Tracking gentrification evolution in the neighborhoods of Paris

Abstract

The democratization of machine learning in different fields has allowed researchers to understand some of the most complex phenomena, including gentrification. This process always has been studied with qualitative methods by social studies. This paper aims to understand the evolution of this urban phenomenon in Parisian neighborhoods between 2006 and 2019 through machine learning methods. Based on census data from the French National Institute of Statistics and Economic Studies (INSEE in French), we will first identify the various sociodemographic characteristics for each neighborhood according to the median income through the years with multiple linear regression. And then detect which one has gentrified with the random forest method.

1. Introduction

The British sociologist, Ruth Glass, introduced the concept of gentrification in the '60s while studying the housing problem in London. She described it as the transformation of the working class or/and poor neighborhoods in the cities by, and I quote, "the process of middle-and upper-income groups buying properties in such neighborhoods and upgrading them." With this definition, she warned the London administration of that time about this process's side effect: the displacement of the lower social classes from the city(Raman, 2014). Her work on this phenomenon has influenced other social science fields to take an interest, specifically in human geography. For example, Carpenter. J and Lees. L could identify and compare gentrification in New York, London and Paris. They concluded that just like urbanization, gentrification affects all significant cities in the world but differs according to historical, political, economic, and cultural contexts (Carpenter Lees, 1995). Most of the studies on neighborhood changes in western cities relate to a qualitative method such as; literature search, field surveys, or interviews. According to Borton. M, these methods are not suitable for having a better understanding of gentrification, identifying which neighborhoods have been gentrified, and at the same time, 'may overlook areas that experienced similar changes to those more widely recognized as gentrified" (Barton, 2016). Still, according to Borton. M, if we mobilize more quantitative methods than qualitative ones, we can track the neighborhoods and changes and, thus, predict those who will be affected (Barton, 2016). Predicting neighborhood changes could also prepare the municipal authorities to put in place political initiatives against the drifts of gentrification. This is what Jonathan Reades, Jordan de Souza, and Phil Hubbabrd did, in 2019, with the neighborhoods in London. Their paper titled "Understanding urban gentrification through machine learning" aims to predict areas going through this process. My report will not have the ambition to predict the neighborhoods that this effect could touch, but it will show this process's evolution through several periods (2006, 2011, 2016, and 2019) in the neighborhoods of Paris. First, I will present the data that I mobilized for this work. Secondly, I will explain the methodology inspired by the article written by Reades al. Then, I will display the result based on my methodology and finish with a conclusion.

. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Data

I used the dataset from the French National Institute of Statistics and Economic Studies (INSSE in Frech). Their mission is to collect and analyze data about the French economy and society and then communicate their results through reports or databases. This is the organism in charge of the population census's population at different scales (national, regional, districts, arrondissements, and cities). As mentioned earlier, gentrification is a phenomenon that is observable at a neighborhood scale, and INSEE needs to study at this scale because this entity is quite challenging to delineate. To solve this problem, INSEE created in 1999 a new study scale called IRIS (this a French acronym, but in English, it stands for "aggregated units for statistical information"). There are three types of IRIS;

- residential IRIS: this is for the population and owns an average of 2000 residents per unit
- business IRIS: this IRIS is for economic activities and accounts for 1000 employees per unit

- and miscellaneous IRIS, representing an area with few or no inhabitants (parks, forest, mountains...)

Municipal authorities mainly use these units. For the city of Paris, there are approximately 1000 IRIS.

After we could identify at what scale our data is used, we need to select the variables. Since gentrification implies the replacement of a social class by a much higher one, which entails a rent increase in an urban area, we need variables that explain the demographic structure for each IRIS (neighborhood). So we collected different types of the census;

- Population: it includes age, gender, nationality, level of occupation, and education
- Education: the level of education
- Accommodation: give a complete description of the accommodation type for each IRIS (number of rooms, bedrooms, bathroom, presence of balcony or not, ...) housing income

I had to collect those four censuses for each year (2006, 2011, 2016, and 2019), which resulted in 16 censuses. As you can see, we did not include the rent price because INSEE does not collect this information. So I have to use the accommodation census produced by INSEE.

3. Methodology

3.1. Literature

Machine Learning applications on urban analysis are well explicitly documented regarding recent research on this topic. However, there needs to be more literature on ML application gentrification. Just like Burton. M mentioned that gentrification is more widely studied qualitatively than quantitatively. In addition to its rarity, there is no article with ML application on Paris gentrification since this process acts differently depending on the city. Despite the lack of work on the case of Paris, I found one paper that deals with my problem. As I have already mentioned, the article written by Reades and al. tries not only to predict neighborhood changes but also to understand and detect the patterns of those areas. They used census datasets on Lower Layer Super Output Area (LSOA) (similar to IRIS but for British people). In total, their dataset contains more than 160 variables. After collecting their data, they used the Principal Component Analysis (PCA) to reduce the dimensionality of the variables and select the predictor variables for their machine learning model. The model they chose was the Random Forest to classify the LSOA based on the scoring results from the PCA. Furthermore, they use GIS tools to communicate their results through maps.

3.2. Workflow

After importing the dataset to my notebook, I regrouped all the demographic datasets together (population, education, and accommodation) for each year. Those datasets will be our predictors variables, and each regroups at least 180 variables. Then, for our predicted variables (the median income), I merge them.

Figure 1. Scree shot of demographic dataset (from 2011) after the merge

IRIS	P11_POP	P11_POP0002	P11_POP0305	P11_POP0610	P11_POP1117	P11_POP1824	P11_POP2539	P11_POP4054	P11_POP5564	P11_POP6579	...	P
751010201	2382	37	77	60	76	286	853	506	274	146	...	
751010202	1745	66	43	36	48	170	660	285	184	162	...	
751010203	2390	52	56	126	97	200	739	517	255	254	...	
751010204	2476	74	63	76	164	259	742	562	208	231	...	
751010301	2928	97	79	71	117	351	755	539	446	328	...	
...	...	...	...	...	...	...	...	...	...	...	...	
751208022	1775	61	36	81	108	196	334	339	223	158	...	
751208023	2265	100	68	156	130	246	703	409	250	143	...	
751208024	2785	103	85	146	128	247	712	524	343	295	...	
751208025	2604	130	85	159	205	176	626	628	320	186	...	
751208026	1907	50	35	76	85	162	624	461	214	163	...	

865 rows × 173 columns

Figure 2. Scree shot of median income (from 2006, 2011, 2016 and 2019) after the merge

	IRIS	RFMQ206	RFMQ211	DISP_MED16	DISP_MED19
6594	751010201	28062.0	31759.0	29744.0	30010.0
6595	751010202	33049.0	41638.0	33466.0	36960.0
6596	751010203	31039.0	35971.0	31726.0	32720.0
6597	751010204	32003.0	36257.0	28284.0	32360.0
6598	751010301	36872.0	40250.0	34345.0	36830.0
...	...	...	...	...	...
7454	751208022	29893.0	31897.0	19767.0	20680.0
7455	751208023	26337.0	29407.0	20226.0	22900.0
7456	751208024	27960.0	32221.0	24603.0	25530.0
7457	751208025	28300.0	31893.0	23841.0	24500.0
7458	751208026	27389.0	31487.0	27139.0	29220.0

865 rows × 5 columns

The first step of my methodology was to determine which variables best explain the median income for each IRIS with multiple linear regression. The best way to optimize the model was to split the data into three datasets (training, testing, and validation). Moreover, to measure the model's reliability, I used the mean squares error. Since the main objective was to classify the neighborhoods by their median income, a random forest was more suitable for this task. However, we need to use a baseline model for our case multinomial logistic regression before using it. To verify if the logistic regression has correctly classified my neighborhoods, I used the accuracy and F1 scores. And then, I could compare both the logistics regression and random forest with the same metrics.

4. Results

While I was trying to train to fit the data into the models, I realized that I needed to create classes to classify the neighborhood. Also, when I use the metrics for both multiple linear regression and logistics regression, Python sends me an error message like the figure below.

Figure 3. Scree shot of one message error from the notebook

```
TypeError                                 Traceback (most recent call last)
<ipython-input-136-9260cd9808c9> in <module>
     23 print("The Accuracy for the validation set : %.2f" % llog.score(x_val, y_val))#0.0
     24 #The F1 score
---> 25 print('F1 Score for the trainig set: %.3f' % f1_score(y_train, predictrain_19, average='weighted'))#0
     26 print('F1 Score for the test set: %.3f' % f1_score(y_test, prediction_19, average='weighted'))#0
     27 print('F1 Score for the validation set: %.3f' % f1_score(y_val, predic_v19, average='weighted'))#0

                                  7 frames
/usr/local/lib/python3.8/dist-packages/sklearn/utils/validation.py in _num_samples(x)
    257     if hasattr(x, "fit") and callable(x.fit):
    258         # Don't get num_samples from an ensembles length!
--> 259         raise TypeError(message)
    260
    261     if not hasattr(x, "__len__") and not hasattr(x, "shape"):

TypeError: Expected sequence or array-like, got <class 'sklearn.linear_model._logistic.LogisticRegression'>
```

Thus, I could not even see which features importance from the predictors could explain those predictor values.

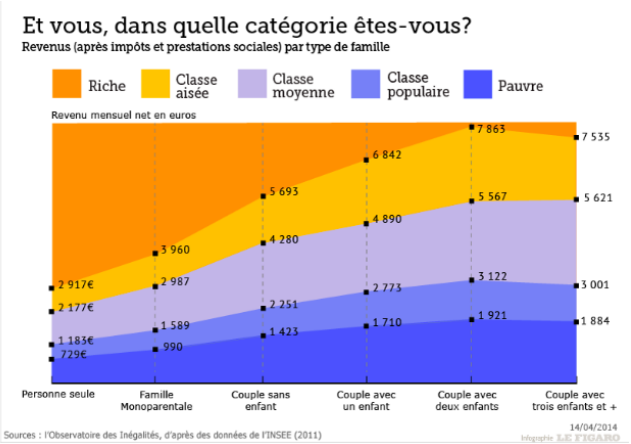
5. Discussions

If I am in this situation, it is because I spent too much time at the process the data. I also overestimated my ability in Python and did not ask for help at the right time when I needed it. If I could redo this project, I would first create a new variable that would provide information on the social classes for each neighborhood based in their median income;

- Poor class
- Working - class
- Middle class
- High middle class
- Wealthy class

The threshold would be established by one of the INSEE's studies that estimate who has estimated in a quantitative way those five classes (figure 4). And then, I would reuse the same method that I present in this paper.

Figure 4. Chart representing the social classes in France based on the median income and the number of people per household



References

Barton, M. (2016). An exploration of the importance of the strategy used to identify gentrification. Urban Studies, 53(1), 92–111. https://doi.org/10.1177/0042098014561723

Carpenter, J., Lees, L. (1995). Gentrification in New York, London and Paris: An international comparison. International Journal of Urban and Regional Research, 19, 286–286.

Raman, S. (2014). Gentrification. In A. C. Michalos (Ed.), Encyclopedia of Quality of Life and Well-Being Research (pp. 2509–2512). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5$_1$157

Reades, J., De Souza, J., Hubbard, P. (2019). Understanding urban gentrification through Machine Learning: Predicting neighbourhood change in London. URBAN STUDIES, 56(5), 922-942. https://doi.org/10.1177/004209801878905

5.1. The data that I used

Statistiques et études — Insee. (n.d.). Retrieved December 23, 2022, from https://www.insee.fr/fr/statistiques?debut=0theme=1categorie=3geo=ICQ-1

5.2. The script

https://github.com/Aaeilo/2022$_M L_E ES/tree/main/Project$