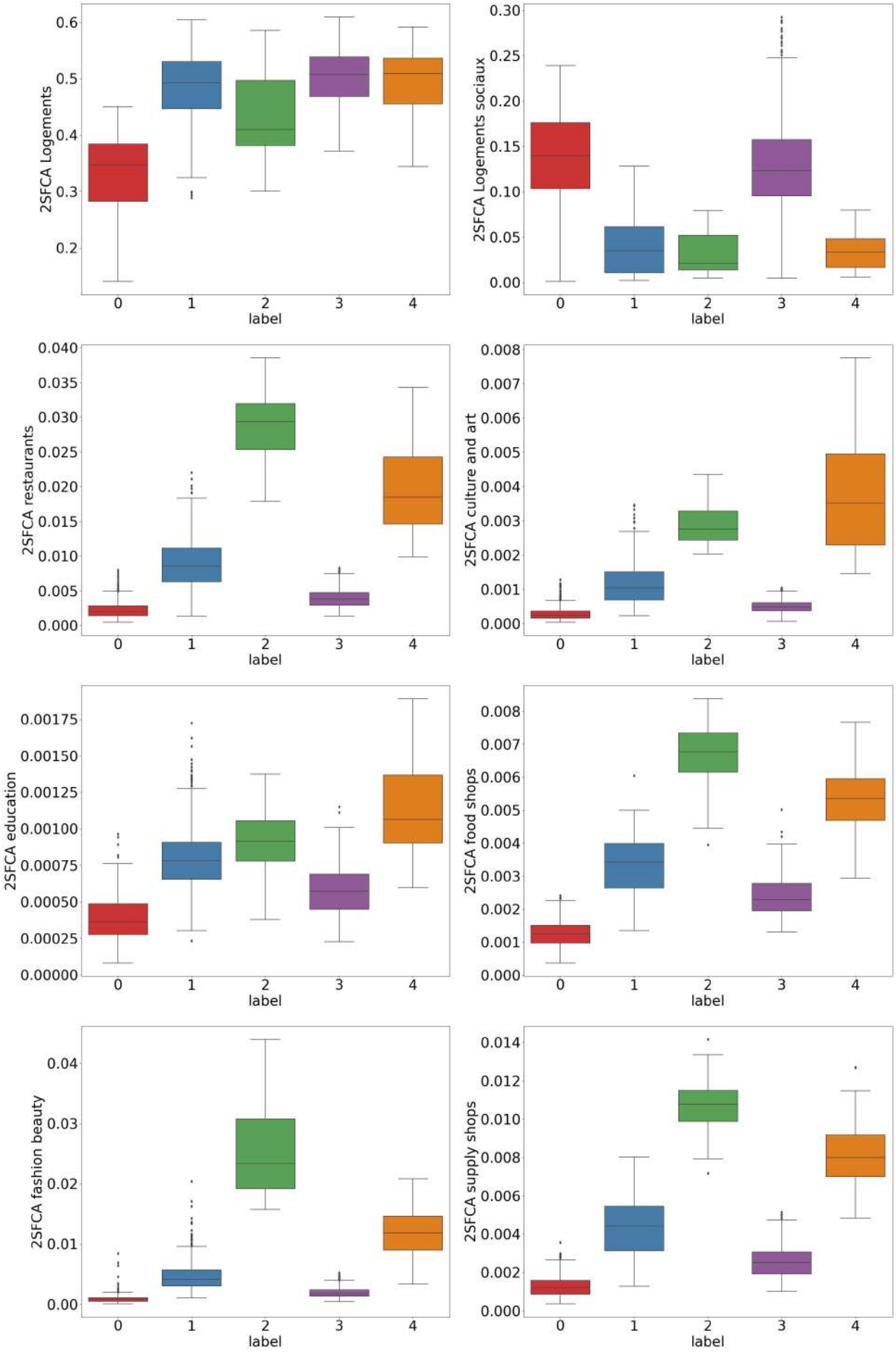


Figure 4.2: Box plots of the MiniBatchKMeans clusters for the different accessibility scores

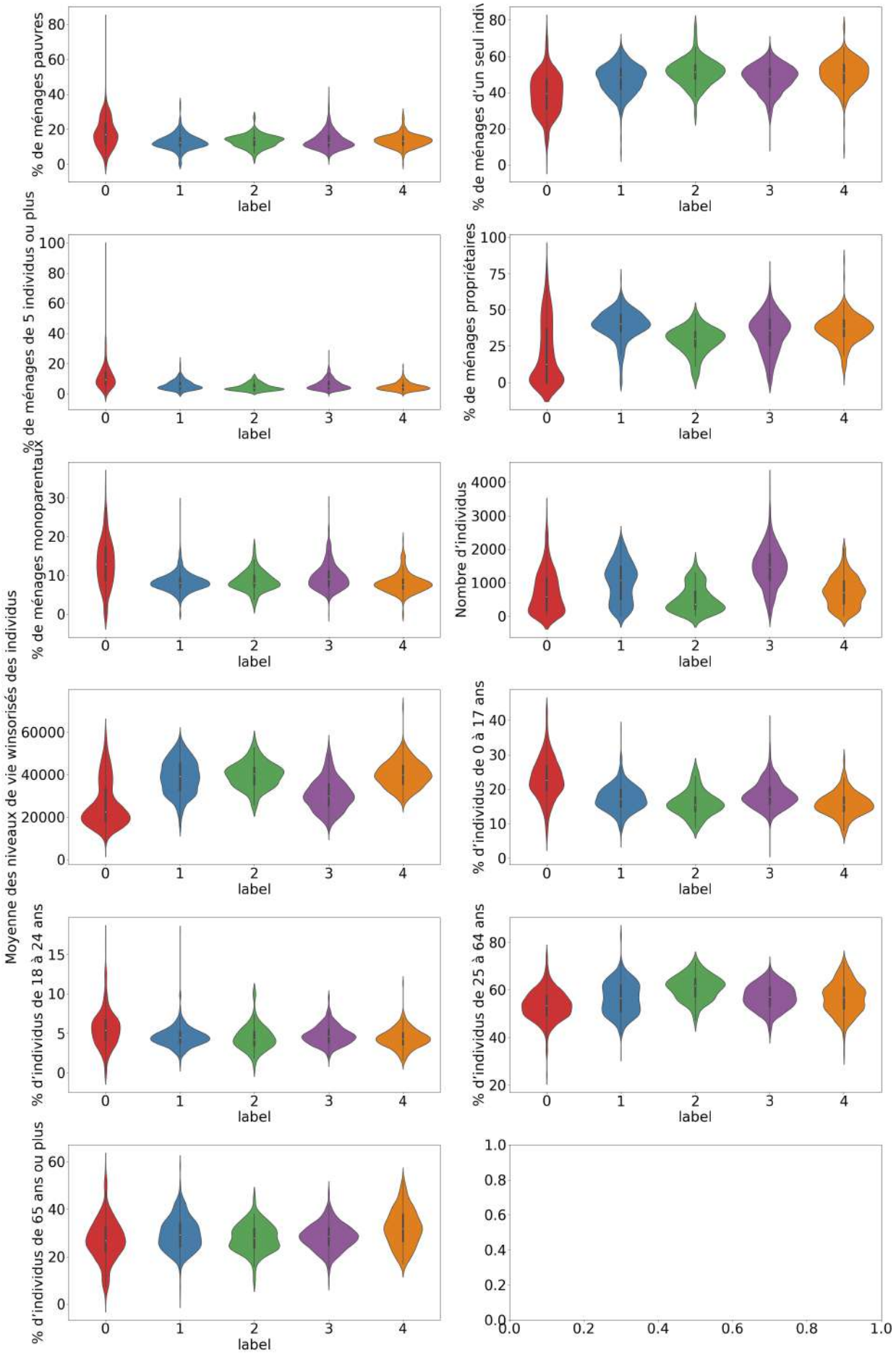


Figure 4.3: Violin plots of the MiniBatchKMeans clusters for different socio-economic variables

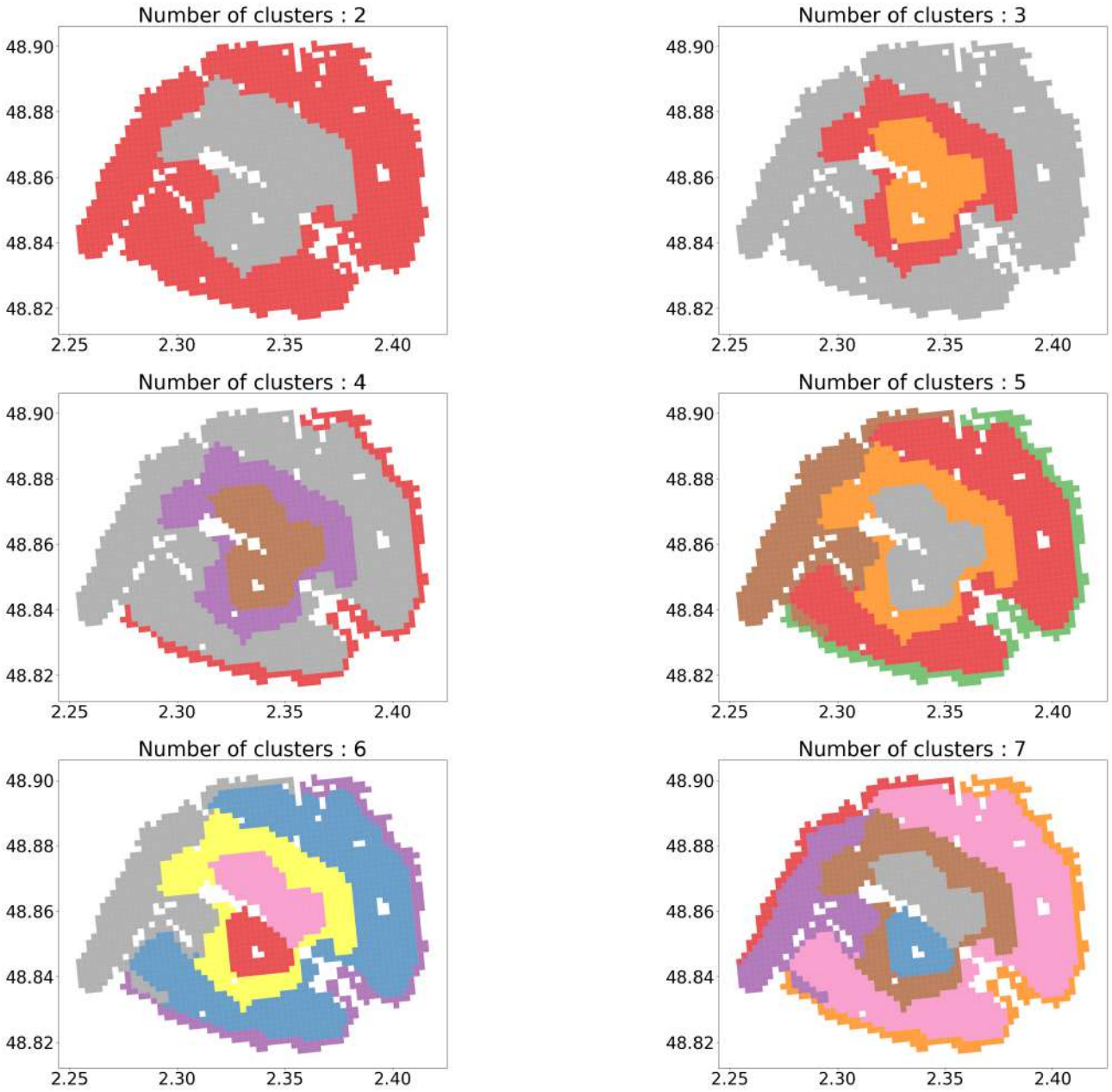


Figure 4.4: Paris *Russian dolls* maps of the different clustering obtained with the AgglomerativeClustering method with an increasing number of clusters. The last map is the chosen one for our analysis, also shown on 4.8

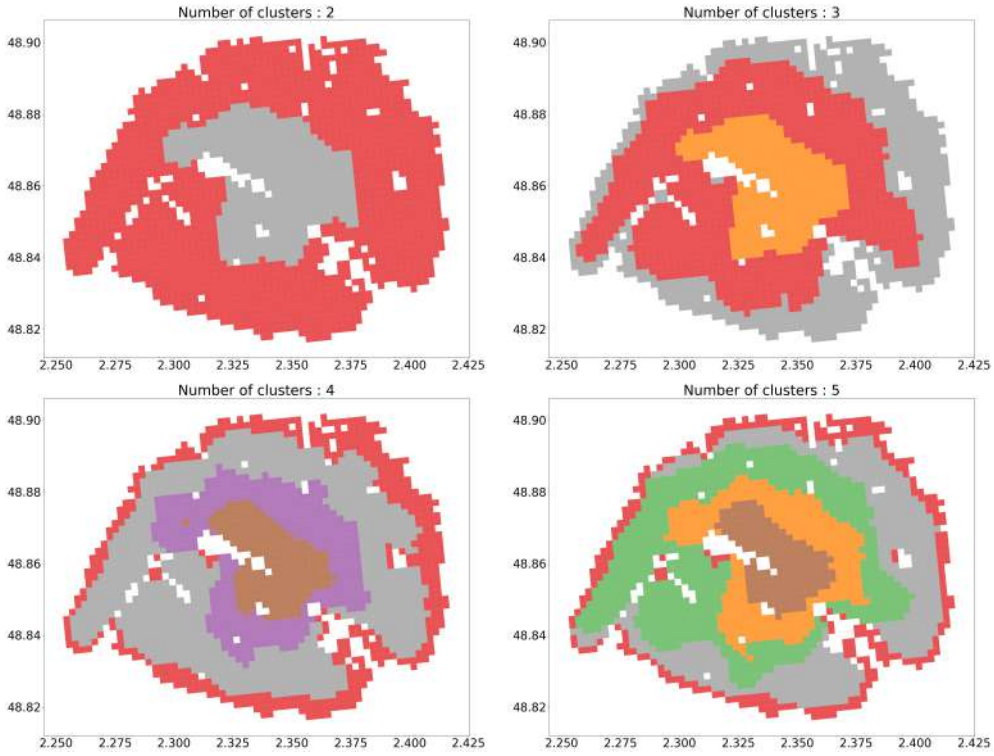


Figure 4.5: Paris *Russian dolls* maps of the different clustering obtained with the KMeans method with an increasing number of clusters

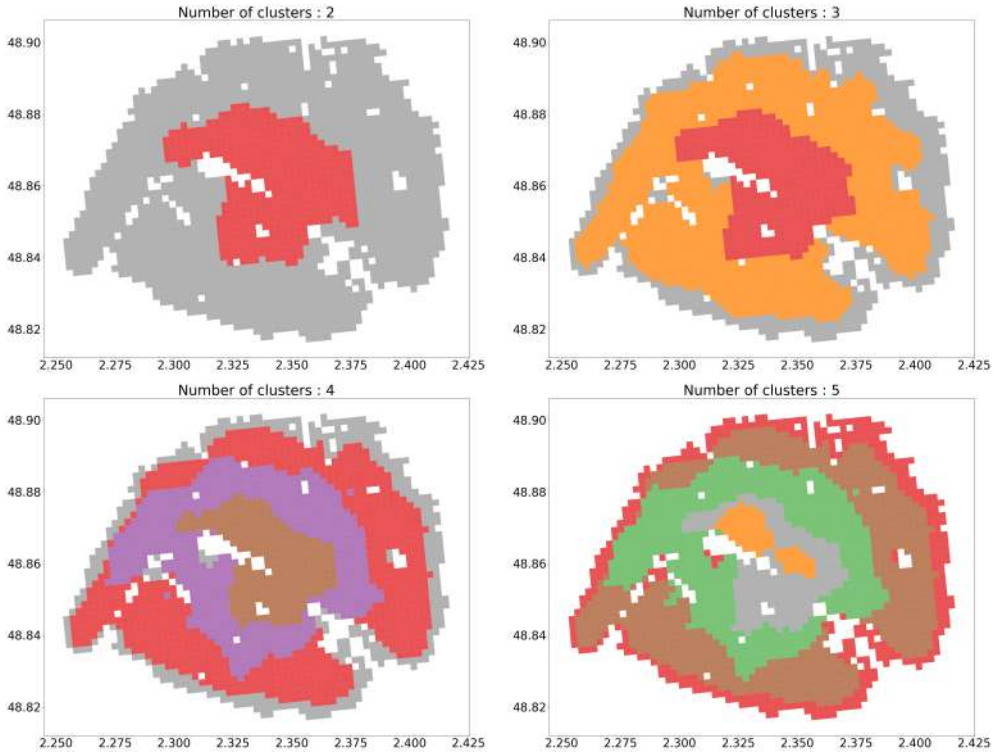


Figure 4.6: Paris *Russian dolls* maps of the different clustering obtained with the MiniBatchKMeans method with an increasing number of clusters

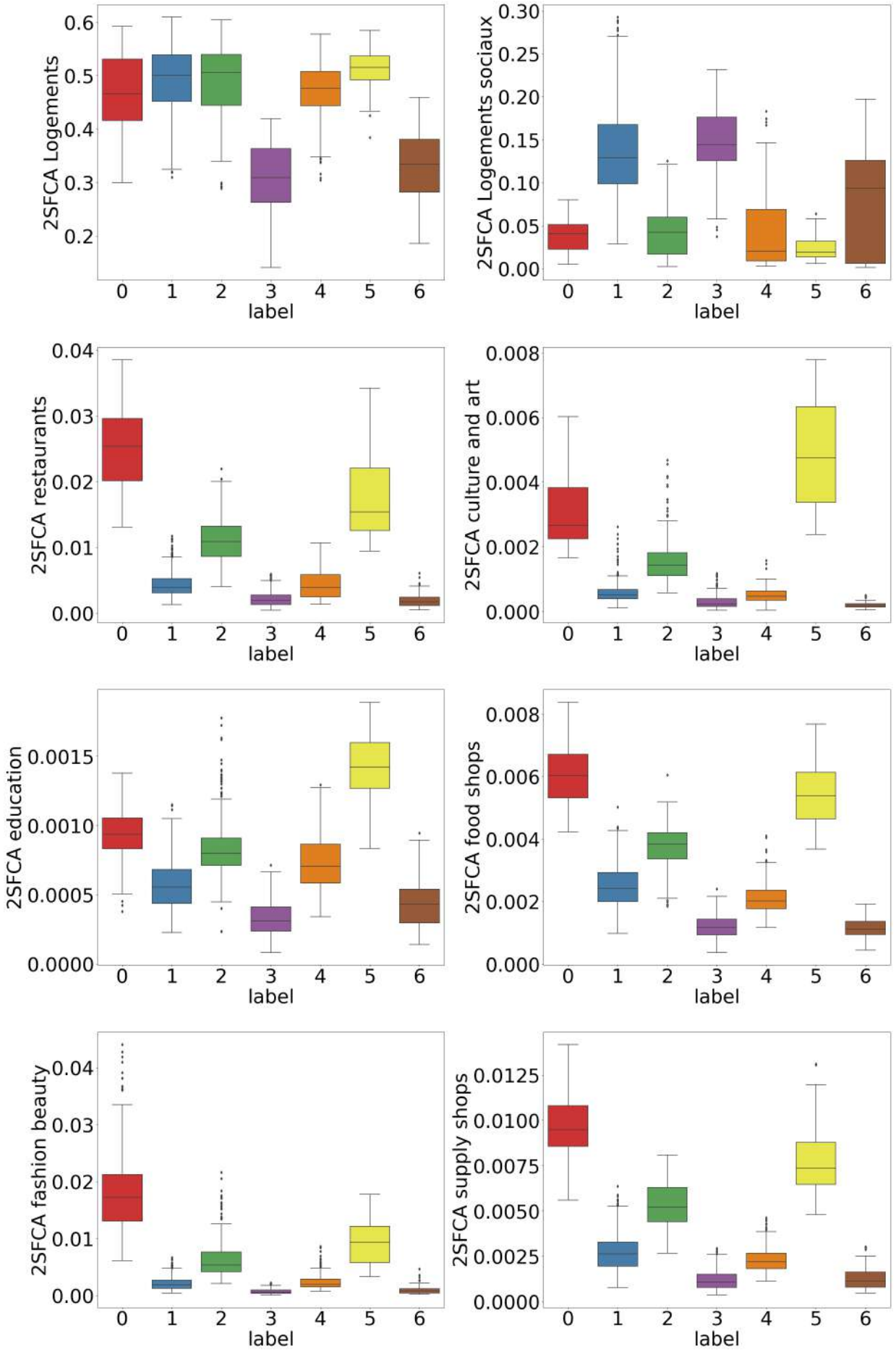


Figure 4.7: Box plot of AgglomerativeClustering method with 7 clusters for the different accessibility score

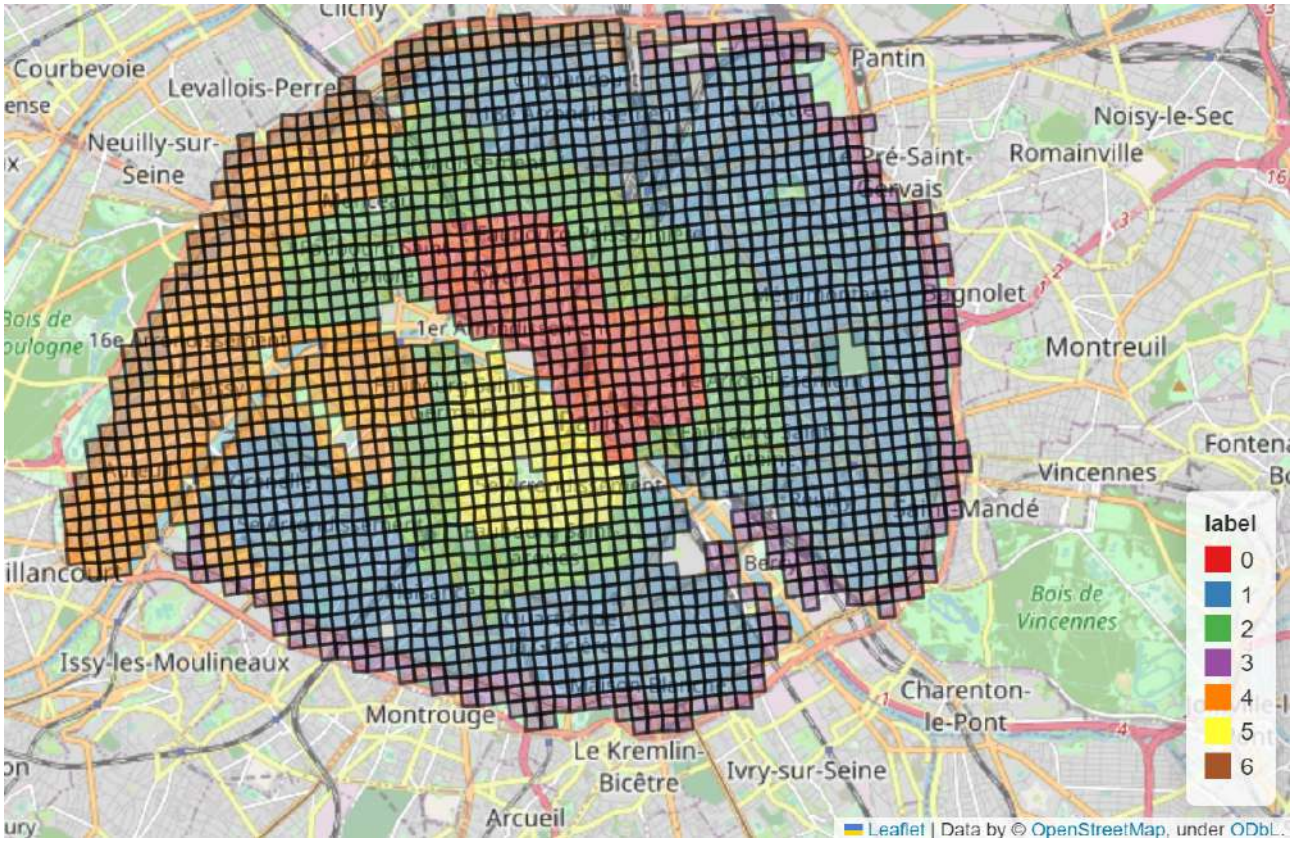


Figure 4.8: Map of AgglomerativeClustering method with 7 clusters

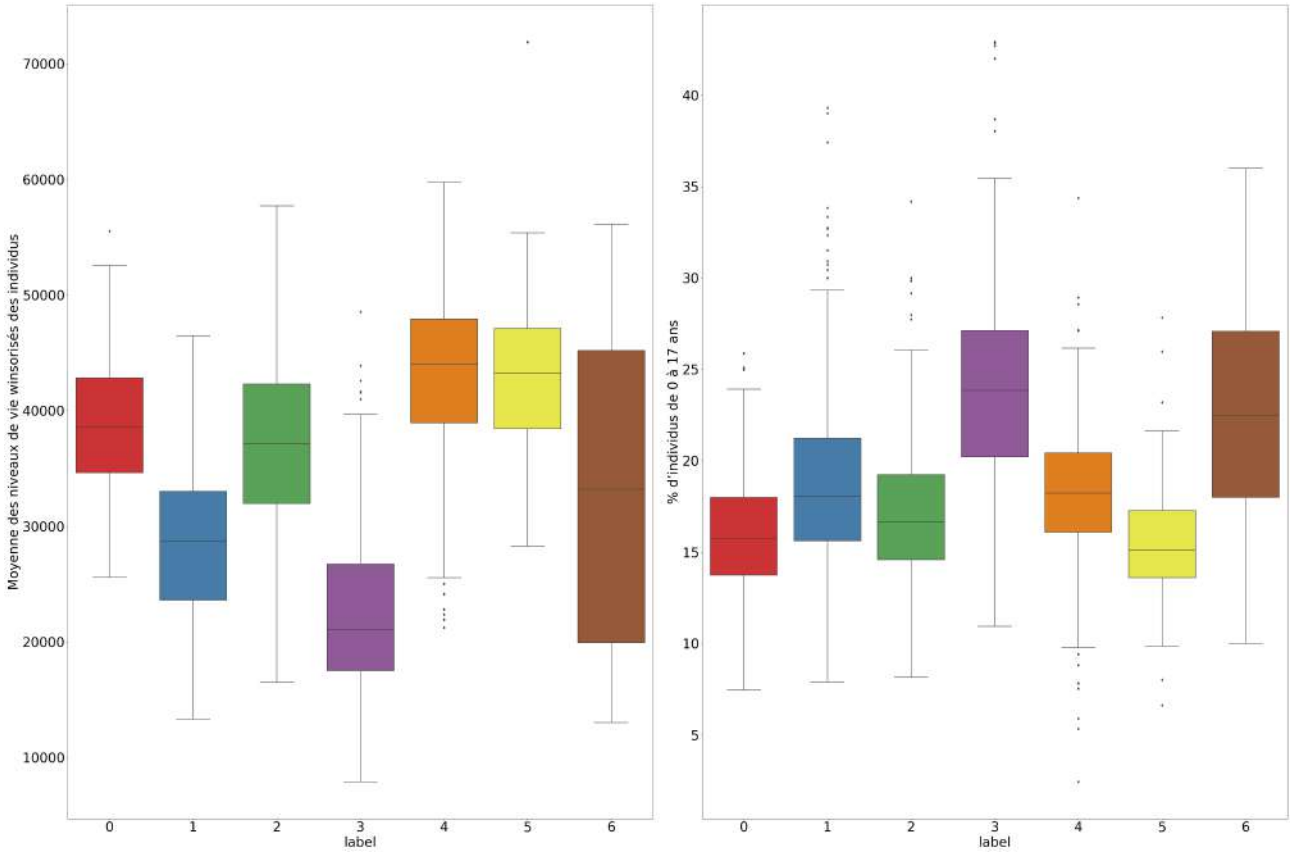
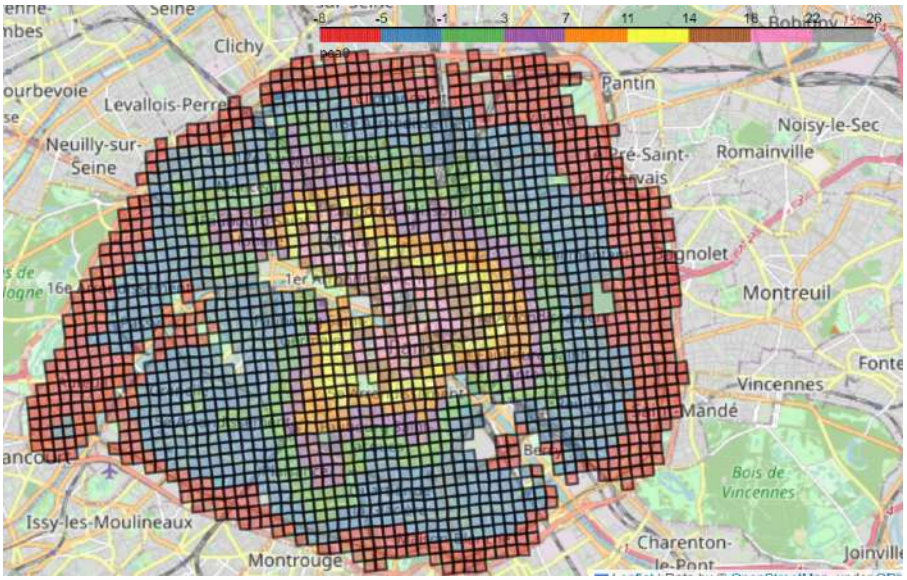


Figure 4.9: Box plots of the AgglomerativeClustering clusters for the mean income and the number of minors

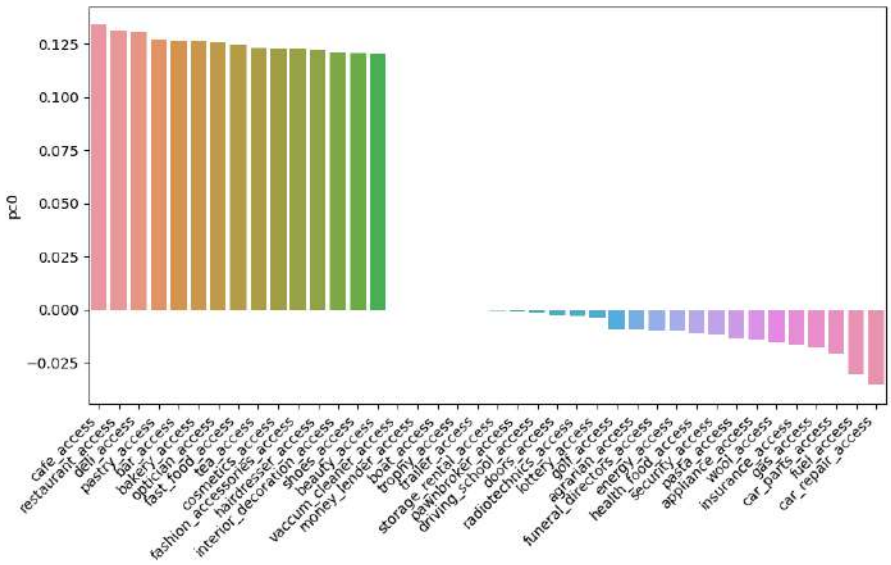
	Model A			Model B		
Variables	β	St.Error	Sig.	β	St.Error	Sig.
Constant	0.201	0.041	***	0.096	0.046	*
% soc. minimum	0.014	0.004	***	0.005	0.005	
% welfare	0.394	0.148	**	0.302	0.151	*
% $\geq$ 65 years	-0.034	0.097		0.045	0.099	
% $\leq$ 15 years	-0.520	0.121	***	-0.369	0.140	**
% immigrants	-0.112	0.073		-0.157	0.075	*
D train station	-0.043	0.010	***	-0.028	0.010	**
% $\leq$ 1945				0.010	0.003	***
% $\geq$ 1994				0.006	0.003	*
% apartments				0.075	0.036	*
Income				0.005	0.073	
Pop .density				0.083	0.025	***
Rho (lag)	1.446	1.467	***	1.384	1.658	***
Spatial dependency residuals						
Global Moran's I z-value, p-value	1.323	0.186		1.201	0.230	
Model fit						
LL	113.90			135.69		
AIC	-211.79			-245.38		
R2	0.362			0.488		

*** = P $\leq$ 0.001
** = P $\leq$ 0.01
* =P $\leq$ 0.05

Table 5: Knap and al's[6] regression results

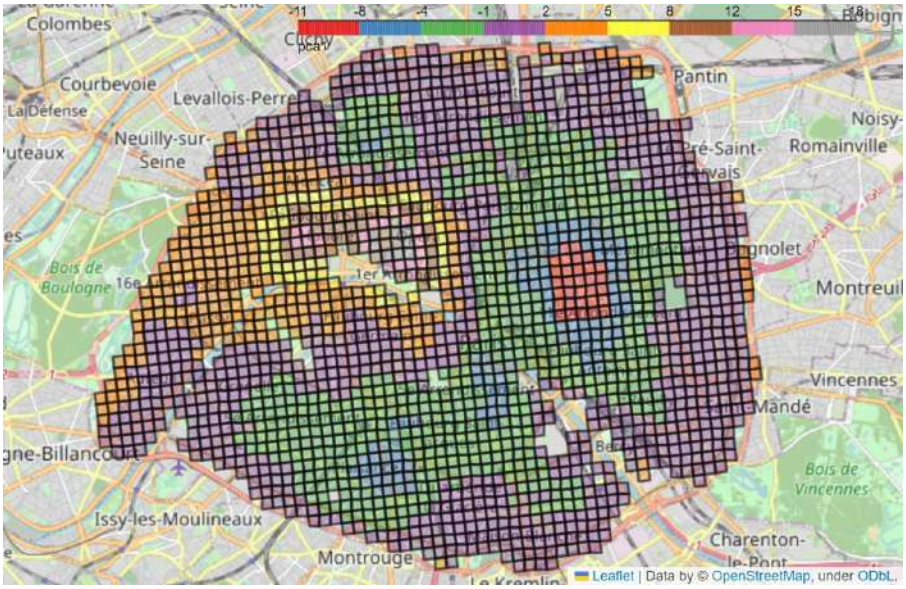


(a) Map of first principal component

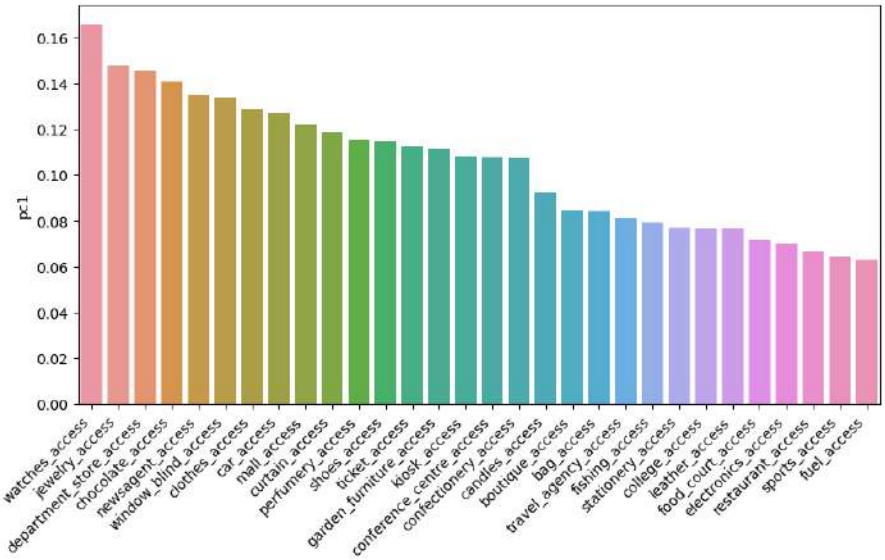


(b) Top15 of the features with the biggest coefficients in the first PC, bottom25 of the features with the most negative coefficient in the first PC

Figure 4.10: First principal component

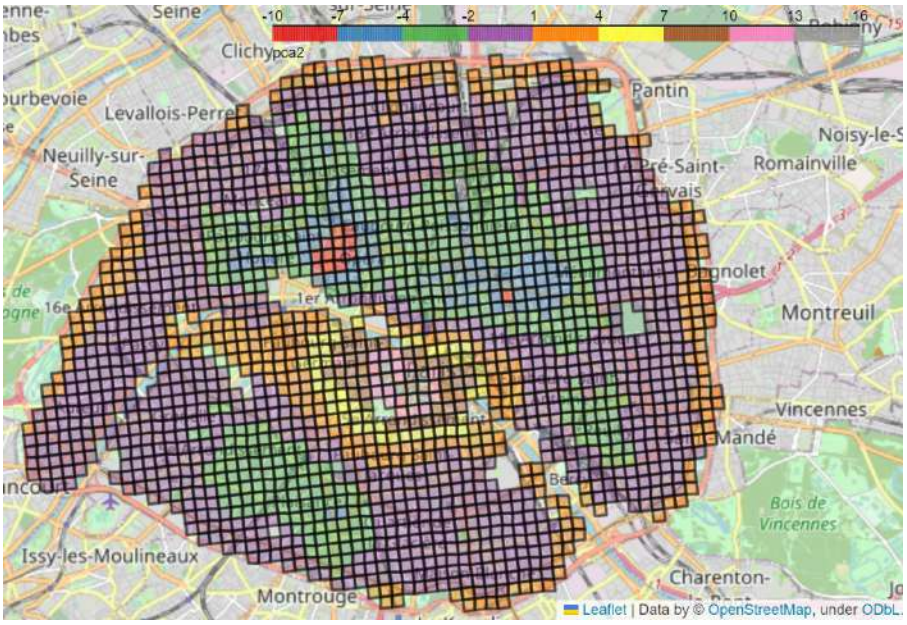


(a) Map of second principal component

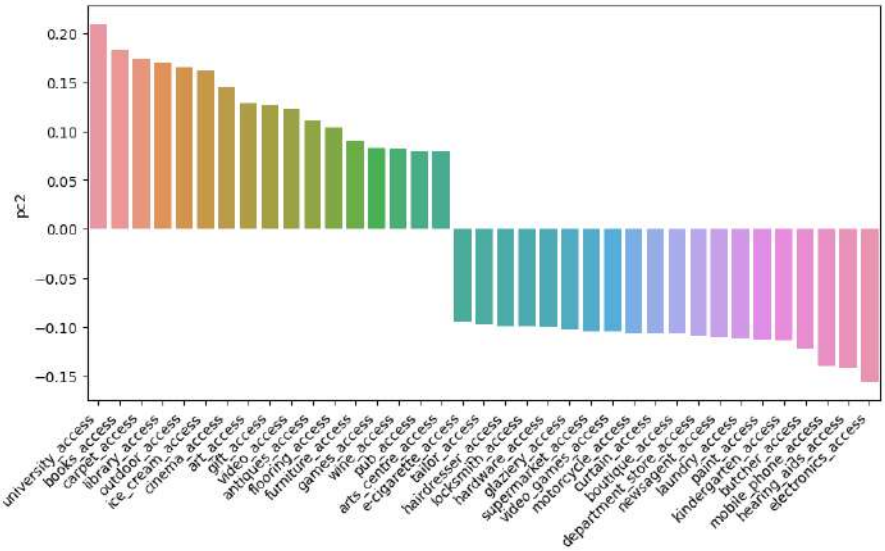


(b) Top30 of the features with the biggest coefficients in the second PC

Figure 4.11: Second principal component



(a) Map of third principal component



(b) Top20 of the features with the biggest coefficients in the third PC, bottom20 of the features with the most negative coefficient in the third PC

Figure 4.12: Third principal component