Classifying and Predicting Gentrification in New York City

Edward Cen, Vaibhav Ramamoorthy, Richard Huang, Chuck Tang

UC Berkeley

---

1. Executive Summary

Gentrification is a contemporary issue plaguing communities across the country, a phenomenon identified by the reconstruction and renovation of older, deteriorating neighborhoods into wealthy, affluent ones. Gentrification is problematic because it forces current residents to relocate due to higher costs of living and wipes out the community's preexisting culture. As such, it is a controversial topic of discussion and a cause for concern. In this paper, we analyze one unique possible predictor of impending or evolving gentrification among New York City neighborhoods: the quantity of 311 service requests received. People typically call 311 to report for non-crime related issues; i.e. rodent infestations, food poisoning, or air quality concerns. We feel that the number of 311 calls has potential as a predictor for gentrification because an increase in complaints about things like smoking or water quality might indicate that those from a wealthy background have moved in to the area and are calling in complaints about things that they regard as lower class; their standards of living also may be much higher.

Our motivation for this analysis was to help urban planning organizations and the government alleviate the effects of gentrification for displaced communities by helping them understand where gentrification is likely to occur in the coming years. Gentrification can have serious negative health effects for those affected, and we believe that these can be ameliorated if aid organizations know which neighborhoods are most at risk of gentrification. With a better understanding of which neighborhoods were at risk, these organizations might be able to save the livelihoods of so many people whose communities are at risk.

After developing a model to predict gentrification using 311 service calls,

we found that we could predict whether or not a neighborhood would experience gentrification with 74% accuracy. This result reinforced our notion that 311 service calls have strong predictive value for predicting gentrification and can help us identify which communities are at risk.

Finally, we make a series of policy recommendations so that agencies can take action to alleviate the effects of gentrification in soon-to-gentrify neighborhoods to help the local communities. Efforts to combat gentrification should focus their research on water quality, air quality, and sanitation, as these are the complaints that are most prevalent in communities prone to gentrification.

2. Overview of Data

For this analysis, we analyzed several datasets. Both external datasets were downloaded from `https://www1.nyc.gov/site/planning/data-maps/nyc-population/american-community-survey.page`.

- 311_*service_requests* (given): This table contained data on 311 calls placed by people across New York City. We focused on the number of calls per NTA (neighborhood tabulation area) per month for this analysis.
- 2013_*NTA_Demographic_Data* (external): This table contained demographic data for each NTA in New York for the year 2016. The fields we used include Population, Median Household Income, Median Home Value, and Percent of Individuals over the Age of 25 with a Bachelor's Degree.
- 2016_*NTA_Demographic_Data* (external): This table contained the same data as above but for the year 2016. We used the same fields as above.

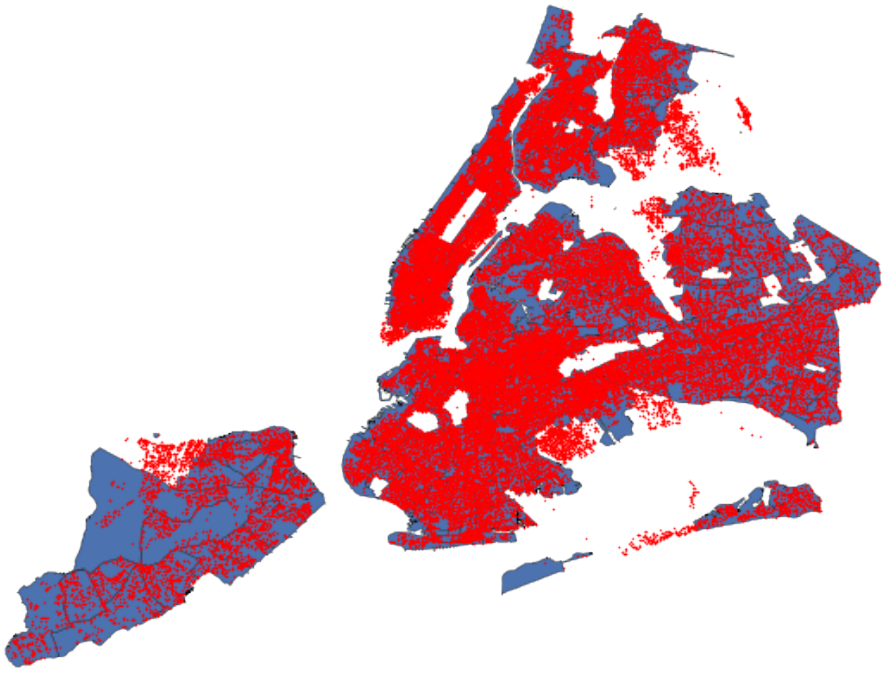
Given the above data sets, we aimed to answer the central question: Can we predict gentrification in different New York City NTAs using the trend of 311 service requests over time, specifically the preceding 36 months?

3. Exploratory Data Analysis

We performed all of our analysis on the geographic unit of NTA (neighborhood tabulation area), as this allowed us to consistently perform analysis

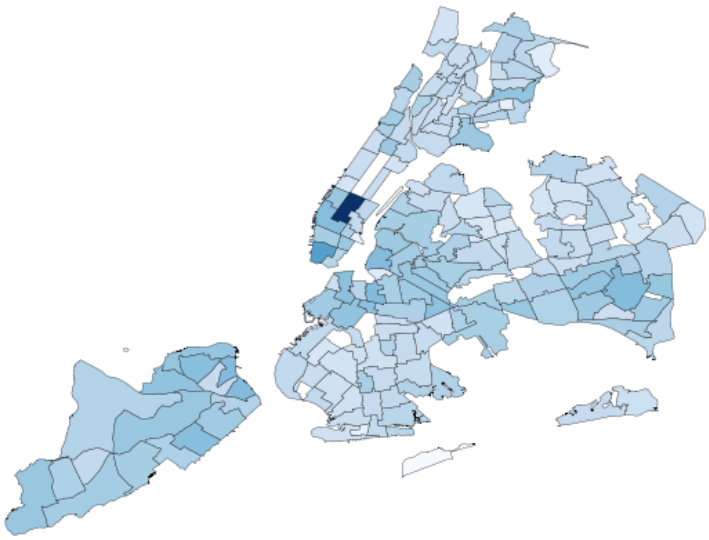
based on specific regions of New York City. Boroughs and zip codes given in the 311 service request table were too broad, and coordinates were too specific, as gentrification is usually defined as occurring on one or two city blocks. We created geospacial polygons in order to cluster our data points into the different NTAs, and then we assigned each 311 request to a NTA based on its coordinates.

We began by exploring the quality of the main data set, the 1.5 million 311 calls recorded in New York City between 2010 and 2018. We focused particularly on two indicators, whether there were many empty fields and how representative the data was of all of New York. For the first property, we found that only around 10% of the data had missing fields, which meant we had at least a million data points to work with. Furthermore, our model only used less than 50% of the fields, meaning only 5% of the data was unusable.



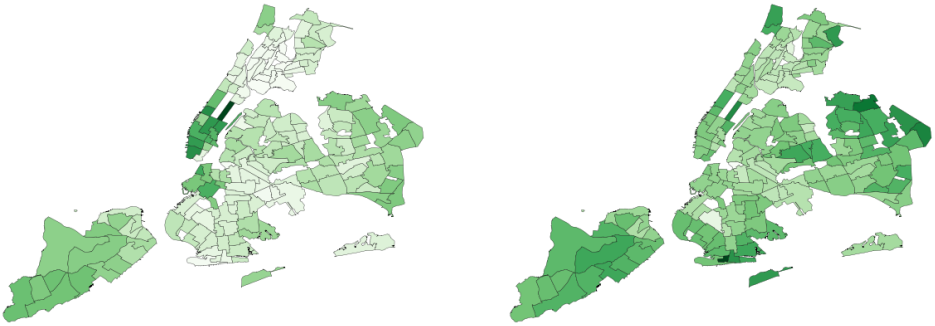
Distribution of 311 Calls

For the second property, we visualized all the service calls onto a map of New York City and found that all five boroughs of NYC seem to be accurately represented. Furthermore, this plot also showed that the 190 NTA's (Neighborhood Tabulation Area) provided in the dataset did not fully cover NYC's land area, as evidenced by the service calls lying outside any NTA region. Thankfully, only around 10% of the data did not lie in any NTA, and it seems regions are excluded from an NTA generally because they were not heavily populated, as evidenced by the missing Central Park block.



311 Calls per Capita

From the 311 Calls per Capita graph, we noticed that there were also clear differences in the number of calls among the different regions of New York. We initially wondered if filtering the 311 Calls per Capita by different complaints and plotting the frequencies of these complaints in areas that our model determined to be gentrified would yield strong associations. After looking at the counts of the different complaints such as rodent infestations, food poisoning, bad water quality, and air quality concerns, we realized that only a few of the complaints dominated the 311 requests. We hypothesized that if we were to plot the frequencies of the dominated requests over time in areas that we thought were gentrified, we could determine a relationship between 311 requests and whether or not a city is gentrified.



Median Income and Median Age

Upon inspecting the Median Income graph in particular, we saw that there is a clear distinction between income and median age. Large disparities can be seen between boroughs that are known to be wealthy like Manhattan, and other, less affluent areas like Fordham South, located North East of Central Park. Gentrification normally occurs in poorer areas, and the clear distinctions in wealth among NTAs and our knowledge of neighborhoods that have become gentrified recently such as Williambsurg and Harlem suggests that there are clear indicators for gentrification.

4. Determining Gentrification for each NTA

To create the dependent variable, we needed to label each NTA as having experienced gentrification or not between some range of years. We defined gentrification similarly to Prof. Lance Freeman of Columbia University, whose definition was republished on `http://www.governing.com/gov-data/gentrification-report-methodology.html`. The criteria for gentrification are as follows:

In order to be eligible to be considered for gentrification, the NTA must:

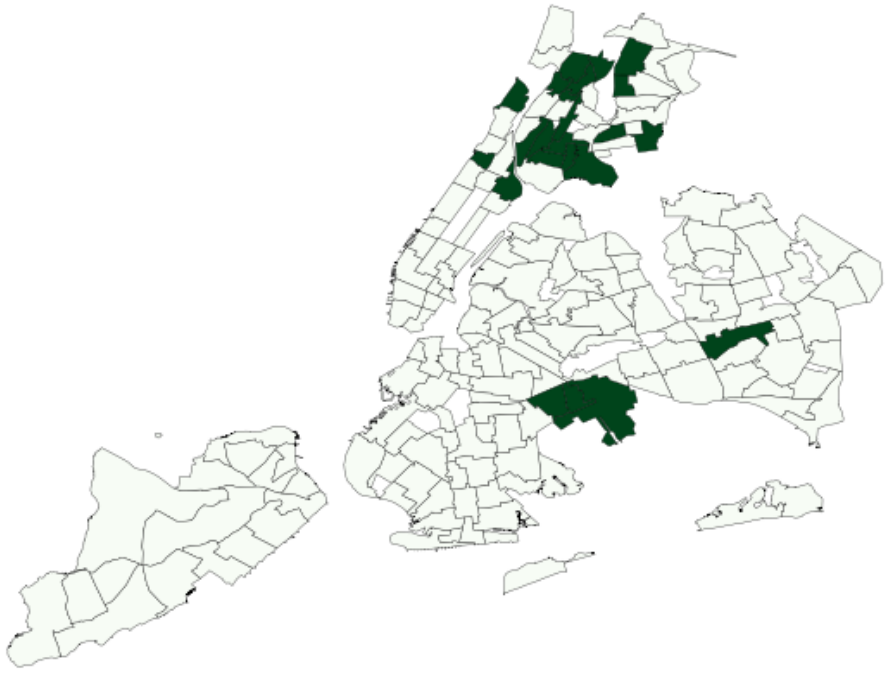
- Have at least 500 residents over the course of the years considered (all our NTAs satisfied this requirement)
- Have a median household income in the lower 40% of all NTAs
- Have a median home value in the lower 40% of all NTAs

Then, in order to actually experience gentrification, the NTA must:

- Have an increase in the percentage of those over 25 with a bachelor's degree over the range of years that ranks in the top third of all NTAs
- Have a percentage increase in the inflation-adjusted median home value over the range of years that ranks in the top third of all NTAs
- Have an inflation-adjusted increase in median home value

After reviewing the data, we chose to use 2013 as our baseline year (since it was the earliest year with clean and operable data from our external data source) and 2016 as our final year (since it was the most recent year with data). We used data from the same source (the American Community Survey as presented by the New York City Department of City Planning) as given from Correlation One/Citadel, but downloaded our own data because the data we were given was not as complete as needed.

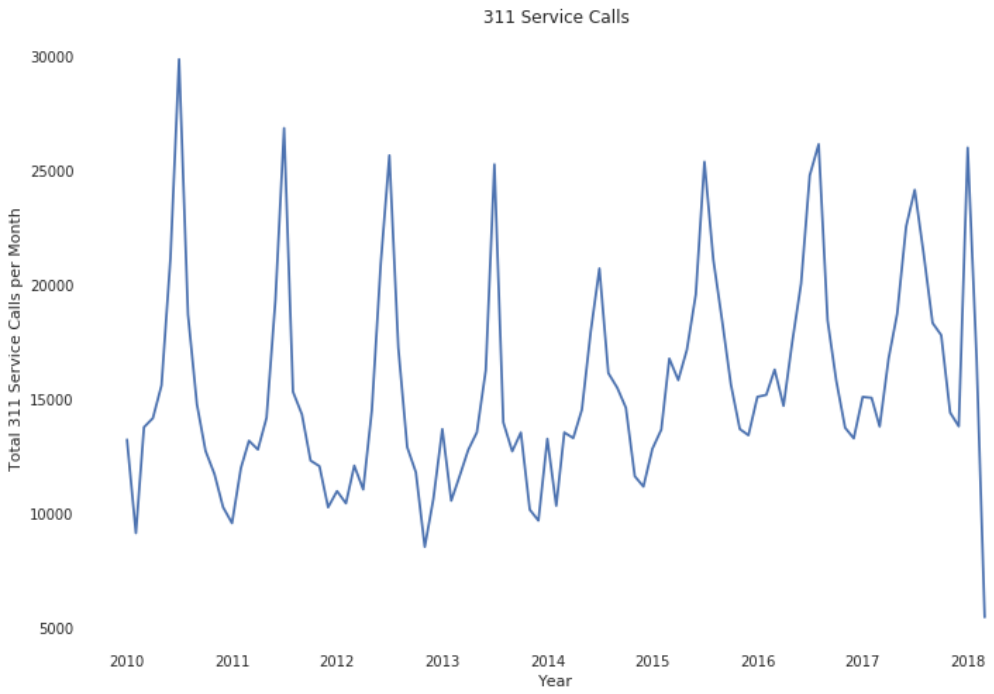
After labelling each NTA with a binary gentrification label, we plotted a map with gentrified NTAs, shown below:



Gentrified NTAs

23 of our 176 NTAs experienced gentrification between the years 2013-2016.

5. Unsupervised Model



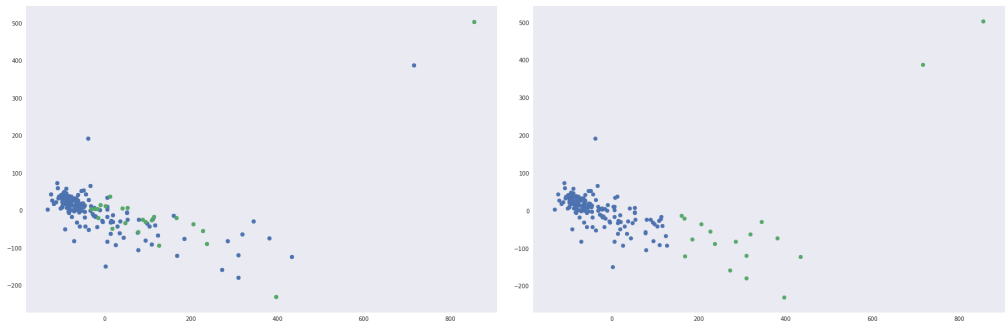
All 311 Calls Over Time

Upon plotting the temporal data, we clearly saw the seasonal trends, but could not identify a clear indicator of gentrification between the years we were interested in: 2012-2016. Because of the periodic nature of our plot, we thought that the DFT (Discrete Fourier Transform) may give us more insight on our data, as the DFT operates in the frequency domain.

DFT Plots of two NTAs, ungentrified (left) and gentrified (right).

Looking at the two plots, we can see that in the DFT domain, changes are much more evident. In the gentrified BK85 NTA on the right, the higher amplitudes in the frequency domain correspond to more changes in service requests. However, after applying the DFT, we were able to clearly see that changes exist; however, they are not easily interpretable in this domain. To clearly identify gentrified NTAs vs. un-gentrified NTAs, we need to move on to classification models.

After applying the DFT to our data, we made sure to demean and normalize our data to make sure that all of our data vectors were of unit length. We then reduced the dimensionality of each Fourier transform using PCA to the first few principal components for simpler k-means clustering. The variance of the principal components declined significantly after the first two, so we reduced each Fourier transform to two dimensions. Finally, we utilized k-means clustering to try and classify the different NTAs as gentrified neighborhoods and un-gentrified neighborhoods in a 4-dimensional vector space, as the the Fourier Transforms were complex valued.



The green dots refer to the gentrified regions (left) and clustered (right).

From a qualitative look at the clustering plots, it seems that the gentrified NTA's are mostly located near the bottom middle of the grid. Thus, considering the green cluster as the cluster most closely resembling the separation between gentrified and ungentrified NTAs, we have an unsupervised method which correctly labeled 8 of the 23 gentrified NTA's, and had 10 false positives, which is around 200% better than null methods as 23 of the 176 NTAs are gentrified.
Our ad-hoc unsupervised model did not have very impressive accuracy nor was it interpretable due to various transformations we applied. However, the big jump in accuracy suggests that there does exist a trend between 311 service call patterns and gentrification. Therefore, we felt confident that a more powerful supervised model could illuminate meaning relationships to make policy recommendations.

6. Supervised Model

We utilized a logistic regression model to determine if the number of 311 service calls in a particular NTA over a range of time is predictive of gentrification in that community. A description of how we built our design matrix and response vector are below:

6.1. Constructing the Design Matrix

To create the design matrix, we calculated the number of 311 service calls received from each NTA for each month in the 36 months preceding 2013. Then, we calculated the difference in number of 311 calls month-over-month to create the final design matrix. Each row represents the time series for a specific NTA and each column represents the difference in calls received month over month for each NTA.

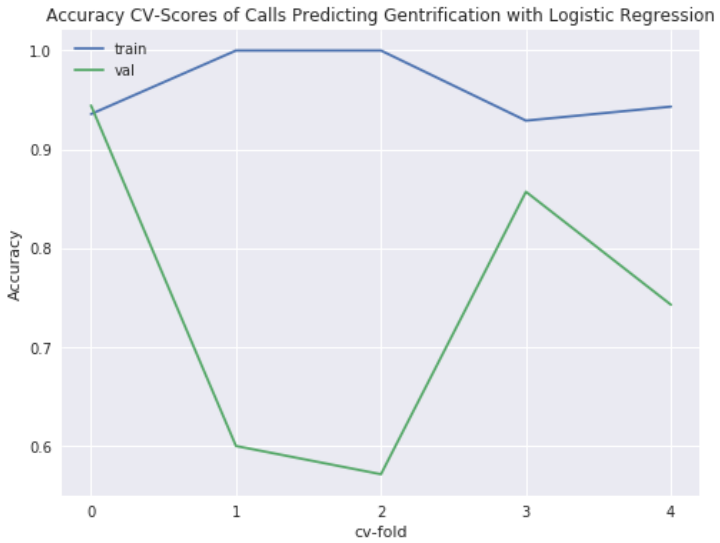
6.2. Logistic Regression Model

We ran a logistic regression model to calculate the probability that each NTA would gentrify over the next few years. We had 176 observations, one for each NTA, and 36 features, one for each month. We performed 5-fold cross-validation to train the model, using a split of the data into 80% training and 20% testing.

We also added 2 features to the model which calculated the maximum and minimum difference between months. These features were meant to identify NTAs which experienced large differences in call volume.

7. Results and Impact

Over the 5-fold cross validation, our average mean squared error was 0.21 with a 0.11 standard deviation and when classifying NTAs with a 50% probability or higher of gentrification as 1 and the rest as 0, our average accuracy was 78% with a standard deviation of 12%. The variance in these results are likely due to the small size of the dataset. These are pretty strong results and show that our model is quite good at predicting gentrification. The plot below shows the result of the cross-validation:



The coefficients for our model are shown below. Note that each variable month_a_b refers to the difference in the number of 311 calls received between month a and month b.

As you can see, the coefficients above show a trend. This is the trend that a soon-to-gentrify NTA would expect to experience in terms of 311 call volume.

Our model performed quite strongly, as the results above show. Therefore, we believe that this approach can become quite impactful in allowing governmental urban planning coalitions to alleviate the impacts of gentrification by being more aware of which NTAs are at risk.

8. Conclusion and Future Work

In conclusion, our next step would be to apply the model to the last few years of 311 call data and predict which NTAs are likely to gentrify in the near future. Agencies could then take action to alleviate the effects of gentrification in these soon-to-gentrify neighborhoods to help the local communities. Efforts to combat gentrification should focus their research on issues around water quality, air quality and sanitation as these are the complaints that dominate in the communities prone to gentrification.

Finally, an extension of this data analysis would focus on separating the 311 service calls into the different categories and finding the individual impact

of each type of service call. Using this instead of the blanket impact of all service calls could allow for more specific policy recommendations, such as targeting noise complaints specifically. Another recommendation for future work would be to apply our model onto other temporal demographic changes, such as racial and industry diversity of different NYC neighborhoods over time.

We conclude this case analysis with a clear recommendation for using 311 service call trends to predict, mitigate and combat gentrification in New York City.