Paris Gentrification Presentation

Paul Joahny and Matei Vasile

Table of contents

1. Research Question

This report attempts to answer the question **how does gentrification in the Paris area affect the "Tissu comercial"?** The objective of this report is to assess how the development and structure of businesses changes from gentrified neighbourhoods to those that are not.

Gentrification is a social phenomenon , as many social phenomenon it has different definitions we'll define it as the process by which a poor neighborhood in a city is replaced by a new population. We can evaluate this by looking at the development of the population of a neighbourhood overtime. This is completed via a process in which a poor neighborhood in a city is changed by people who have money, including especially the improvement or replacement of buildings and businesses.

The idea is to see how gentrification can reshape neighbourhoods in terms of retail/service supply. As higher income populations move into formerly lower income or mixed income areas, we expect to observe a shift in the available businesses and more generally the commercial mix: growth of cafes, restaurants, shops, boutiques, coworking spaces etc with a subsequent decrease in traditional proximity retail. We will also attempt to observe whether new types of businesses emerge specifically tailored to a wealthier population.

We will do this by firstly identifying areas (IRIS units) in which clear signs of gentrification have occurred. To do this we will create some indicators for price per sqm, median income levels, poverty rates, income disparity etc. Then using these to identify areas that for example have been previously below the average price per sqm and are now above that, we can note them as gentrified areas.

Once we have identified which IRIS units have undergone gentrification, we will observe changes in the "tissu commerical" of these areas using the BPE (Base Permanente des Equipements). The BPE dataset is a record of the quantity and type of establishments. By analysing changes in the BPE for our identified gentrified areas we can see whether new business categories emerge as the distribution of income changes, which type of businesses increase in quantity, which types decline.

This method of first identifying gentrified areas, then analysing the change in types and quantities of businesses, allows us to observe the effect of gentrification on the "tissu commerical" of Paris.

2. Dataset Descriptions

a)FILOSOFI

The filosofi database, offers specific estimates of household income, poverty, and inequality at area levels such as IRIS. It combines fiscal data and social benefit data to estimate income before and after redistribution, adjusted per household unit to account for household size and structure. Filosofi replaces older local income datasets and provides useful indicators like income deciles, medians, poverty rates and inequality measures like the Gini coefficient.

b)IRIS Historical Change

The "Historique des codes IRIS" dataset shows a cross-reference of IRIS codes in France from 1999-2022, revealing how these neighbourhood-level statistical units have changed over time. It offers a "table de passage" that links old and new IRIS identifiers following boundary modifications, splits, or regroupings. This resource allows for researches to align socio-economic data over the years despite changes in IRIS definitions.

c)IRIS Shapefile

The IRIS shapefile provides the official polygon boundaries for all IRIS units in France. The 2020 "France entiere" shapefile offers total coverage of France and is the standard geographic reference for mapping IRIS-level data. It is broadly used to link datasets (like FILOSOFI and DVF) to their corresponding neighbourhoods.

d)DVF

The "Demandes de Valeurs Foncieres" (DVF) dataset, which is available through data.gouv.fr, contains property transaction records for France and includes geographic information when available. We have data sets from 2014-2018 and 2020-2025. DVF provides valuable information such as transaction dates, sale values, property types, surface areas. It can be used to analyse real estate information, such as change in price per square metre. In our project, DVF enables us to construct indicators of housing price changes at the IRIS level.

e)BPE

The "Base Permanente des Equipements" (BPE) dataset for Ile de France. It provides an inventory of public and private facilities, amenities and commercial establishments for the region of France.

f)Activity of residents

This base « Activité des résidents » gives data on different characteristics of the workers of a given neighborhood. This could be age , sex , CSP category.

3. Data Analysis

The main part of our analysis was finding a way to measure something you can't directly observe: gentrification. To be able to do that we chose to measure it trough 3 different dimensions. Revenue, Population and Prices.

Population : This whole dimension is looking at who are the people living in said neighbourhoods. In france there's a good indicator called CSP(Catégorie Socio Profesionelle) created and monitored by INSEE this works as a statistical categorisation of jobs of a similar social environment.These categories are

Employees-Employés : This is Ce groupe socioprofessionnel rassemble des professions aux fonctions très variées (administratives, commerciales, de services aux particuliers, de surveillance et sécurité, etc.) dont il est difficile de trouver une définition commune si ce n'est qu'elles n'ont pas ou peu de responsabilité d'encadrement. Ce groupe socioprofessionnel est composé uniquement de salariés, à de rares exceptions près.

Workers-Ouvriers This represents that work in the industrial secrors Ce groupe socioprofessionnel regroupe des personnes qui exercent des fonctions d'exécution dans le cadre d'une division poussée du travail dans les secteurs industriels, de services à l'industrie (nettoyage, maintenance, tri, expédition, etc.) ou des tâches manuelles dans les secteurs artisanaux ou

agricoles. Il ne comprend que des salariés, qui peuvent être employés par des établissements de nature publique ou privée.

This is very useful for our analysis some of theses categories represent a

We chose to measure trough an indicator with a score, gentrification at least as we see it isn't a binary variable it's a relative notion. It makes way more sense to see it as which neighbourhood is more gentrified than an other instead of simply saying which one is or isn't. To do this we built an indicator that itself is composed of 3

a. Prices

This indicator was built from 2 distinct DVF databases. One from 2014 to 2018 and one from 2020 to the first semester of 2025. Respectively "DVF_Paris_2014_18" and "DVF_Paris_2020_25". These two bases represented all property transactions records for the Paris region (this was done after a good amount of filtering). It was important to make sure these two bases had treatable data so we had to limit a few parameters, getting rid of anything too small or anything too big (needed a certain amount of square meters excluding closets or mansions). We could then build a price/squaremeter column this was done quite simply

$$\text{prix\_m2} = \frac{\text{valeur fonciere}}{\text{surface\_reelle\_bati}}$$

Even after our filtering some data made no sense (prix_m2=600 or prix_m2=50000) so we had to get rid of a part of the distribution to only keep realistic data. This mistake is most certainly the result of a typing error we're not excluding extremely cheap neighborhoods only unrealistic ones.

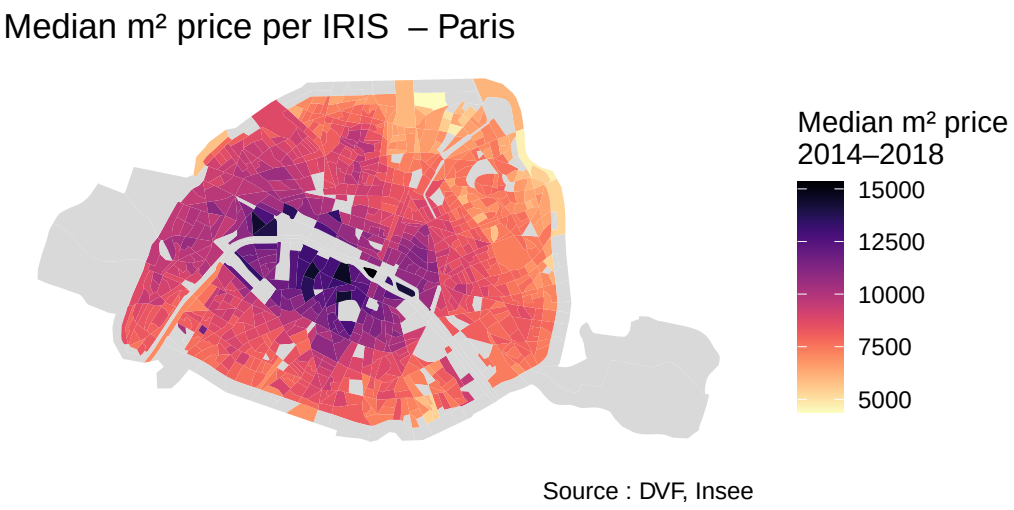
The big step was now the Spatial reattachment. Our whole goal is to look at neighborhoods but we only have transactions how can we link them into neighborhoods ? Our data for DVF_Paris_2020_2025 are transactions but we have longitude and latitude. Those transactions function as points on a map thanks to a iris shape file we'll be able to get the irises that will function as polygons and every transaction that fits into one of these polygons (or close to the border) well get reattached to a corresponding IRIS code from 2020.

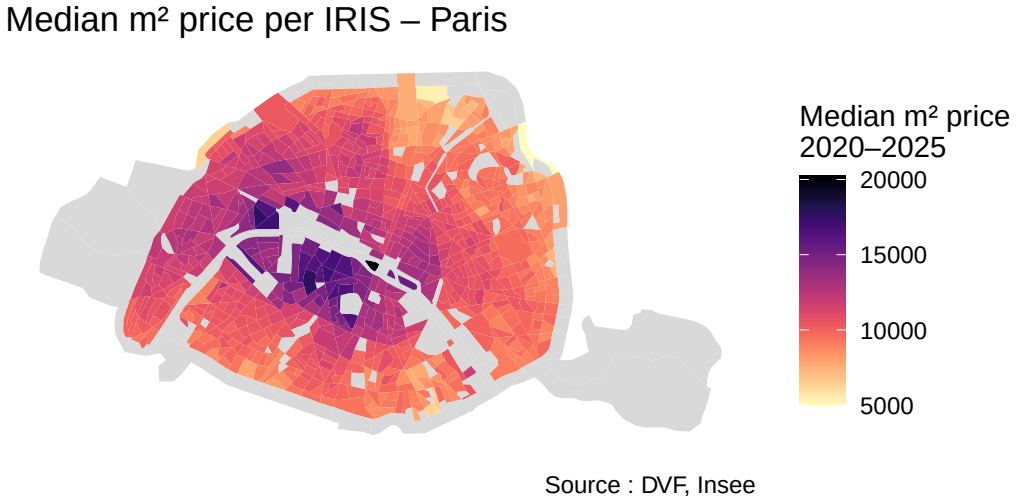
But for DVF_Paris_2014_18 this wasn't as easy as we didn't have any prior IRIS code or geographical coordinates We had to rebuild from the corresponding columns a complete postal address for every transaction.We had to then extract the set of unique addresses that we had just built, as we didn't want to search for the same address for transactions that happened at the same place during different years. We then used the banR package to reverse geocode coordinates from our addresses. This functions thanks to the French national address database (BAN). Every address given by our data set receives geographical coordinates and a confidence

score to gives us an idea how accurate the match is. To ensure spatial accuracy, we retained only geocoded addresses with a confidence score of at least 0.6. This threshold was found with trial and error, manually checking what level was precise enough.With the corresponding validated coordinates they got merged back into our transaction dataset. This meant that transactions with no coordinates disappeared. We find ourselves in the same position as beforehand with coordinates we only need to see which ones fit inside of the corresponding polygons.

Once all transactions were spatially linked to IRIS neighborhoods, we aggregated the data by IRIS and year. For each IRIS-year pair, we were able to compute : the median price per square meter, the mean price per square meter, the number of transactions. To ensure we had representative resultes, we only kept IRIS-year observations with at least five transactions. This final dataset was then merged with the 2020–2025 data to create a continuous price indicator covering the 2014–2025 period.

Everything described beforehand can be found in the file "DVF_cleaning.qmd" in the file "Cleaning'.





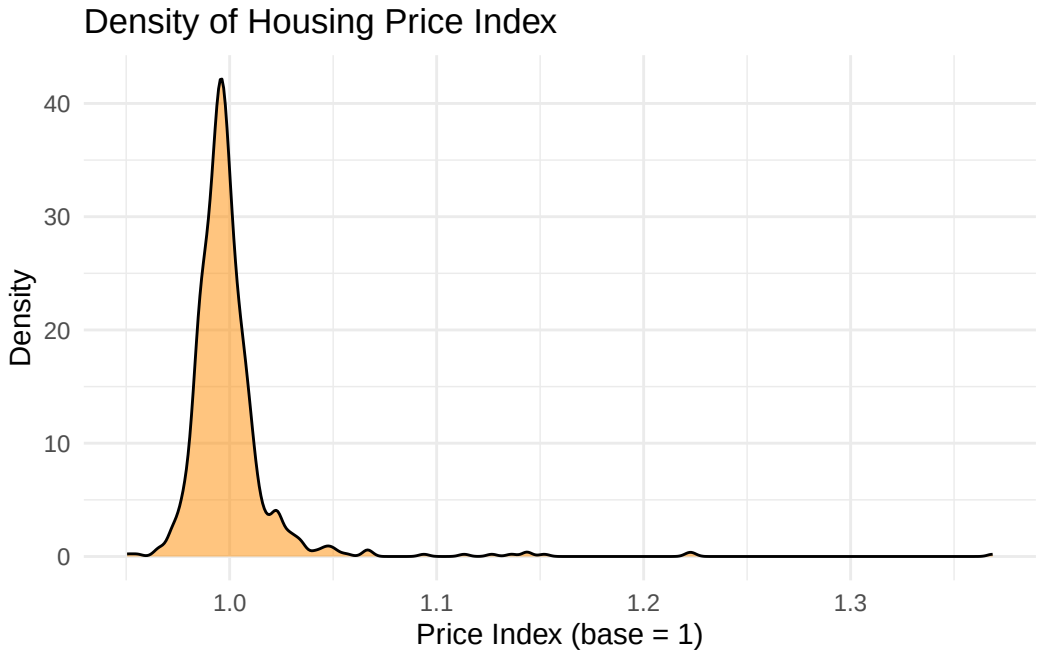
Using the DVF dataset, we construct a housing market indicator that measures relative price changes at the IRIS level. The aim of this indicator is to see whether housing prices in a given neighbourhood have increased at a faster rate than the Paris wide market and whether this occurs consistently over time.

First for each IRIS and year, we compute a representative price per square meter using the median values. Next we compute annualised price growth rates between observed years. This expresses the price changes as a per year growth rate, making it possible to compare across different gaps of time. Housing prices are subject to influence from macroeconomic effects and city wide occurrences. To isolate the real changes in price dynamics we compute for each year the Paris wide average growth rate. This benchmark average represents the normal evolution of housing prices in that year.

Now for each IRIS year we can define relative price growth. Relative price growth is the deviation from the city wide benchmark that we calculated earlier. So for a specific IRIS year a positive deviation would mean price evolution that increased faster than the market, negative meaning prices that lagged behind the market. Then finally, relative price growth rates are accumulated over time to form a final index. This is index is constructed such that the base value 1 represents following the Paris wide price movements exactly, an index greater than 1 would result in cumulative price outperformance and an index less than 1 indicates cumulative underperformance.

```
# A tibble: 10 x 2
   CODE_IRIS index_price
```

	<chr>	<dbl>
1	751156009	1.37
2	751166124	1.22
3	751135015	1.22
4	751207910	1.15
5	751103801	1.14
6	751156010	1.14
7	751207906	1.14
8	751197317	1.13
9	751186925	1.11
10	751207801	1.09



b. Revenue

This indicator was built from the FiloSofi databases. The goal of this cleaning was to create a panel of different income information at the IRIS level, harmonized across years and restricted to the Paris area. We used files from 2015, 2016, 2018, 2020, and 2021.

For each year, we first read the Excel file and cleaned the column names to harmonize them, only keeping the ones we found interesting. The biggest harmonization that had to be done was with the iris codes.Iris codes evolve with time and this has to be taken into account. We had chosen prior to sink our analysis with iris code 2020. This is why we used a passage table

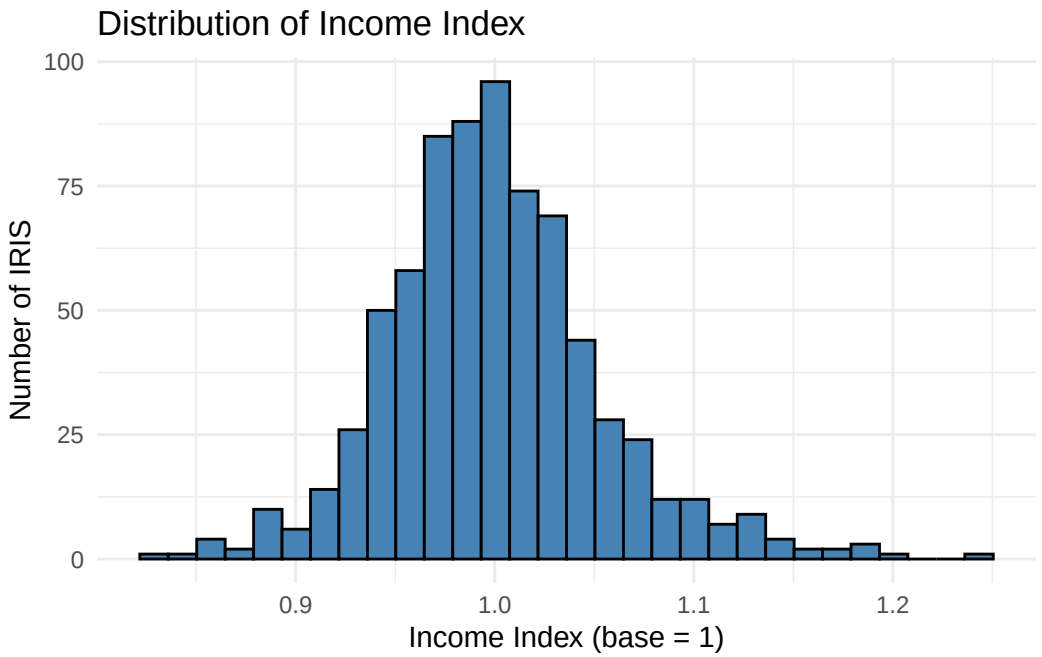
to harmonize all codes to the same date.This allows us to track the same geographic units consistently across years.Finally, all yearly datasets were merged into a single panel, with a column indicating the year. The final dataset contains IRIS-level income data for all years from 2015 to 2021, ready for indicator construction.

Everything described beforehand can be found in the file "Cleaning_FILOSFI_Files.qmd" in the "Cleaning" file The final dataset is saved as filo_final.rds.

From the Filosofi dataset we constructed another indicator that captures changes in come structure at the IRIS level. This indicator is composed of two parts: a median income component and a poverty composition component.

We first look at mdeian income as a measure of economic change that can reveal gentrification. For each IRIS and year, we observe the median income. We compute annualised income growth rates between observed years, this produces a per-year growth rate that we can compare across time. From this we can create for each year a Paris wide average income growth. We define relative income growth as the distance of an IRIS from this city wide average. Positive values indicate that income in an IRIS has grown faster than the Paris average in that year, with negative values showing underperformance comapred to this average. Finally, we aggregated these distances over time into a cumulative relative income index with a base value of 1. An index value of 1 indicates that the median income in the IRIS followed the Paris wide average exactly over the period of time, values over 1 indicate overperformance and those udner 1 indicate underperformance. This accounts for half of the total indicator created from FiloSofi.

```
# A tibble: 10 x 2
   iris      income_index_score_1
   <chr>                    <dbl>
 1 751114112                 1.24
 2 751187107                 1.20
 3 751187103                 1.19
 4 751187017                 1.18
 5 751197401                 1.18
 6 751187105                 1.18
 7 751187203                 1.17
 8 751187109                 1.16
 9 751187202                 1.15
10 751197308                 1.15
```



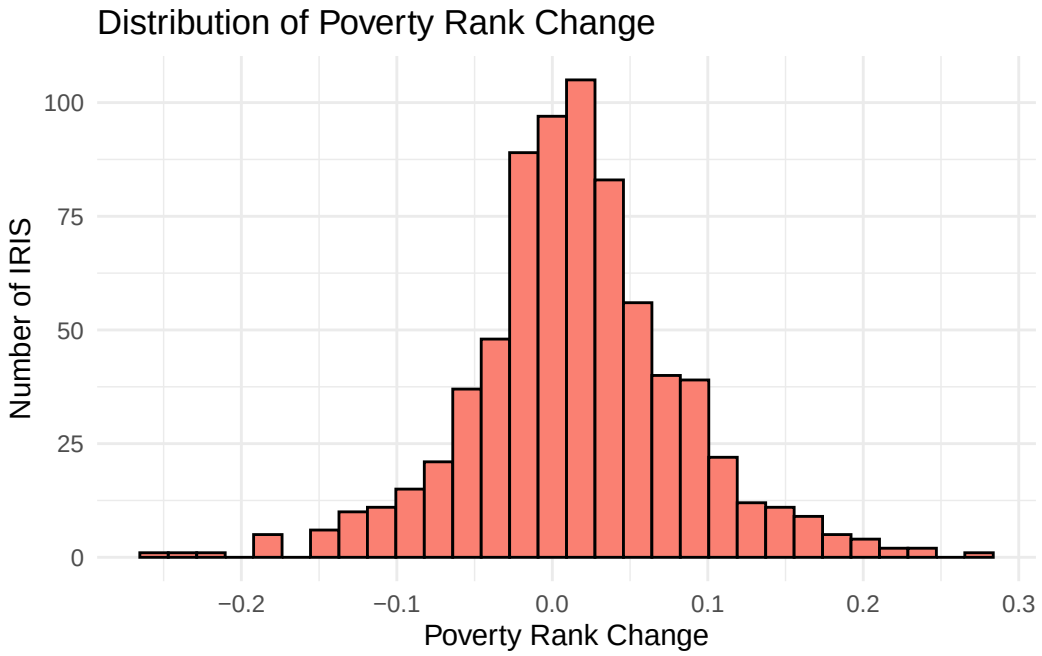
The second part is constructed from the percentage of the population who have an income deemed under the "poverty line".

For each IRIS and year, we observe a poverty share P. For each year, all IRIS are ranked by their poverty share thus higher ranks correspond to higher poverty and lower ranks correspond to lower poverty. We then invert the ranking such that higher values indicate lower poverty (improvement). To observe change over time, we compute the change in relative poverty position between the first and last observed years.

Positive values indicate that an IRIS moved toward the low-poverty end of the Paris wide distribution, while negative values indicate increasing relative poverty. This captures relative displacement or upgrading which can show gentrification.

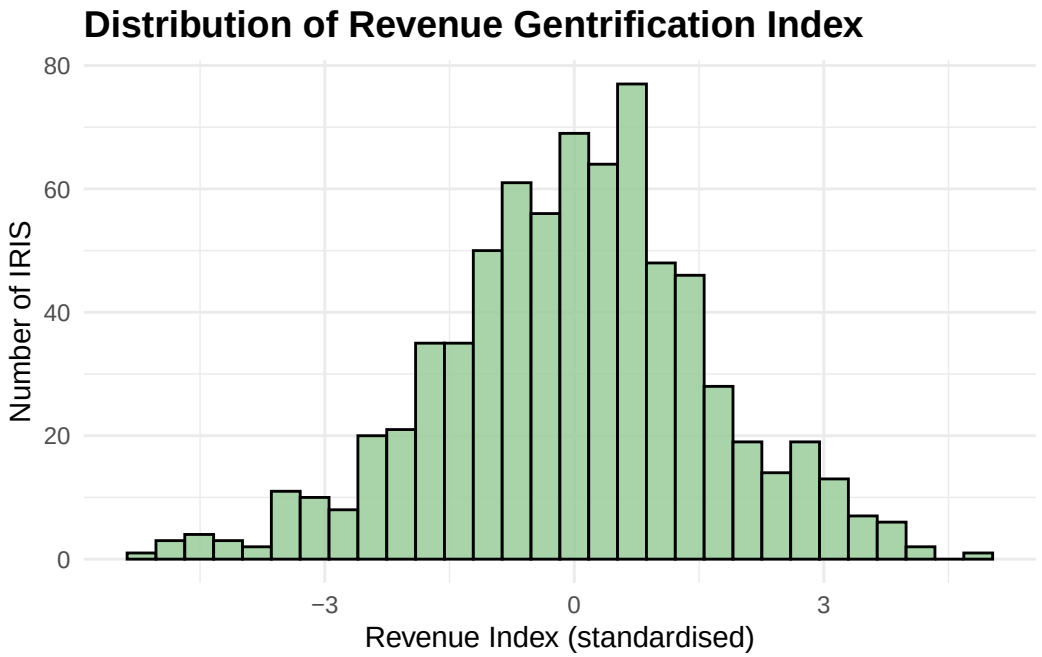
```
# A tibble: 10 x 2
   iris      poverty_rank_change
   <chr>                   <dbl>
 1 751103903               0.267
 2 751187003               0.240
 3 751114115               0.238
 4 751093502               0.215
 5 751155909               0.214
 6 751010204               0.210
 7 751114203               0.203
 8 751187013               0.196
```

```
 9 751114314              0.192
10 751062204              0.186
```



The two indicators constructed from income and poverty capture changes in socioeconomic factors that can unveil gentrification. The income index reflects upwards income change and the poverty rank change reflects changes in the composition of the population.

Because the two indicators are measured on different scales and have different distributions, each is standardised to have mean zero and unit variance. Thus we create a weighted linear combination concluding with the final index $S = 0.6Z(\text{inc}) + 0.4Z(\text{pov})$, a higher S indicates a stronger improvement than average of income changes and poverty share changes relative to Paris wide averages.



c. Population

This indicator was built from the resid_activity databases.The goal of this indicator was to measure how does the population inside a given neighborhood evolves. This is especially interesting for a gentrification perspective to look across time if the composition of a neighborhood has drastically changed (hint of a wealthier population coming in).

To do so we used 4 databses from 2018 to 2021 which we restricted to the Paris area. Like for the rest of our analysis we needed to harmonize the iris codes to a same Year format and have temporal compatibilty, we kept the IRIS-2020 like for the rest of our previous databases.

The important aspect was now to know which variables to keep further on for our analysis. Except the ones to help identify (iris , year, ...) we choose to keep :

-pop1564 <- This represents the total population of the neighbourhood between 15-64.

-act1564 <- This represents the population either working or unemployed (actively searching for work) of the neighborhood between 15-64.

-actocc1564 <- In act1564 this represents the number people in the neighborhood with a job.

-chom1564 <- In act1564 this represents the number of unemployed unemployed people in the neighborhood (act1564 = chom1564+ actocc1564)

-sal15p <- this represents the number of residents of the neighborhood that are receiving a salary (in contrary to being independent and being your own boss).

-etud1564 <- this represents the number of residents of the neighborhood that are students between the age 15-64. (They are not considered as actives).

-cadres_1564 <- This represents a sub group of actocc1564, those are residents with a job but more specifically that are executives.

-prof_inter_1564 <- This represents a sub group of actocc1564, those are residents with a job but more specifically that are

-employes_1564 <- This represents a sub group of actocc1564, those are residents with a job but more specifically that are employees

-ouvriers_1564 <- This represents a sub group of actocc1564, those are residents with a job but more specifically that are workers (Closer to manual labor,factory work for example)

Table 1: First 5 lines of Merged base

iris_2020	annee	pop1564	act1564	actocc1564	chom1564	sal15p	actocc15p	etud1564	cadres_1564	prof_inter_1564	emplo
751010101	2018	653.55	522.41	458.13	64.29	393.32	472.48	75.78	196.02	129.80	
751010102	2018	100.41	68.23	59.38	8.84	43.66	71.39	23.27	33.38	8.84	
751010103	2018	144.94	106.06	91.60	14.46	59.69	98.49	16.75	60.42	19.28	
751010104	2018	3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
751010105	2018	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

All of the previously mentioned work can be found in file “Cleaning_resid_actity.qmd” which is found in “Cleaning” file.

After creating this database , it was time to build an indicator. We felt it wasn’t very interesting to look at absolute values but more interesting to look at shares of a given CSP(Socio-Professional Category). We made sure our denominator was always bigger than 0 (Trying to avoid dividing by 0).

We then created 3 new variables:

share_csp_sup = (cadres_1564 + prof_inter_1564) / actocc1564 Csp_sup is a new category we built to look at how more “privileged” or “high-end” professions would be entering the neighborhood. The fact we made it a share reads as the percentage of these high end profile in respect of all workers in the neighborhood. If this share increased this could be seen as more and more of these individuals entering a certain neighborhood, a good lead for gentrification.

share_populaire = (employes_1564 + ouvriers_1564) / actocc1564, Populaire is the opposite side of this analysis. This would be the profile you’d expect to be living in a future gentrified neighborhood and would expect their number to decrease as they are being replaced by CSP_sup. You’d expect that these shares evolve in opposite trends when looking at a potential candidate for gentrification.

share_chom = chom1564 / act1564 This represents the unemployment rate for all active residents.

After having created these three new variables, we then focused on looking at how these shares evolved over time rather than just their absolute values. To do this, we calculated the slope of each share for each IRIS. To do so we used a linear regression of each share to to find yearly trend. THe goal was to see if the proportion of privileged profession were increasing, if the more popular profession was decreasing and how unemployment was evolving. The goal was to look at dynamic changes since gentrification is a long term process we need some overtime tendancies.

Next, we filtered IRIS with at least 4 years of data to ensure the trends were reliable (available all of our databases timespan 2018-2021). Then, we standardized the slopes (converted them to Z-scores) to put them on the same scale. This step is important because it allows us to combine these different trends into a single indicator without one variable dominating the others simply because of its scale.

Finally, we created a composite population index (index_pop) defined as:

$$\text{index\_pop} = Z_{CSP+} - Z_{populaire} - Z_{chom},$$

where:

- Z_{CSP+} = standardized slope of high-end professions,
- $Z_{populaire}$ = standardized slope of popular professions,
- Z_{chom} = standardized slope of unemployment.

A high score in this index should logically indicate a good neighborhood candidate for gentrification because it should show a certain iris code where the proportion of privileged professions is increasing while the proportion of popular professions and unemployment are decreasing. These tendencies are very consistent with gentrification. On the opposite side a low score should either show a stable neighborhood (already gentrified or not concerned) or opposite trends which can't be seen as gentrification.

We then sorted IRIS by this index to identify neighborhoods showing the strongest signs of socio-economic change, and saved the results for further analysis.

d Gentrification score

This indicator was built by combining the three previously constructed indices: housing prices, population dynamics, and income/poverty changes. The goal was to create a single measure

capturing the overall gentrification dynamics at the IRIS level. First, we harmonized IRIS codes across the three datasets to make sure we were comparing the same geographic units. Only IRIS that were present in all three datasets were kept for the final index. Next, each indicator was standardized to have mean zero and unit variance. To avoid extreme outliers skewing the results, we capped all standardized values at ± 3. Finally, we constructed the composite gentrification score as a weighted sum of the three standardized components: $0.4 \times$ housing price index $0.3 \times$ population index $0.3 \times$ revenue index The resulting score gives a single measure of relative gentrification: higher values indicate stronger upward changes in housing prices, population composition, and income/poverty structure relative to the Paris-wide average. Everything described beforehand can be found in the file "Gentrification_Score.qmd" in the "Cleaning" folder. The final dataset is saved as gentrif_indices.rds.

e BPE

These data sets was built from the BPE (Base Permanentes d'Équipements) databases for the years 2016, 2018, and 2024. They give us info of one the geographical position of different establishments in France but for our analysis we limit ourselves to businesses.These databases function with a specific nomenclature with a letter and three number A405 to define what does the "equipment" do. The letter indicates a broad category and the first number a sub category. After filtering all of the bases to the Paris region and only keeping necessary columns, we face the similar problem of spatial reattachment. Both 2016 and 2024 base had geographical coordinates so similarly to the DVF beforehand it was just a matter of putting them inside of the corresponding IRIS_2020 neighborhoods. The 2018 one hade an IRIS code so using the previously mentionned passage table (in the Revenue section) we were able once again to harmonize all bases to a same time frame for IRIS2020.

All of the previously mentioned work can be found in file "Cleaning_BPE.qmd" which is found in "Cleaning" file.

We then use our three previously worked BPE datasets.They describe the same type of information but the variable names and nomenclatures differ across years. We notice as time goes on the nomenclature gets more and more specific declining into smaller more precise categories. Once again in order to be able to compare them, a harmonisation step is required. The BPE 2024 dataset only contains equipment type codes. To make the data workable, we merge the official BPE 2024 nomenclature in order to retrieve the corresponding labels. This makes it possible to work directly with readable activity names which are more easily understandable. Each yearly dataset is then reduced to a common structure. We only keep iris <- the harmonized iris code year <- the year the equipment was registered in code <- this the four character to design type <- clear label of what this equipment is nb <- how many of those are found in this area

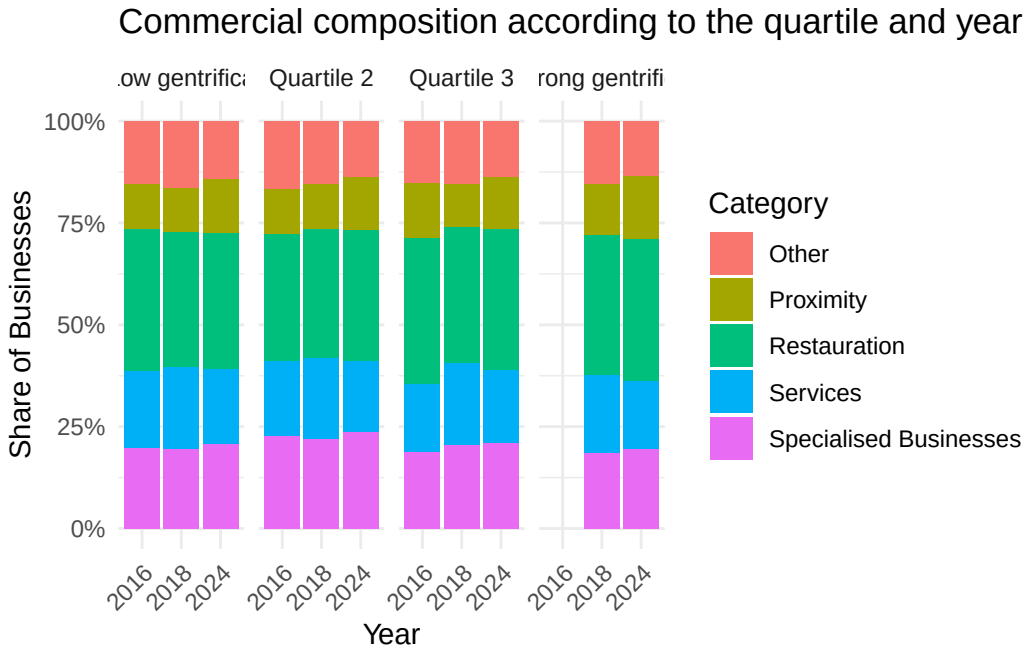
The problem is that the nb in 2024 had a lot of missing data so we assume a value of one when there's an NA, which corresponds to the presence of a single establishment.

Once the three datasets are harmonised, they are merged into a single BPE panel. To avoid artificial differences due to spelling, accents or casing, all equipment labels are normalised by converting them to lowercase and removing accents.

Our analysis focuses on the effect of gentrification on surrounding businesses (tissu commercial). For this reason, we restrict the dataset to business-related activities only. We keep categories B (Commerces) and A5 (Other services), which include proximity retail and service activities. All other equipment types are excluded. Finally, some activities still appear under slightly different names across years. We merge these cases into common categories when the differences are minor (for example restaurants, beauty services or specialised retail).

4. Conclusion

While we have clearly been able to contruct an index that observes gentrification dynamics from residential prices, income structure, and population composition, these observations do not translate into substantial results of changes in the BPE.



From the representation it is apparent that the share of types of businesses for each quartile of our gentrification score do not differ significantly, nor do they differ significantly across time.

Although our analysis is succesful in identifying neighbourhoods that have undergone gentrification it is not strongly reflected in the BPE.

Firstly we must address that in the BPE we have a fairly large missing column of data in 2024. The column for number of equipments is missing which specifies the number and type of businesses in the area. This means that it is assumed in 2024 meaning any substantial change in the amount of restaurants for example is not visible.

Overall when it comes to the BPE it can be for multiple reasons that we do not observe changes. First, commerical change often happens more slowly than residential change. People can move into a neighbourhood relatively quickly as housing prices and income levels change but shops and services are generally much more stable. Many businesses are under long contracts and face high relocation costs. As a result even if the population or income structure changes, the types of businesses present may remain the same for several years.