POLITECNICO
MILANO 1863

Analysis of San Francisco Housing Market

Methods and tools to act against gentrification and displacement

Academic Year 2022/2023

Pietro Bogani, Tomaso Castellani, Sara Tonazzi
8 CFU 8 CFU 5 CFU

Abstract

The San Francisco housing market has undergone significant changes in recent years, leading to increased gentrification and displacement of low-income and working-class families. To understand and address this pressing issue, we conducted a study aimed at finding practical solutions. Our goal was to gain insights into the phenomena of gentrification and displacement and to identify strategies for addressing them.

1.1 Introduction

Gentrification is a process that transforms the socio-economic character of a neighborhood, often driven by the influx of wealthier individuals who push out and displace the current residents. In San Francisco, the booming tech industry and demand for housing by tech workers has fueled a significant rise in housing prices, making it increasingly difficult for low-income and working-class families to afford to live in the city.
The consequences of gentrification and displacement are far-reaching and have negatively impacted the social demographic structure of the city. San Francisco has become one of the most expensive cities in the US, leading to a shift in its population and the displacement of long-time residents.
In light of these developments, it is crucial to find practical solutions to curb gentrification and displacement in San Francisco and other cities facing similar challenges. Our study provides valuable insights into the dynamics of these phenomena and offers strategies for addressing them.

1.2 Methodology

Our study focuses on the years 2011-2018.
For this analysis we used data on rent prices, evictions and new constructions.
In the first part of our study we focused our attention on understanding how the San Francisco housing market has changed in the latest years. In particular, we explored the

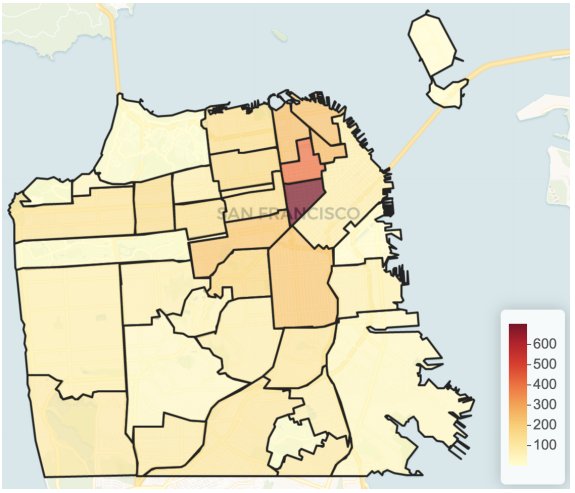
main trends of rent prices and evictions through a functional perspective and by visualizing data in maps.
In the second part of the study, several semi parametric models have been designed. The models were firstly developed at a neighborhood granularity (data aggregated at a neighborhood level) and were later modified to be implemented at a parcel granularity, allowing to focus on the local effect, down to only 100m. The aim is to understand how the number of new houses built affects over time the rent prices and the evictions.
Building new housing is controversial because its impact on rents and rates of displacement and gentrification nearby is ambiguous. Increasing the housing supply could ground soaring housing prices and slow demographic change. However, building new, high-quality housing could also increase demand for nearby housing by improving neighborhood quality. If these demand effects are larger than the supply effects, new construction could accelerate local displacement. We call this phenomena “renovation effect” vs “supply effect”. The results provide precious guidance for the local government to implement the proper policies to address the phenomena of gentrification and displacement.

2 Dataset

The majority of the data were gathered from the official website of the government of San Francisco, specifically from the *DataSF department*. *DataSF* serves as a centralized platform that provides access to data generated by the city and county of San Francisco, making it readily available to the public.

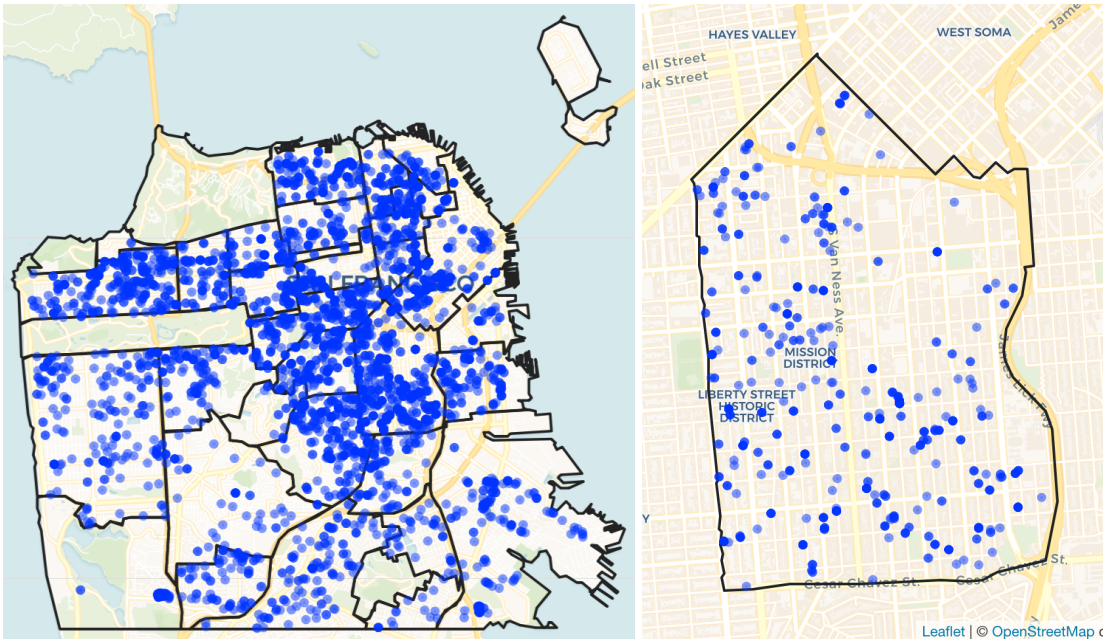
2.1 Evictions

The original eviction data, obtained from *DataSF*, comprised information regarding the neighborhood of properties where tenants received eviction notices and the date of the notifications.
The data cleaning process was designed to organize the observations into neighborhoods consistent with the other dataset and time periods (annual and monthly eviction numbers for each neighborhood).



2.2 New constructions

New building permits data was gathered from *DataSF*. These data included information about the exact location, the date of emission, proposed and existing units of new constructions. From the original data we kept only the permits associated with residential buildings and for which we had an increase of residential units.



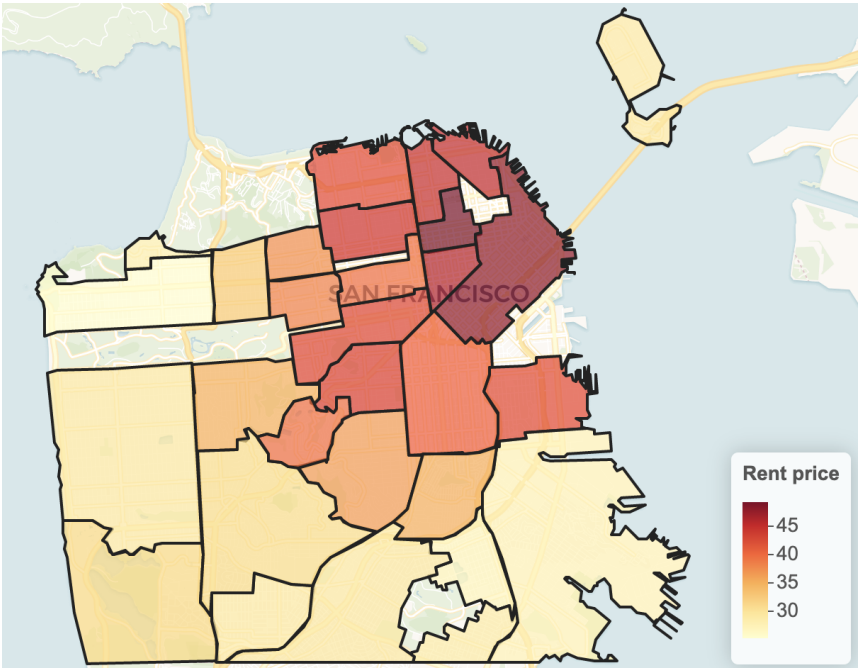
2.3 Rent

The city government does not provide to the public data on rental prices, therefore alternative sources had to be sought. We've been able to use the data gathered by Kate Pennington[2]: data from a comprehensive scrape of Bay Area rental posts using the Wayback Machine[3].

These data were obtained from Craigslist, a well-known website in the United States that hosts rental advertisements. It should be noted that these data may not provide a comprehensive representation of the rental market in San Francisco, as they may exclude the higher end of the market (which is typically managed by real estate agents) and the lower end (which is largely based on word-of-mouth). Furthermore, Wayback Machine stored information only sporadically and for this reason there are time intervals in which no data were collected.

It's important to notice that among the available data, 90% of them did not include precise geographic location information. However, the neighborhood information was available, allowing for the creation of monthly and yearly averages for each neighborhood.

In addition to the monthly rent price and apartment dimensions, the information available also included the computation of the rent price per square foot, which offers a more nuanced understanding of price dynamics.

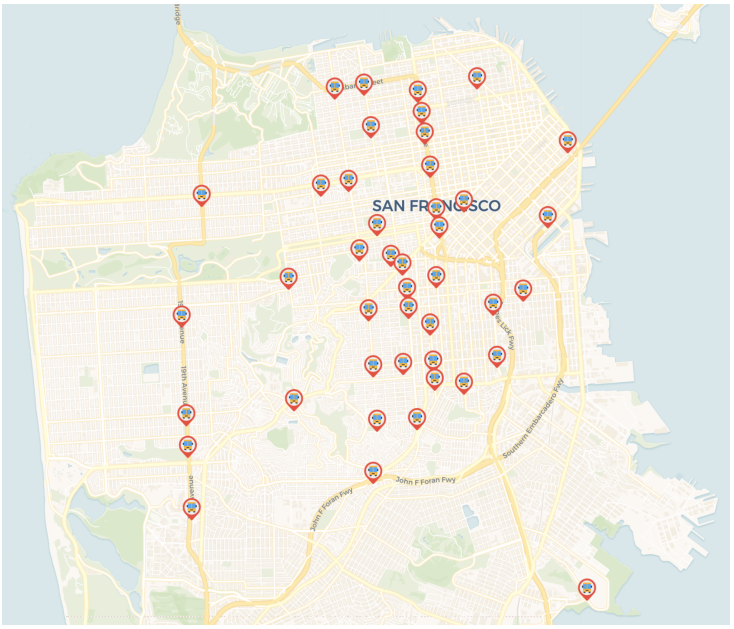


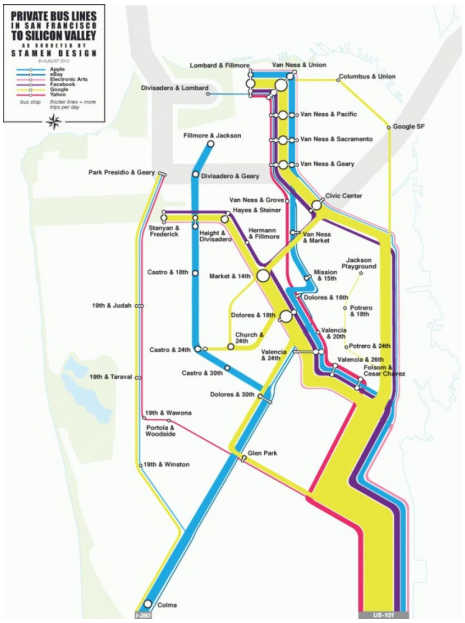
2.4 Geographical data of the city

San Francisco is divided into neighborhoods and parcels. Neighborhoods are geographical areas within the city that are defined by cultural, historical, and social boundaries. They are often used to describe the local community and its characteristics. Parcels, on the other hand, are individual pieces of land that are identified by a unique identifier and are used for property ownership and taxation purposes.
San Francisco's neighborhoods and parcels datasets were obtained from *DataSF*.
They both contain the coordinates of the vertices of the polygons.

2.5 Google Bus Stops

Google bus stops in San Francisco refer to the stops that were established by tech companies such as Google, Yahoo and Twitter for their private employee shuttle program in the city. Critics argue that the use of public bus stops for private transportation exacerbates the displacement of lower-income residents and that it accelerated the process of gentrification in San Francisco.[4]
In fact, the rapid growth of the technology industry has led to an influx of highly paid workers into the city, driving up housing costs and making it more difficult for low- and moderate-income residents to afford to live in the city.
For these reasons we decided to take into consideration the location of Google bus stops in our analysis.
Unfortunately, companies have never published official maps of routes and stops . In 2012, Stamen Design[5] tracked the movements of tech shuttles and created a map of them, so we were able to build manually our dataset with coordinates of big tech companies' bus stops. From now on, we will refer to them as Google bus stops.





2.6 Data processing

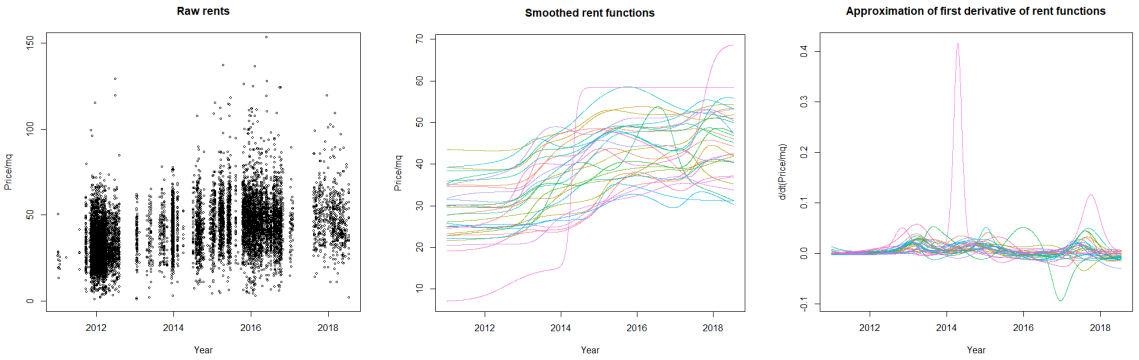
The original datasets were inconsistent with regards to the localization information. The new construction dataset contained neighborhood information and postal addresses of the permits, which were then used to obtain the latitude and longitude coordinates through the assistance of the Geoapify[6] software.
Although the evictions, parcels and rents datasets also contained neighborhood information, they did not have the same set of neighborhoods. This was due to differences in the level of granularity between the datasets: one dataset might consider a neighborhood as a separate entity while another considers it as part of a larger neighborhood. To ensure consistency in

the information across all datasets, it was necessary to align the neighborhoods in all sources.
The outcome of this process was to consider the neighborhoods illustrated in the following map:

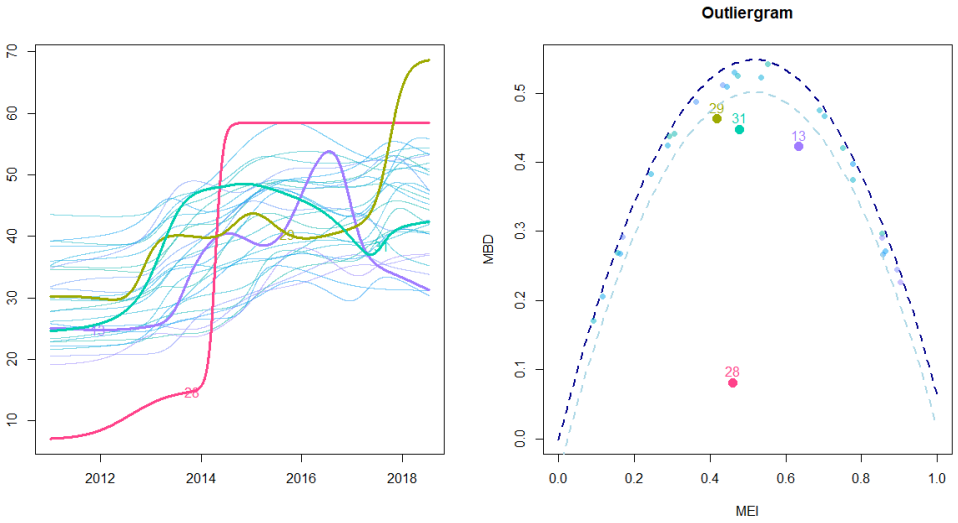
3 Trend Analysis

In this part we wanted to understand the main trends in time of the rent prices and evictions in the different neighborhoods and, in order to do this, we decided to see the two variables as functions of time.
Due to the scraping technique that was used to create the rent prices dataset[7], that data has a lot of missing observations for several periods of time. For this reason we opted for a gaussian kernel regression model to perform the smoothing taking advantage of the not compact support of the kernel function to avoid problems of the bad sampling of the data.
Once we had computed the functions, we were able to look not only at the main trends of functions but also in their derivatives.

3.1 Trends among rents functions

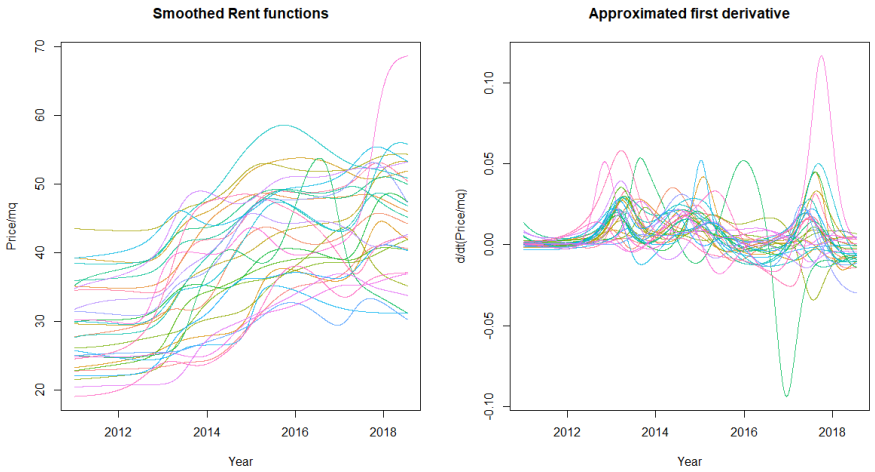


Starting from the functions of the rent prices (one for each neighborhood), we found an increasing trend common among all the neighborhoods but before proceeding with an interpretation of this result, we tried to perform an outlier detection.



Here we reported the outliergram for the rent prices. With this tool we spotted some neighborhoods as shape outliers and looking at them singularly, we found out that the most outlying of them was *Treasure Island,* an artificial island close to the north-east coast of SF, which is very scarcely populated and for which we had only a few observations. For this reason we decided to omit it from the analysis. Performing the same analysis on the derivatives, we spotted the same neighborhoods : Lakeshore (13), Twin Peaks (29), Western Addition (31) and Treasure Island (28).

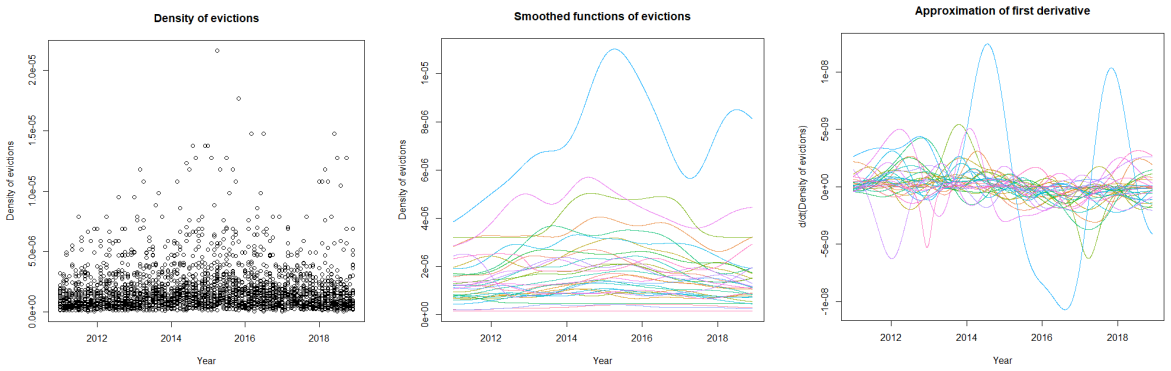
We decided to keep all the spotted neighborhoods except for Treasure Island, since they carry important information for the study.



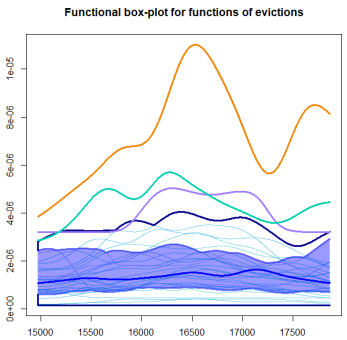
Observing the “cleaned” functions of rent, we can now confirm that there is clearly a positive trend in the prices.

3.2 Trends among evictions functions

In order to compare the trends between rent prices and evictions, we performed the same analysis on the evictions data. In order to take into account the different population of each neighborhood, we tried to normalize the number of evictions by the area, the number of parcels or the number of residential units of the neighborhood. Here we report the analysis of the eviction density (i.e. considering them normalized by the area), which provided the most reasonable results.



We performed also in this case an outlier detection and we spotted some magnitude outliers.

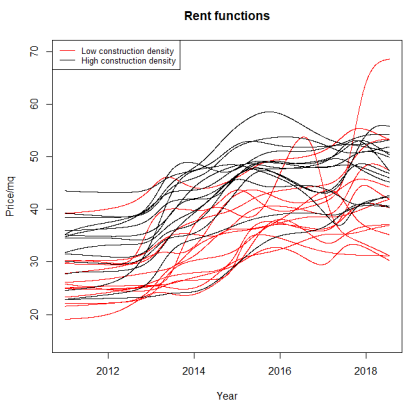


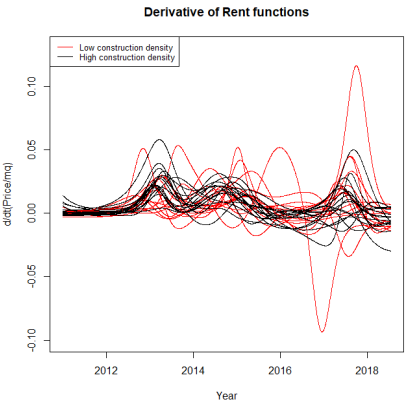
It is important to note that all the neighborhoods spotted as magnitude outliers (Japantown,Tenderloin and Nob Hill) belong to the north-east area of San Francisco, suggesting this area could be gentrifying faster than the others.
As before, we didn't omit these neighborhoods from the next analysis since they carry important information.

From this first analysis we can already see that the main trends of the rent prices and eviction functions seem very different and this suggests that there could be not a particular relation between them.

3.3 Inference on rent functions through permutational tests

In order to investigate the relation between rent prices and the new constructions, we partitioned the neighborhoods based on the density of new constructions in the whole period of study and then we tested if the two partitions seem to come from the same family of functions. Using as test statistics the integral of the absolute value of the difference of the two sample medians (computed with respect to the modified band depth) we obtained the following results:

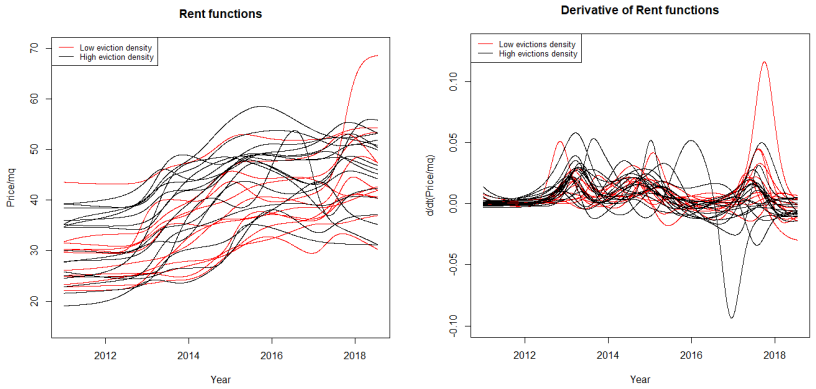




p-value = 0.031 *p-value = 0.561*

From this naive partition we can already see that there is a statistically relevant difference in the history of the prices between neighborhoods with high and low new constructions. On the other hand, we have no evidence to say that there is difference also in the derivatives. This can suggest that neighborhoods where we had a high density of new constructions, had different (higher) rent prices but the general trend was similar.

In order to further investigate also the relation between rent prices and evictions, we performed a similar test but with a partition based on the density of evictions instead of the constructions. We obtained the following results:



p-value = 0.378 *p-value = 0.748*

In this case the tests suggest that there is no difference between the two partitions both in the rent prices and in their derivatives. This seems to be coherent with the previous visual inspection of the general trends.

4 Constructions effect

4.1 Constructions effect on evictions at neighborhood granularity

We continued our study by implementing a semi parametric model having the number of evictions in a certain neighborhood (scaled by the neighborhood area and grouped by year) as the target variable. Our main goal was to understand how the average rent prices in the neighborhood and the new house units built in the past years affect the number of evictions, which is considered to be a proxy for displacement and gentrification. In particular, a high number of evictions is usually associated with an area experiencing gentrification as new and richer people are moving there and there are incentives for the landlords to evict the current (and poorer) tenants.
The first model we implemented was:

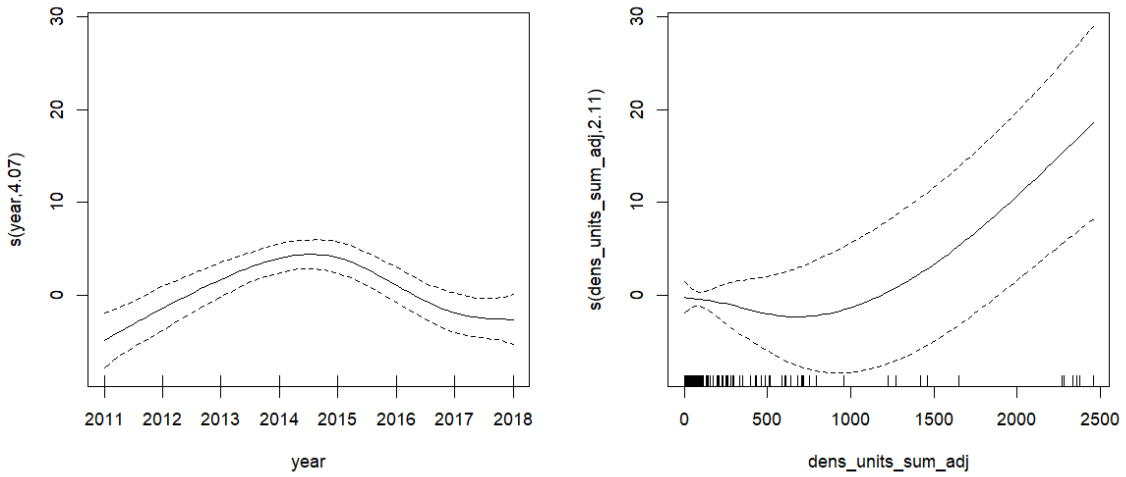
$$evictions_{year,nhood} = \beta_0 + \beta_{nhood} + \sum_{i=0}^{4} f_i(density\ new\ constructions_{nhood,year-i}) + f_5(year)$$

This model is significant , but the results on the variables related to the number of constructions for different years are not easily interpretable, and for this reason we decided to group them, obtaining the following model:

$$evictions_{year,nhood} = \beta_0 + \beta_{nhood} + f_1(density\ new\ constructions_{nhood,last\ 5\ years}) + f_2(year)$$

A smoothing with cubic splines for the time and construction density variables was used in both models .The results are satisfactory as the model is significant (R^2_{adj} is around 0.85) and it provides useful insights. Among the significant variables we find the covariates related to the space and time domain (namely the neighborhood and the year of the observation) and the number of constructions in the neighborhood in the 5 past years. Interestingly, the average rent price in the neighborhood does not affect the number of evictions in the

neighborhood, as the variable is not significant. This is supported by our previous findings, such as the non-significant permutational tests on rent, conducted on the partition induced by the evictions (3.3). Also a simple exploration of the plot of the trends of evictions and rent prices (3.2) lead to the same conclusion.
To sum up, we conclude that with respect to the city of San Francisco and the chosen time interval, rent prices and evictions are not equivalent measures of displacement and gentrification, but rather carry different information and thus they may bring to different conclusions.
The effect of the constructions in the last 5 years and of the time are shown:



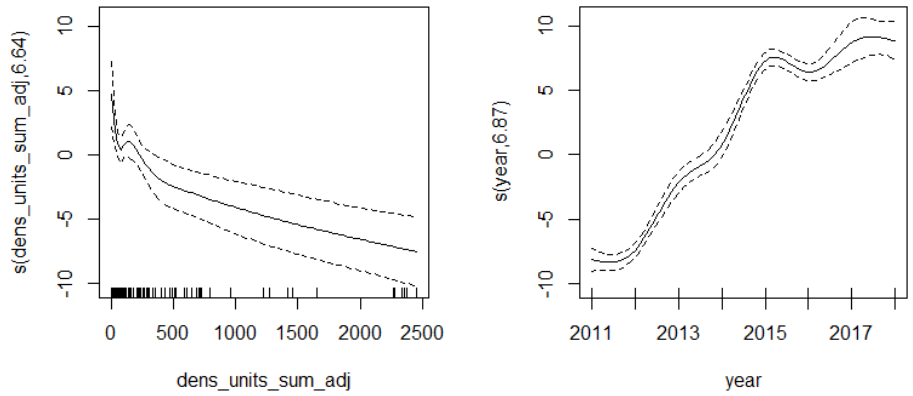
We see that a high number of new house units is associated with an increase of evictions, which would suggest a “renovation effect” in the neighborhood, leading to higher gentrification. We notice though that just a few observations have a high number of new constructions and this have likely altered the results, but there may be more. By a closer look, we notice that the only neighborhoods with such a high number of constructions are located in the north-east of SF, for example the Financial District, Tenderloin and Western Addition. This is the most expensive area in the whole city and rent prices have been rising significantly in all the years of study, making it one of the fastest gentrifying areas in the city. The areas in which properties are more expensive are more attractive to house builders, who have incentives in building there as their margins will be higher with respect to other areas in the city.
By further investigating these 3 neighborhoods, we notice that among them only Tenderloin has consistently a high number of evictions (it was marked as outliers in terms of evictions (3.2)), while Financial District and Western Addition are comparable to the other neighborhoods.
By removing Tenderloin from our dataset, the model maintains its goodness but the effect of new constructions on evictions is not considered to be significant anymore.
We conclude that our initial model is heavily influenced by a single group of observations (the ones in Tenderloin neighborhood) and that in all the other neighborhoods there is no clear relation between new constructions and eviction notices. Considering that our main goal is to study the relationship between new constructions and the gentrification process, we decide to focus only on rent prices from now on.

4.2 Constructions effect on rent prices at neighborhood granularity

In order to further investigate the impact of the new constructions (of the five last years) on the rent prices we built some semiparametric models.
The first attempt has been to study the effect of new constructions separately for each year. The results were not satisfactory and difficult to be interpreted and we moved to a different model, grouping five years of constructions in a single variable, obtaining the following semiparametric model:

$$rent\ price_{nhood,year} = \beta_0 + \beta_{nhood} + f_1(constr\ density_{nhood,year}) + f_2(year)$$

using a cubic splines basis for the time and construction density variables.
The model produced in this way has a low R^2_{adj} (around 0.37) but the dependence from the constructions (and also time and membership to a particular neighborhood) seems to be significant.

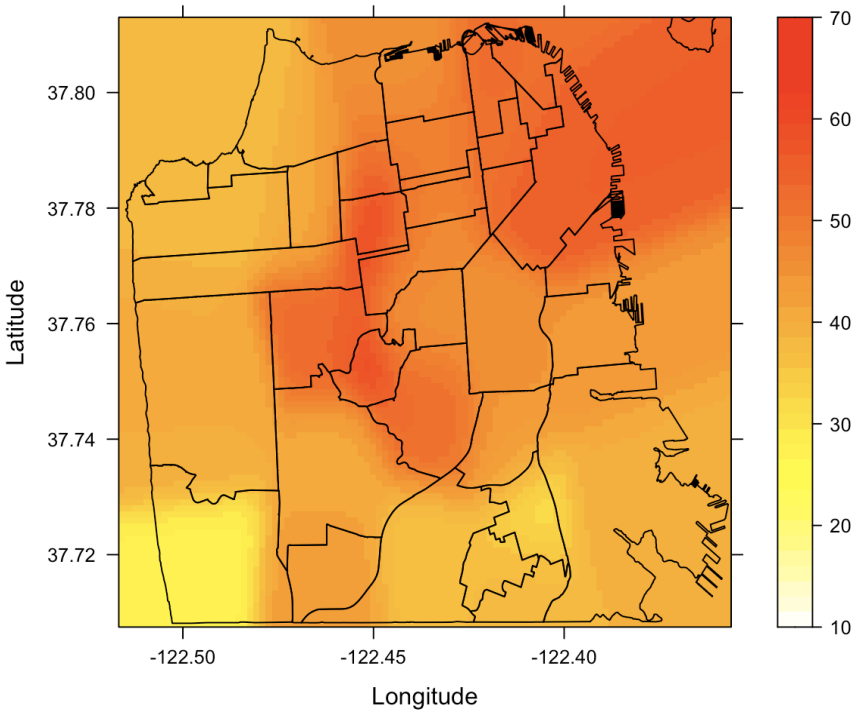


If we take a look at the relationship between rent prices and density of new constructions, we note that (keeping all the other covariates fixed) for low densities we can expect higher prices than for high densities. In addition, the general trend seems to be decreasing with high slope in the first part of the domain. This behavior can suggest that (again, keeping all the other covariates fixed) for low densities of new constructions the effect on prices is high because of a predominance of the "renovation effect" but as the value on the x-axis increases, the "supply effect" becomes more influential driving prices down.
Looking at the relationship between rent prices and year, we can find an increasing trend with a higher growth between 2012 and 2015 and a more stable behavior in the first and final periods.
In order to improve the quality of the model and to capture local effects of constructions, we decided to move from a neighborhood granularity to a parcel level.

4.3 Estimation of parcels rent prices through two-dimensional smoothing

In order to investigate the local effects of the new constructions, we needed the localization of the rent data. Unfortunately only a small fraction of the observations had the coordinates (or the address to be geocoded). Other studies on the topic[8] proposed to use a linear interpolation of the average prices in the neighborhoods to estimate the prices on the rest of

the city. Taking inspiration from this idea, we tried to overcome the problem of " not located advertisements " creating a (gaussian kernel) smoothing based on the averages of the rent prices of the neighborhoods for each year. Basically for each year, we computed the average price for each neighborhood and then used these observations (centered in the centroid of the correspondent neighborhood) to create a function with a 2d domain (*Longitude x Latitude*) representing the prices on the whole area of SF for a particular year.
In order to extract all the information we had in our data, we improved the model we have just described by introducing the small fraction of geolocalized rent data as observations (and not as part of the average rent price for their neighborhood) for the smoothing, giving different weights to the observations. In particular, the weight was proportional to the number of rent advertisements corresponding to the observation, namely the averages were weighted with the number of advertisements in the neighborhood for that year and the geolocalized advertisements with 1.
Here we have reported an example of this smoothing for one year.



4.4 Constructions effect on rent prices at parcels granularity

With the aim to deeper investigate the effect of building permits on rent prices, we performed a more complex semi parametric model at the parcel level.
Once we built the smoothing on the rent prices over the whole area of SF, we were able to estimate the prices on the location of the parcels.
Then we computed the number of new constructions located near each parcel. To determine the proximity of the new constructions we builded 4 circles of different radii (100 meters, 500 meters, 1 kilometer, and 2 kilometers) for each parcel, with the centroid of the parcels as centers. For each year and parcel, we counted the number of new constructions within the 100 meter circle and in the adjacent ranges of 500 meters to 100 meters, 1000 meters to 500 meters, and 2000 meters to 1000 meters.

This process allows us to quantify the number of permits in proximity to each parcel, for each year and in four different ranges.

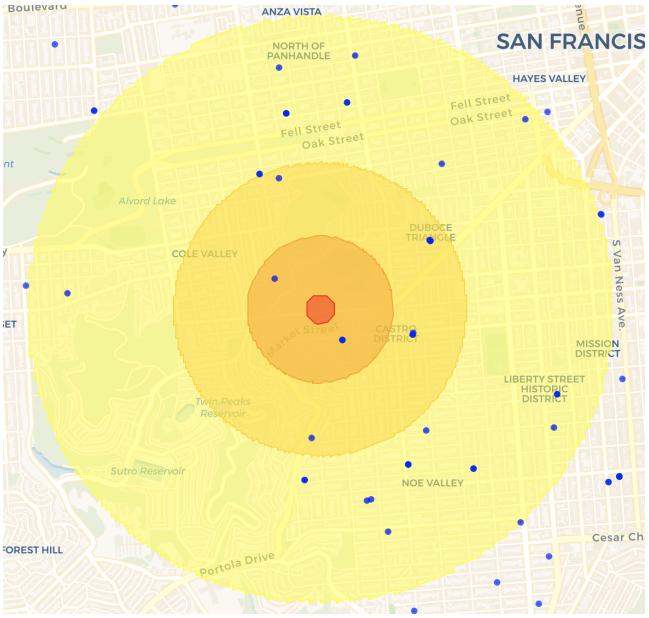


Fig.1

For each parcel we also calculated the distance from the nearest Google bus stop, to examine the impact on the rent prices.
With this setting we were able to build the following semiparametric model:

$$rent\ price_{year,location} = \beta_0 + \beta_1 \cdot min\ dist\ Google\ Bus + \beta_{nhood} + f_1(year) + f_2(latitude, longitude) + f_3(\#new\ constructions\ in\ 100m_{last\ 5\ years}) + f_4(\#new\ constructions\ between\ 100 - 500m_{last\ 5\ years}) + f_5(\#new\ constructions\ between\ 500 - 1000m_{last\ 5\ years}) + f_6(\#new\ constructions\ between\ 1000 - 2000m_{last\ 5\ years})$$

To estimate all the nonlinear functions, we used cubic splines basis except for f_2 for which we used a thin plate spline basis.
Passing from neighborhood level to parcel level we still find a significant dependence on the new constructions but we also increased a lot the R^2_{adj} which is around 0.85.
Regarding the effect of the nearest Google bus stop, we found that just a linear relation is sufficient to explain exhaustively the dependence. In particular we have that the associated estimated coefficient is β_1 = -0.74 meaning that increasing the distance (by 1km) makes the prices lower (by -0.74 $/mq). The effect is statistically significant but its impact is less important than the one of the other predictors.

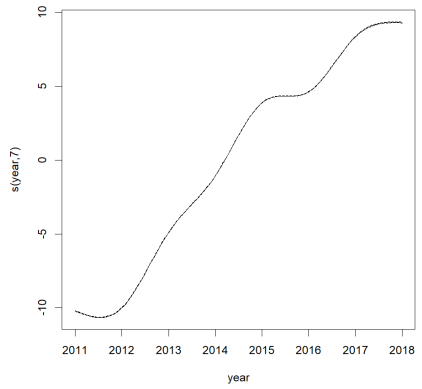


Fig.2

For what concerns the time dependence, we found an increasing trend with a stabilization in the periods 2015-2016 and 2017-2018 (as found in section 3.1).

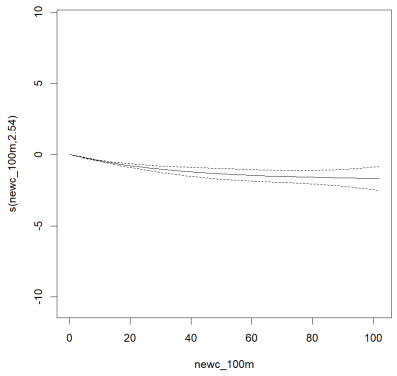


Fig.3

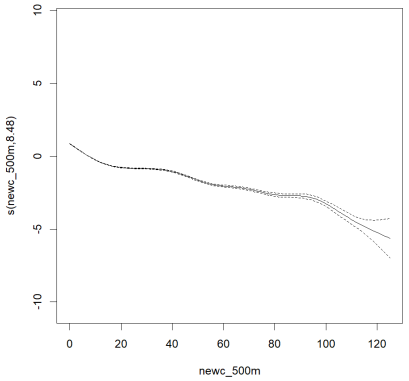


Fig.4

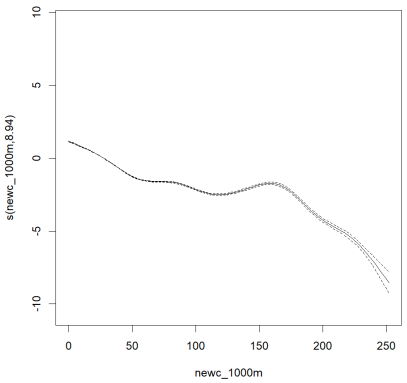


Fig. 5

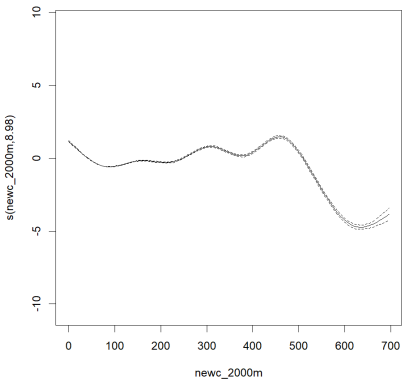


Fig.6

From figures 3-4-5 we can see that the main trend between the new constructions within 1 km and the rent prices is decreasing, namely we can expect lower prices where the supply of new houses is higher. However, considering the constructions in the “outer ring” (namely the area between 1-2 km) we can see that the prices become lower only for a consistent number of new constructions. This difference between the main trends of constructions within 1km and within 1-2km underlines how the effect of the constructions is local.

5.1 Conclusions

In our study we focused on two different proxies for displacement and gentrification, the rent prices and the evictions number. We were interested in understanding if these two measures have the same trend and if they lead to the same conclusions.
Considering the data visualization, the results of the functional permutational tests and the semiparametric model built at point 4.1, we can conclude that rent prices and eviction notices aren't two equivalent measures of gentrification.
The core of our study is the effect of new constructions of gentrification and in particular on its proxies, rent prices and evictions number. As we've just said, these results are different.
Specifically, the impact of new constructions on eviction notices results negligible for the neighborhoods, signaling that, in San Francisco and in our time period, building new houses did not lead to an increase or a decrease of eviction notices.
On the other hand, new constructions appear affecting the rent prices, both at neighborhood and at parcel level. It can be concluded that building new house units has lowered the rent prices and that this effect is larger whenever lots of new units have been constructed. By looking at the parcel granularity model, all the constructions within 1000m seem to decrease the price in the parcel similarly (whether they're in the 100m or in 500m range for example), while constructions further than 1 kilometer do not appear to significantly affect the rent price in the parcel, unless the number of construction is very high. This illustrates the importance of studying these effects locally, and that the neighborhood perspective is not sufficient to properly capture the complexity of the phenomenon.
A reasonable policy for tackling the gentrification and displacement phenomena is to encourage the construction of new house units in the city. In the meanwhile, monitoring the rent prices over the evictions number is to be advised, as rent prices trends allow to track the magnitude at which the city is gentrifying.
Regarding Google bus stops, we have found that they have a negative impact on the prices: the further away the nearest bus stop, the lower the prices.
This effect can be interpreted in different ways. Firstly, the landlords could consider applying higher rents, knowing that the houses near a Google bus stop could be attractive for tech workers (who can afford higher prices).
On the other hand, it could be also reasonable to assume that these bus stops are located by the companies near their employees' accommodation. Under this assumption, the increase of the prices due to the bus stops is actually driven in principle by the high concentration of tech workers in the area. This result suggests that tech workers are one of the driving factors of the rent prices increment and ultimately, gentrification.

5.2 Further improvements

The analysis could be improved considering the following ideas:

- Obtaining the official data on rents collected by the government with complete information about the whole housing market.
  This allows to avoid the intermediate step of the estimation of the prices on the parcels (Chp. 4.3), which obviously adds some error to the next models and is performed only to obtain geolocalized data.

- Obtaining data on evictions with precise information about the exact location
- Implementing survival models to estimate the time-to-displace, using a dataset with the recordings of the addresses of a group of SF citizens. Unfortunately these data are very sensitive and hard to obtain
- Considering a functional regression model to better capture the time dependence for the last model, namely considering both the response and the variables as functions of time
- Investigating the effects of different types of new constructions: affordable and not affordable new buildings could have different impacts on the prices.

References

1. https://sf.gov/ and its department DataSF
2. https://www.katepennington.org/data
3. https://archive.org/web/
4. https://www.reuters.com/article/us-google-protest-idUSBRE9B818J20131210; https://en.wikipedia.org/wiki/San_Francisco_tech_bus_protests
5. https://stamen.com/the-city-from-the-valley-57e835ee3dc6
6. https://www.geoapify.com/geocoding-api
7. https://www.katepennington.org/clmethod
8. https://www.katepennington.org/research