Big Data Management
Spring 2017
Final Paper

Analyzing Noise Complaints in New York City

Xueqi Huang, Kristin Korsberg, Maisha Lopa, Ziman Zhou

I. Background

Noise in New York City remains an under-studied phenomena despite it being a key area of concern for many residents. In 2016 alone, the city's 311 system, which houses all complaints, received 405,282 noise services requests. The data collected is fairly robust, but existing research using that data falls short in many capacities. For example, it does not quantify noise using basic standardization methods, such as normalizing for population. Additionally, few publications study noise complaint trends through space and time. There is a lot of space for novel and interesting analysis.

II. Research Scope

The scope of this research project is twofold. First, we attempt to rectify methodological shortfalls (i.e. population standardization) of existing 311 noise complaint analysis by using a normalized metric to quantify noise volume. Second, we aim to identify specific factors to predict noise complaint volumes. The ultimate goal of this research project is to develop foundational knowledge of NYC's noise complaints, which can be used by researchers to continue developing novel and sophisticated analysis. The scope of work is outlined below:

- Correct methodological shortfalls by normalizing noise complaint volume by population size
- Identify the influence of land use, time, and population demography on noise complaint volumes
- Use the above-referenced features to build models that predict noise complaint volumes
- Visualize results

III. Methodology

A. Data Sources

311 Service Requests

NYC's 311 agency publishes service requests from residents between January 1, 2010 through the present. Our data captures all activity between January 1, 2010 and March 27, 2017. Each record includes information about the service request including date, time, type, and location. The dataset is over 5GB in size. Below is a snapshot of the fields required to generate noise complaint counts: (table put into LaTex as **Table1**, **Table2**)

zColumn Name	Unique ID	Incident Time	Incident Type	X Coordinate (State Plane)	Y Coordinate (State Plane)
Content	Complaint indicator	Time of complaint filed	Type of the complaint	Geographic latitude coordinate	Geographic longitude coordinate
Data Type	String	String	String	Float	Float

mapPLUTO

The mapPLUTO data captures land use and geographic composition at the tax lot level. There are over 45 features associated with each of New York City's 800,000 tax lots, so we selected six that captured the physical makeup of each lot: number of buildings, height of the tallest building, total tax lot value as defined by the NYC Department of Finance (estimated full market value by a uniform percentage for the property's tax class), building age, the amount of building area allocated as commercial, and building dimensionality as defined by building depth multiplied by building front.

American Community Survey
The United States Census' American Community Survey generates annual population estimates. From this database, we collected demographic data at the census tract level for each county in NYC. The features included population count; white, black, Asian, Hispanic/Latino, other populations; population over the age of 65; and median household income. There were 2,167 unique records in each ACS table.

US Census Cartographic Boundaries
The United States Census also produces shapefiles for different spatial units. We used a shapefile at the census tract level.

B. Data Processing

1. MapReduce on 311 Service Requests

The reduce job was implemented in four steps:

1) As mentioned above, each noise complaint record included geographic coordinates. The X and Y values from that dataset were zipped together to create a geographic point and then spatially joined to a census tract. The spatial join generated a new column associating each complaint record with a census tract ID. Due to the high volume of data, we implemented the R-Tree spatial indexing algorithm to improve efficiency.
2) Convert String to DateTime format
Each record in the complaint dataset also had a time column. The time column data was converted from *string type* to *DateTime object*, at which point time attribute methods could be called to quantify complaints by different temporal increments.
3) Create Keys
Before reducing on the whole dataset, each record needed to be assigned a unique key by time and census tract. The unique key was generated in the following format: (time, census tract number), e.g.: (hour = 17, '10782').
4) Reduce on the whole dataset
Once each complaint record was associated with a key, a reduce job was called using the 'add' function. The 'add' function generated a count of each unique key. We reduced the dataset into four separate temporal levels: hour, day, month, and year.

2. Spatial join on census tract and mapPLUTO

We started with three spatial units: incident-level data from 311 Service Requests, tax lot data from mapPLUTO, and US census tract data from the ACS. Incident level data was aggregated to the census tract with the above-referenced reduce method. The mapPLUTO data was aggregated to the census tract level using the R-Tree algorithm. Since each census tract was associated with more than one tax lot, we averaged all tax lot features within one census tract.

3. Unique record merge

With the tax lot data joined to the census tract shapefile, we merged each population demographic table from the ACS by unique census tract ID. Finally, we merged hourly, daily, monthly, and annual complaint count data to this dataframe, again using unique census tract ID. Our final outputs were four unique csvs, with population, land use, hourly, daily, monthly, and annual complaint volumes for each of NYC's census tracts. We dropped census tracts without any population, which reduced the size of our dataframe from 2,167 to 2,101.

(already in LaTex as **Table3**)

Year	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*
Counts	*188807*	*191035*	*218888*	*249372*	*326956*	*373777*	*408261*	*78012*

C. Modelling Workflow

Spatial Analysis

As referenced above, to date, analysis of 311 noise complaint data has been basic. As such, we built a model which not only predicts a volume of complaints per person but also accounts for spatial autocorrelation. Spatial autocorrelation evaluates whether behaviors exhibited by proximate geographies are random or not.

<<model explanation in place!>>

For the spatial analysis, we tested three models that corrected for spatial autocorrelation to predict complaint counts. We then selected the optimal model (out of three), which was the Spatial Lag Model and selected the optimal weight matrix (out of four), which was KNN-5.

Temporal Analysis

For the temporal analysis, we clustered 2,101 census tracts by the proportion of noise complaints per hour. We implemented two different clustering techniques, the first was KMeans and the second was hierarchical. Both KMeans and hierarchical clustering algorithms were run with several different cluster numbers. We visualized the results of both methods with color-coded maps and line graphs.

As a final step, we combined the results from both analyses and calculated the average cluster value for each demographic and geographic feature, plotting the difference in means in bar charts.

IV. Results and Findings

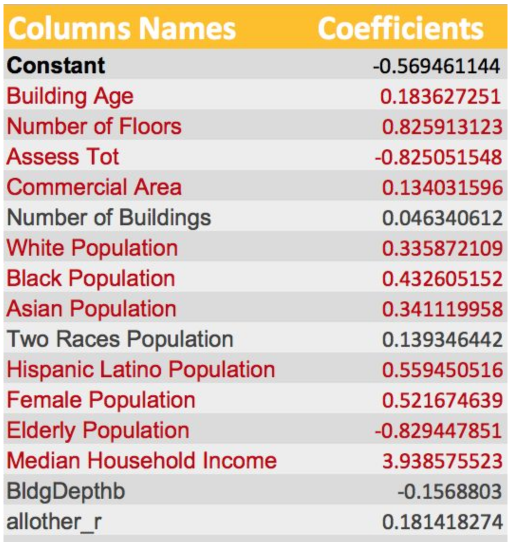
1)Spatial Analysis Result

Spatial Regression

The predictive model we build aimed to estimate noise complaint counts normalized by population. All demographic features were normalized by population, while physical features were standardized to a range between -1 to 1. This was done to avoid inaccurate distance measurements, which would arise based on the various scales of feature values. To understand spatial autocorrelation, we calculated spatial weight matrices that only consider direct neighbors.

By pairing each of the four weight matrices—KNN-3, KNN-5, rook, and queen— with the Spatial OLS, Spatial Error, and Spatial Lag models and comparing all the results, we find the Spatial Lag with the 5 nearest neighbors weight matrix to slightly outperform the other models-weight combination. The R-Squared was 0.570.

(Already in LaTex as **Table)**



Columns Names	Coefficients
Constant	-0.569461144
Building Age	0.183627251
Number of Floors	0.825913123
Assess Tot	-0.825051548
Commercial Area	0.134031596
Number of Buildings	0.046340612
White Population	0.335872109
Black Population	0.432605152
Asian Population	0.341119958
Two Races Population	0.139346442
Hispanic Latino Population	0.559450516
Female Population	0.521674639
Elderly Population	-0.829447851
Median Household Income	3.938575523
BldgDepthb	-0.1568803
allother_r	0.181418274

Table 4 shows the coefficients of the optimal Spatial Lag model for all features. Features in red are significant at the 0.05 significance level. It is worth noting that among the significant features, only the Assessed total financial value of a tax lot (AssessTot) and elderly population are negatively correlated with complaint counts. This means that for a census tract, the higher the financial value of the associated buildings, or the larger the proportion of elderly population living in that area, the fewer the complaint counts per person would be recorded. Furthermore, median household income is weighted the most, with a coefficient of approximately 3.98, which is much larger than that of the other features. This indicates that census tracts with higher income level are likely to have more reported noise complaints.

Spatial Autocorrelation

The Spatial Lag we derived can be visualized on a map as well:

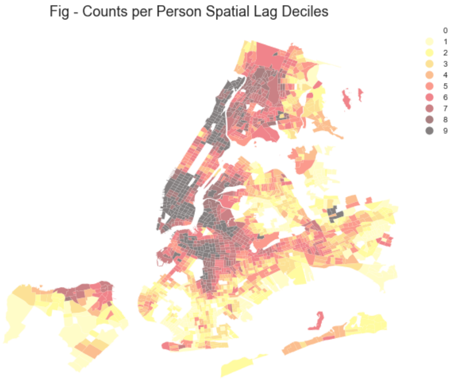


Figure 1: The map shows the study area that is colored based on 10 quantiles of **counts per person** spatial lag. Darker color indicates more counts of the spatial lag quantile.

The calculated Moran's I is 0.250, with an extremely low P-value: 2.72e-50. This indicates that spatial autocorrelation exists, but it is not very strong. Furthermore, using Local Moran’s I, namely LISA (Local Indicator of Spatial Autocorrelation), the correlation can be visualized in **Figure 2**: there are substantial amount of data points gathering in the bottom left region (the cold spot region) and the top right region (hot spot region). Yet there are still many points fall into the areas in which we are uncertain about the spatial effects associated with these points.

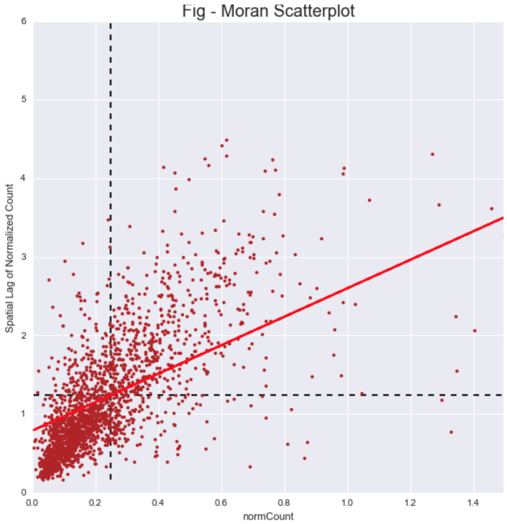


Figure 2: The scatter plot shows the correlation between normalized complaint counts and normalized count spatial lag with a fitted line. The dotted PCI lines divide the areas into four where the top right area contains the **hot spot** data points and bottom left are **cold spot** data points.

Spatial autocorrelation identifies census tracts where the count of complaints were similar to those of adjacent tracts. We find 236 hot spots and 424 cold spots at 5% significance level (see **Figure 3**).

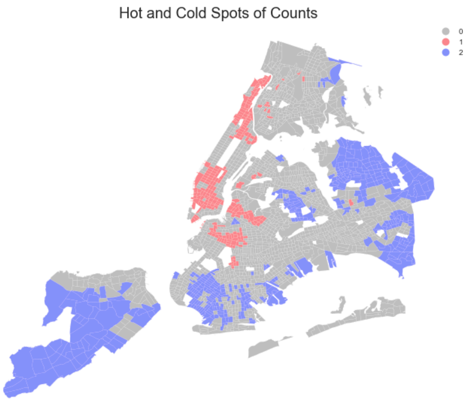


Figure 3: The heat maps show the hot spots and cold spots of noise complaint counts(normalized by population) based on Moran's I Statistics(Local), indicated by the red and blue colored regions.

It is obvious that spatial autocorrelation exists between neighboring census tracts (neighbors are weighted). The cluster of red regions indicates that the tracts with high complaint counts are likely to have adjacent tracts with high number of complaints; the cluster of blue regions shows that the tracts with number of complaints are more likely to have low counts in neighbor tracts.

2) Temporal Analysis Results

We decided to conduct two clustering techniques for temporal analysis. Of the two, KMeans and hierarchical, KMeans identified temporal trends with more interpretability than did hierarchical clustering. As depicted in **Figure 4** below, KMeans identified three distinct hourly noise complaint trends. The proportion of hourly noise complains peaks in the evening hours for all three clusters, dipping around 5AM and leveling off until about 7PM. Additionally, temporal trends in Clusters 0 and 2 are more similar than Cluster 1, which experiences lower peak volumes in the evening and higher complaint volumes during the day than its counterparts.

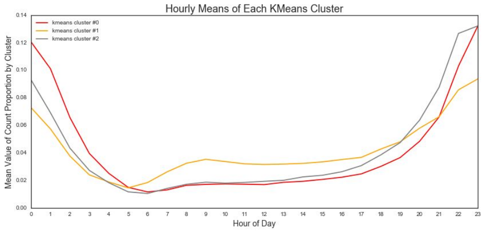


Figure 4: Line graph of the proportion of complaints per hour for each of the three KMeans clusters.

We decided to plot the results in a color-coded map of NYC as well, revealing the trends in Figure 5.

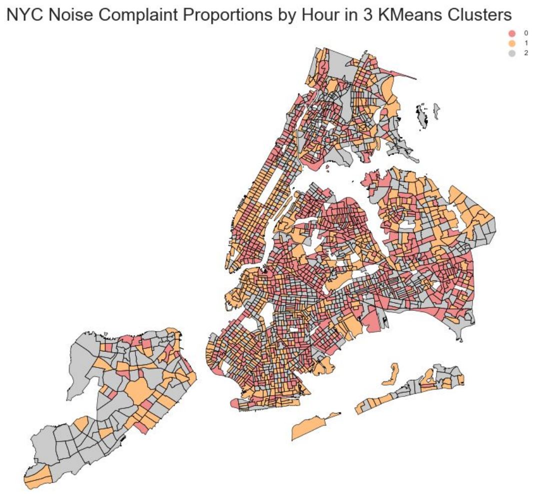

NYC Noise Complaint Proportions by Hour in 3 KMeans Clusters
0
1
2

Figure 5: Color-coded map of KMeans cluster by proportion of complaint calls per hour.

Here we can see that a majority of KMeans Cluster 1 is comprised of census tracts in Midtown Manhattan, Downtown Brooklyn, and LaGuardia Airport.

Unlike KMeans clustering, results of the hierarchical clustering were uninterpretable. As seen in Figure 6, when splitting the data into four separate clusters, hierarchical clustering placed 2,082 of the 2,101 census tracts into the same group, leaving some clusters to consist of just one census tract.

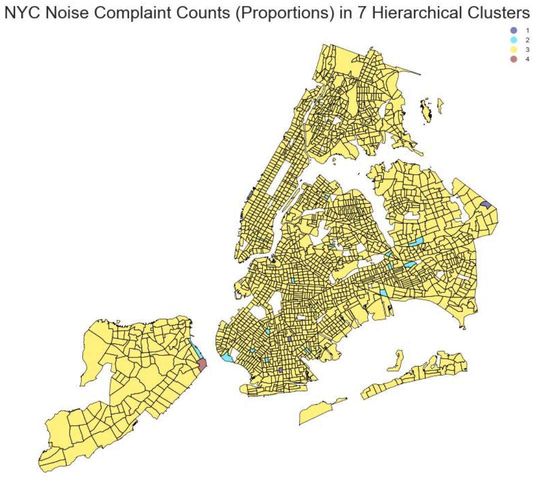

NYC Noise Complaint Counts (Proportions) in 7 Hierarchical Clusters
1
2
3
4

Figure 6: Color-coded map of hierarchical cluster by proportion of complaint calls per hour.

A review of a line plot of the proportion of complaint calls per hour per hierarchical cluster, was similarly discouraging. There were no discernible trends that we could explain with our features or domain knowledge.

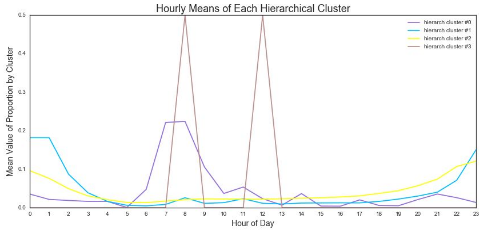


Figure 7: Line graph of the proportion of complaints per hour for each of the seven Hir clusters.

V. Big Data Challenges & Pipeline Bottlenecks

Our original dataset is comprised of 311 service requests between 2010 and 2017. The first dimension of our dataset can be defined as the number of records, which totalled over 2,070,000. The second dimension of the complaint dataset was the number of features per complaint, which was 51. The resulting dataset was over 5GB in size.

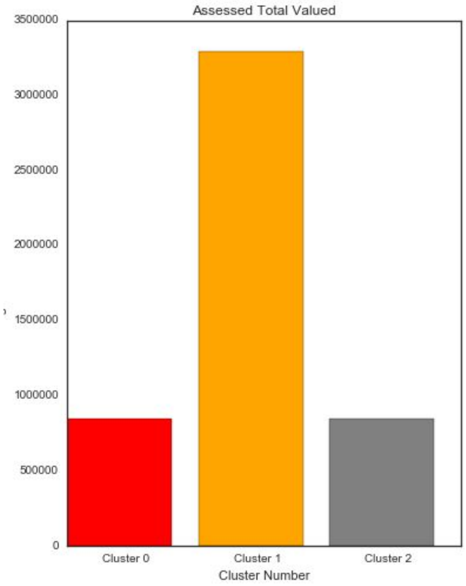
Our objective to analyze noise complaint counts required that we aggregate incident level data to a larger spatial unit. As such we aggregated complaints to the census tract level across four spatial units. The MapReduce framework was implemented to do so. We made use of the High Power Computing Clusters provided by NYU and ran a *PySpark* MapReduce job using cloud computing. It performs four times faster computing power than local machines and avoids the unnecessary memory space.

VI. Conclusions / Insights

Results of our spatial and temporal analyses generated unique insights into noise complaint trends across New York City. Specifically, the temporal analysis clustered the predominantly commercial areas of Midtown Manhattan, Downtown Brooklyn, and LaGuardia Airport into one group, which exhibited different trends than non-commercial census tracts. Specifically, non-commercial census tracts exhibit higher complaint volumes per person between 9PM and 5AM. At 5AM, commercial census tracts register greater complaints per person until 7PM. This result is unsurprising, as it mimics population flows to and from centers of business during working hours.

The results of our spatial analysis reveal that autocorrelation does exist and is significant in noise complaint patterns across New York City. It suggests that future regression analyses of complaint volumes should include a spatial weight matrix. It also identifies 11 census tract features that significantly impact such behavior, suggesting that population composition is a strong indicator of noise complaints volume. Every demographic feature we tested was statistically significant.

While our models tested different hypotheses, the results validate one another. For example, we did not use any land use or population features to cluster the hourly complaint data, however we identified differences in cluster means across the variables which were found to be statistically significant in spatial model. This can be seen in the figure below.



The average assessed total tax lot value in Cluster 1 (predominantly commercial census tracts) was 3,294,867 while the average value for tax lots in Clusters 0 and 2 were 847,238 and 845,539 respectively. This is interesting since it the spatial model found that higher assessed lot value was associated with lower complaint volumes, which is what we see in Cluster 1 of the temporal analysis.

Through our spatial and temporal analyses we were able to achieve our objectives. We predicted noise complaint volumes with the spatial model, identifying population and land use features which significantly affect such volumes. We also identified temporal trends trends across New York City, which validate the results of our spatial model.

VII. Future Work

As mentioned in the conclusion section, the temporal analysis revealed patterns consistent with population flows to and from places of work. A future model might capture population demographics not by place of residence, as the American Community Survey does, but by place of work. Similarly, we were unable to identify trends from the hierarchical clustering method, but this method could serve as anomaly detection. Additional work would have to be conducted to understand the relationships within each cluster.

VIII. References