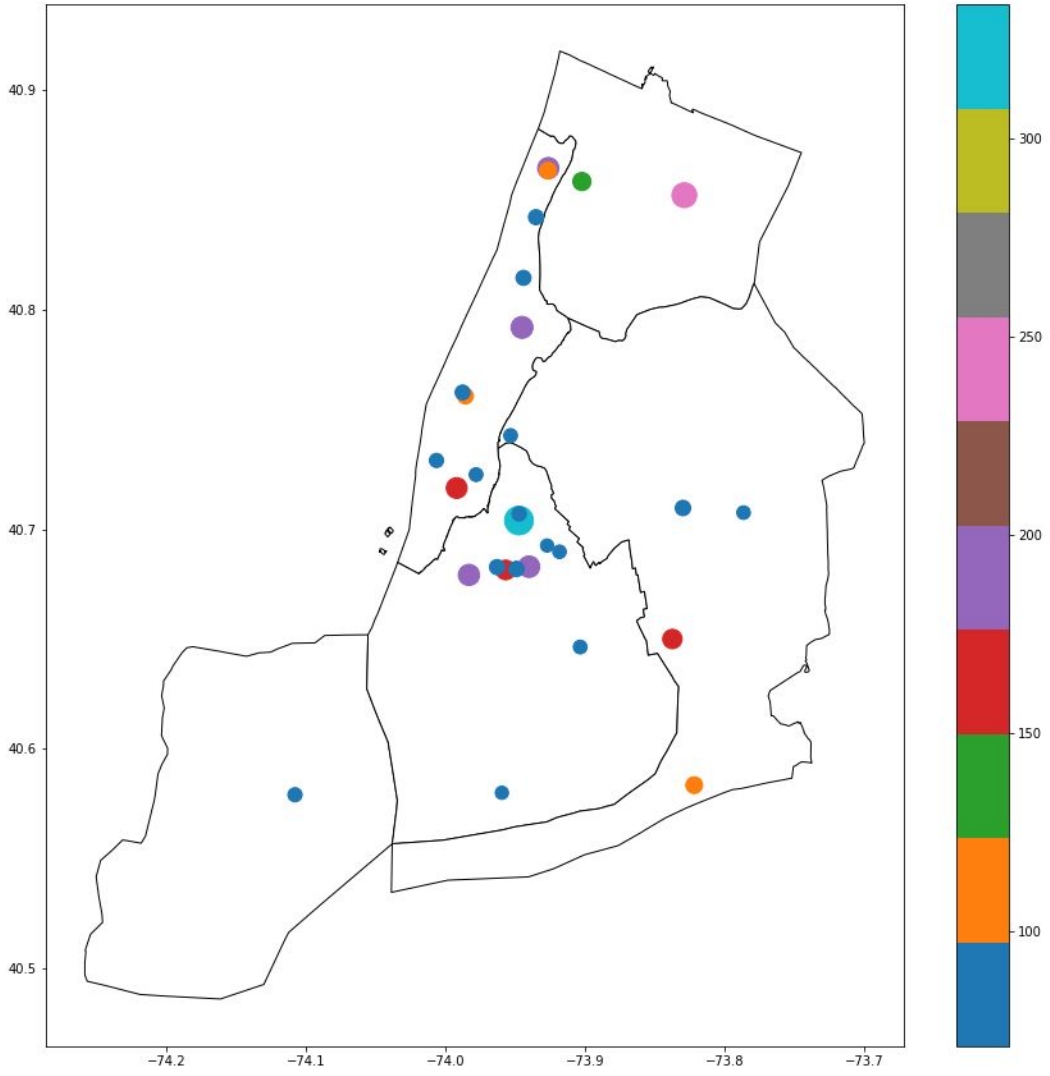
URBAN SPATIAL ANALYSIS PROJECT 2

IDENTIFICATION OF LOCAL HUBS OF NIGHT-TIME ACTIVITY OF NEW YORK CITY USING APPLIED SPATIAL ANALYSIS.



A PROJECT BY:

VAIDEHI VIDHYADHAR THETE

NYU CENTER FOR URBAN SCIENCE AND PROGRESS

MS URBAN INFORMATICS 2018-19

vvt221@nyu.edu

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank Dr. Professor Paul Torrens for giving me an opportunity to work on this project and also imparting the necessary knowledge and resources about the various clustering techniques without which the success of this project would not have been professor. Additionally, I would also like to thank teaching assistant Hai Lan for clarifying any doubts pertinent to this project.

DECLARATION

The following project submission is mostly my own work. It adheres to New York University's guidelines on plagiarism and Student Code of Conduct. All the materials used and referenced have been duly acknowledged in the bibliography and references.

TABLE OF CONTENTS

INTRODUCTION

PROBLEM STATEMENT:

Using spatial analysis, identify local hubs of night-time activity in New York City. Your response to this question must make use of at least three different forms of spatial analysis. Your analysis must make use of data that are initially non-spatial, but which you have made spatial to inform your analysis. Your analysis must also make use of data that are already spatial in form.

APPROACH:

I decided to identify the pubs, bars and restaurants of New York City as points of night-time activity as these places are ubiquitously frequented by people all the year round to enjoy, dine or simply unwind after a long day.

Using 311 noise data as a proxy for the popularity of the pubs, bars and restaurants, my analysis identifies the corresponding census tracts of these establishments as hubs of night-time activity in New York City.

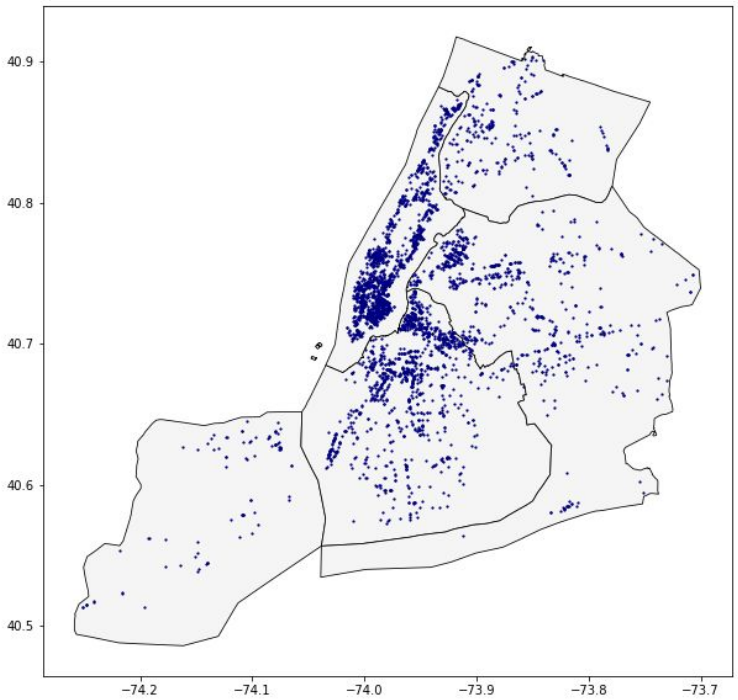


FIGURE 1: Plot showing the locations of all the 311 complaints made in the year 2016 for noise due to Pubs/Bars and Restaurants during night-time.

The spatial analysis techniques used for this purpose are :

1. Local Moran's I
2. Global Moran's I
3. Getis and Ord's G
4. Local Getis and Ord's G and G* statistic

This analysis was done using geopandas, pysal and shapely in a ipynb environment.

DATA COLLECTION AND PREPROCESSING:

The data used in the analysis is the 311 noise data taken from the nyc open data portal [1].
I decided to look at the calls made in the year 2016 at night time from 0600 hours to 1800 hours and I am using only those observations where the complaint type is Noise-Commercial and the location type is Pubs/Bars/Restaurants. I then aggregated the count of the complaints at the location level which is the longitude and latitude of the complaint location.

The rationale behind using the 311 complaints as a proxy for the popularity of the hubs is that if 311 complaints are made for a particular commercial establishment for loud music/party, one can safely conclude that the restaurant/bar receives substantial footfall and is popular among the patrons.

For the spatial data, I made use of the NYC census tracts 2010 as the shapefile and performed a spatial join of the 311 dataset in the course of my analysis.

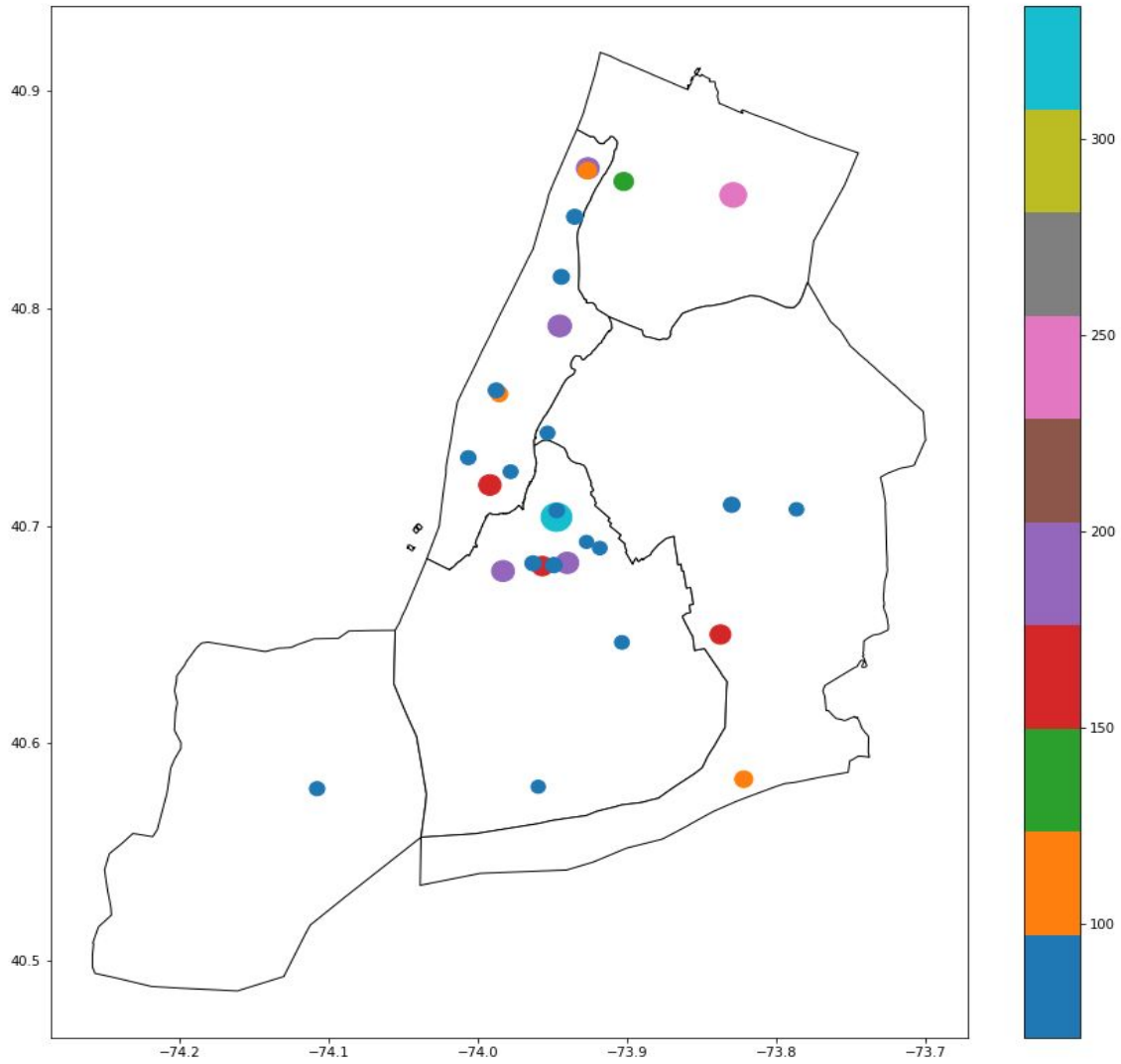


FIGURE 2: Plot showing the locations which recorded the highest number of complaints for bars/restaurants and pubs for the year 2016. The area of the circle increases in proportion to the number of the complaints received for that location and the corresponding counts can be found to the legend on the right.

ANALYSIS TECHNIQUES

BACKGROUND

Spatial autocorrelation pertains to the non-random pattern of attribute values over a set of spatial units. This can take two general forms: positive autocorrelation which reflects value similarity in space, and negative autocorrelation or value dissimilarity in space. In either case the autocorrelation arises when the observed spatial pattern is different from what would be expected under a random process operating in space.

Spatial autocorrelation can be analyzed from two different perspectives. Global autocorrelation analysis involves the study of the entire map pattern and generally asks the question as to whether the pattern displays clustering or not. Local autocorrelation, on the other hand, shifts the focus to explore within the global pattern to identify clusters or so called hot spots that may be either driving the overall clustering pattern, or that reflect heterogeneities that depart from global pattern.

GLOBAL MORAN'S I

Moran's I measures the global spatial autocorrelation in an attribute y measured over n spatial units and is given as:

$$I = n/S_0 \sum_i \sum_j z_i w_{i,j} z_j / \sum_i z_i z_i$$

where $w_{i,j}$ is a spatial weight, $z_i = y_i - \bar{y}$, and $S_0 = \sum_i \sum_j w_{i,j}$.

METHODOLOGY:

1. Spatial Join

First I performed a spatial join of the location of 311 complaints and their respective counts with that of the census tracts and plotted the choropleth of the 311 complaint counts across all the CTs.

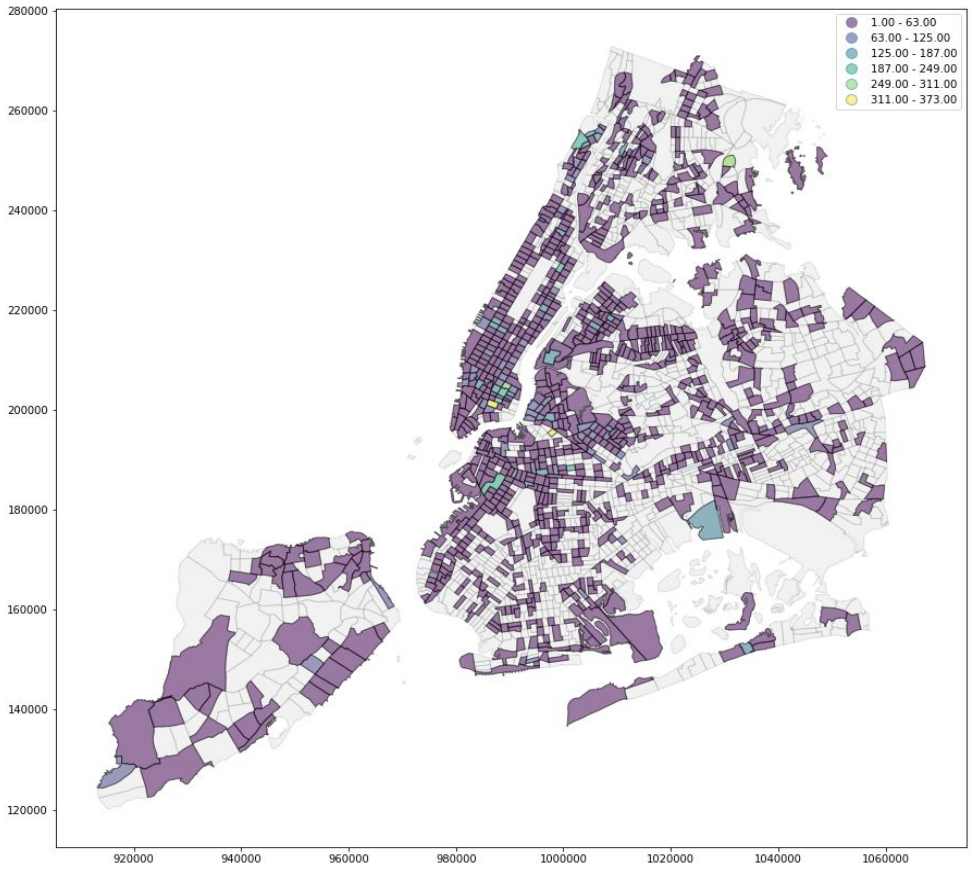


FIGURE 3: The plot above shows the counts of the 311 complaints aggregated at the census tract level. Notice the high noise clusters in Manhattan and the areas of Brooklyn adjacent to Manhattan which have higher complaints than normal.

From the plot above, it can be observed that the number of complaints ranges to about 63 annual complaints for almost all census tracts. However, some census tracts in Manhattan in Brooklyn show a spike in the number of complaints and they appear to cluster together.

2. Spatial Lag

In order to eliminate any randomness associated with the above observation we explore the concept of spatial autocorrelation. Spatial autocorrelation has to do with the degree to which the similarity in values between observations in a dataset is related to the similarity in locations of such observations. Spatial autocorrelation can thus be formally defined as the "absence of spatial randomness", which gives room for two main classes of autocorrelation, similar to the traditional case: *positive* spatial autocorrelation, when similar values tend to group together in similar locations; and *negative* spatial autocorrelation, in cases where similar values tend to be dispersed and further apart from each other.

We now build a row standardised contiguity matrix of these census tracts using the Queens Weights to define the neighborhoods. A spatial weights matrix is the way geographical space is formally encoded into a numerical form so it is easy for a computer (or a statistical method) to understand. Following this operation, we compute the spatial lag of these complaints which is basically the product of the spatial weights matrix and the value of the complaint counts. The spatial lag thus gives us the average count of the complaint counts in the neighborhood of each observation.

Quantiles of the spatial lag was then created to categorize the census tracts according to the spatially lagged complaint counts as follows:

Quantiles

Lower		Upper	Count
	x[i] <=	1.333	105
1.333 <	x[i] <=	2.500	101
2.500 <	x[i] <=	4.570	91
4.570 <	x[i] <=	7.000	104
7.000 <	x[i] <=	10.250	95
10.250 <	x[i] <=	15.000	100
15.000 <	x[i] <=	22.500	98
22.500 <	x[i] <=	33.000	100
33.000 <	x[i] <=	52.333	98
52.333 <	x[i] <=	256.000	98

FIGURE 4: Binning of the spatially lagged variable complaint counts into 10 bins

When we plot the spatially lagged counts onto the census tract, we find that that the higher complaint counts cluster around the southern side of Manhattan and the western side of Brooklyn. The decile map for the spatial lag tends to enhance the impression of value similarity in space.

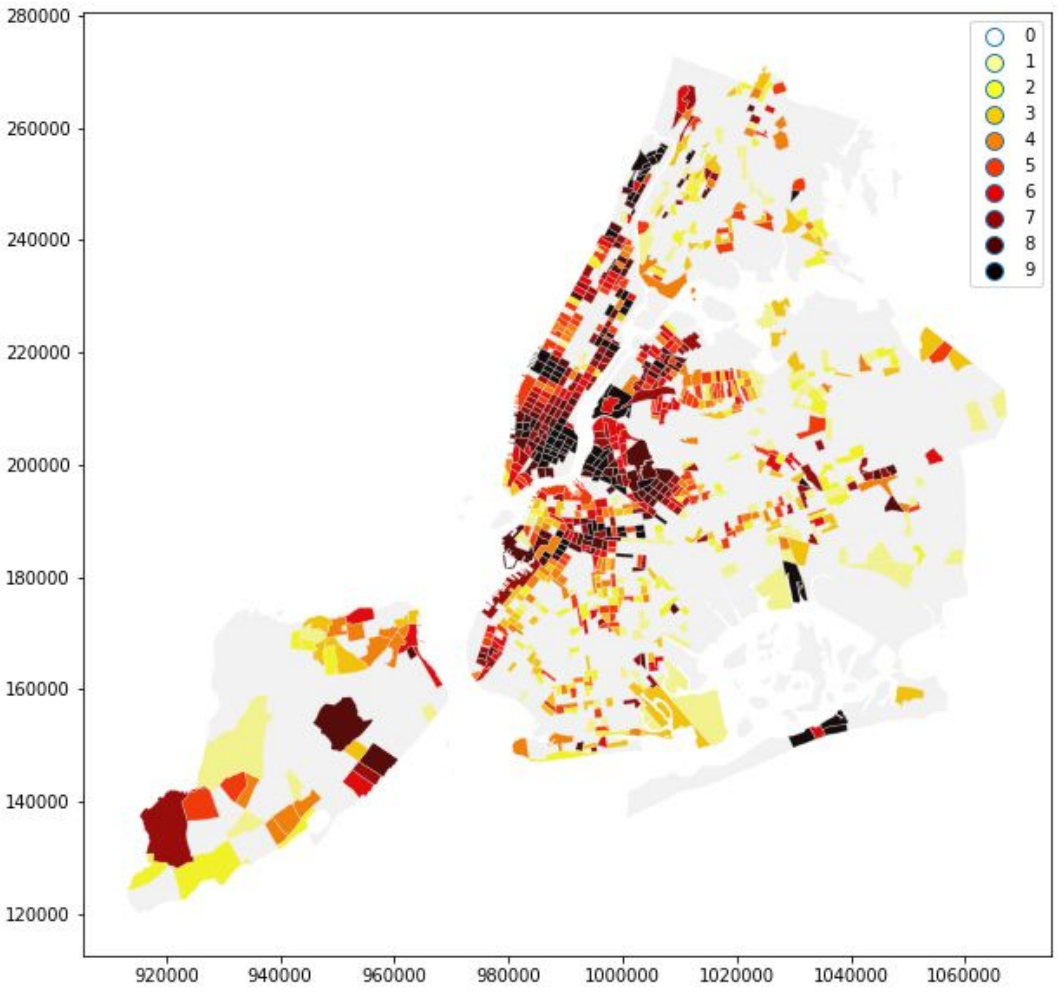


FIGURE 5: Plotting the spatially lagged variable onto the census tracts with the counts divided into 10 bins. Higher spatially lagged values appear to cluster together mostly in southern side of manhattan and western side of Brooklyn.

However, we still have the challenge of visually associating the value of the complaint counts in a census tract with the value of the spatial lag of counts of each census tract. The latter is a weighted average of complaint counts in the focal census tract's neighborhood.

3. Moran Plot

The moran plot is a way of visualizing a spatial dataset to explore the nature and strength of spatial autocorrelation. It is essentially a traditional scatter plot in which the variable of interest is displayed against its spatial lag.

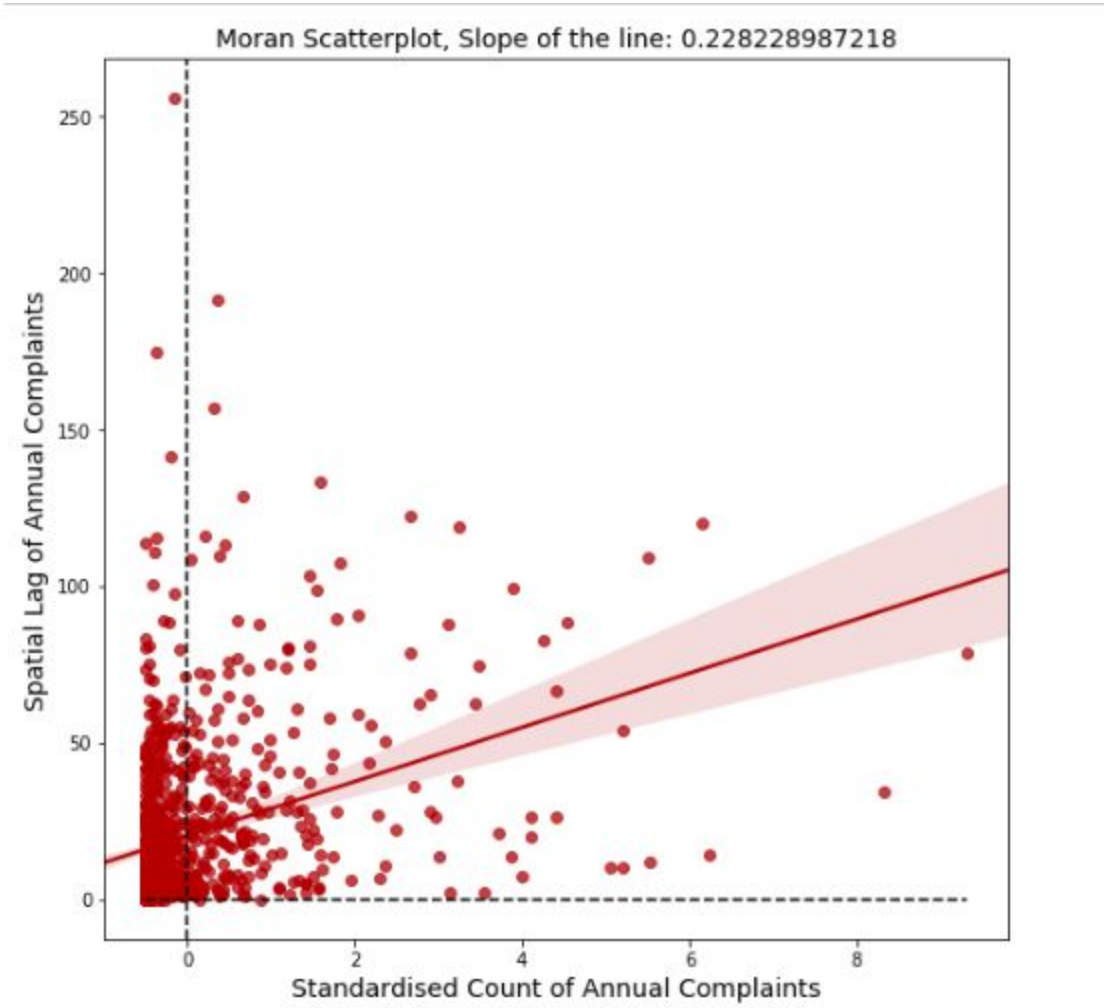


FIGURE 6: Moran Scatterplot displaying a relationship between the variable under observation (Complaint Count) and the spatially lagged value.

The figure above displays the relationship between standardised 311 complaint counts and its spatial lag which, because the weight matrix that was used is row-standardized, can be interpreted as the average complaint count density of cholera deaths in the neighborhood of

each observation. In order to guide the interpretation of the plot, a linear fit is also included in the post, together with confidence intervals. This line represents the best linear fit to the scatter plot or, in other words, what is the best way to represent the relationship between the two variables as a straight line. Because the line comes from a regression, we can also include a measure of the uncertainty about the fit in the form of confidence intervals (the shaded red area around the line).

The plot displays a positive relationship between both variables. This is associated with the presence of *positive* spatial autocorrelation: similar values tend to be located close to each other. This means that the *overall trend* is for high values to be close to other high values, and for low values to be surrounded by other low values. This however does not mean that this is only situation in the dataset: there can of course be particular cases where high values are surrounded by low ones, and vice versa. But it means that, if we had to summarize the main pattern of the data in terms of how clustered similar values are, the best way would be to say they are positively correlated and, hence, clustered over space.

In the context of the example, this can be interpreted along the lines of: census tracts in the dataset show positive spatial autocorrelation in the density of complaint counts. This means that street segments with a high level of complaint counts tend to be located adjacent to other complaint counts also with high number of complaint counts , and vice versa.

4. Moran's I

While the Moran Plot is an excellent tool to explore the data and get a good sense of how much values are clustered over space, it is sometimes hard to condense its insights into a more concise way. We therefore use a statistical measure that summarizes the figure. This is exactly what Moran's I is meant to do.

The Moran's I when calculate using Pysal library yields the following value with its associated p value:

```
In [380]: I_AnnualComplaints = ps.Moran(AnnualComplaints, qW_CT)

In [382]: I_AnnualComplaints.I , I_AnnualComplaints.p_sim
Out[382]: (0.22923755579085198, 0.001)

In [783]: I_AnnualComplaints.EI
Out[783]: -0.0010111223458038423
```

FIGURE 7: Code snippet showing the calculation of the Moran's I statistic along with its associated p value and the expected value of the I statistic under the assumption of normality.

Thus, the I statistic is 0.229 for this data, and has a very small p value of 0.001. I also want to point out that the Moran's I statistic is equal to the slope of the moran plot. Thus, Moran's I captures much of the essence of the Moran Plot.

The I statistic calculated is different from the expected value of I under the assumption of normality. We visualise the distribution of the simulated p values using Kernel density estimation. We plot the I statistics of the simulated statistics and comparing it against the actual and the expected I statistic.

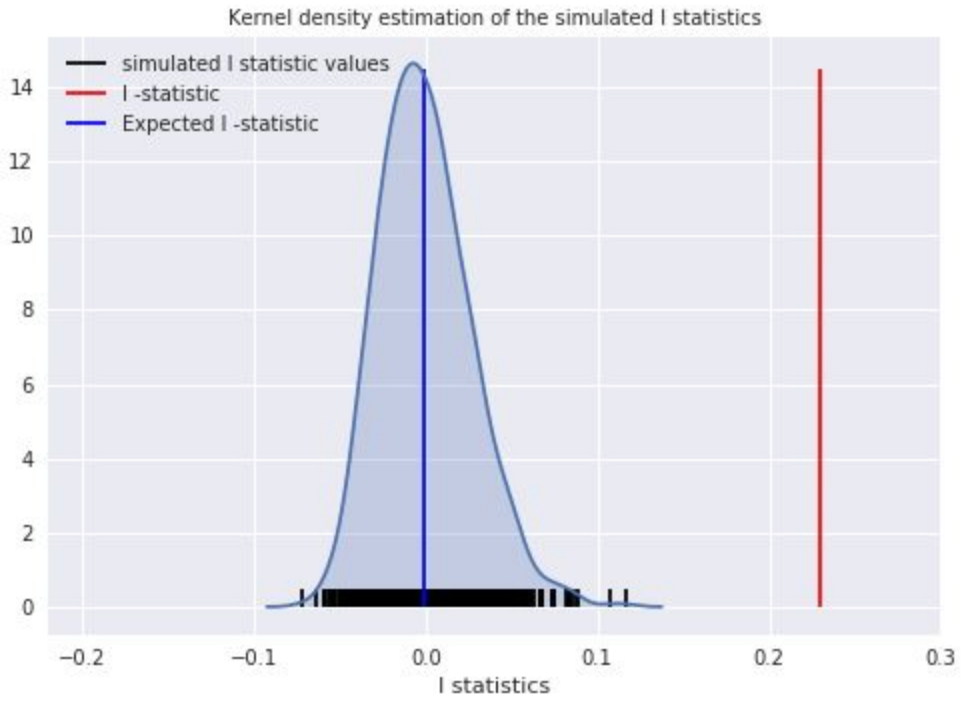


FIGURE 8: Kernel density estimation of the simulated I statistics and the actual calculated I statistic shown in the red line . The expected I statistic value is also shown in the blue line.

The p-value associated with the I static is significant since it is below 5%. It is interpreted as follows: if we generated a large number of maps with the same values but randomly allocated over space, and calculated the Moran's I statistic for each of those maps, only 0.1% of them would display a larger (absolute) value than the one we obtain from the real data, and the other 99.9% of the random maps would receive a smaller (absolute) value of Moran's I. Thus, we reject the null hypothesis that the complaint counts are randomly distributed over the place. Thus, the values are positively correlated with the space. Thus we can conclude that the census tracts are indeed hubs of restaurant establishments which are frequented by a lot of people and are very popular.

LOCAL MORAN'S I

PySAL implements local Moran's I as follows:

$$I_i = \sum_j z_i w_{i,j} z_j / \sum_i z_i z_i$$

which results in n values of local spatial autocorrelation, 1 for each spatial unit.

METHODOLOGY

Global spatial autocorrelation does not suggest the location of clusters. For that purpose, we need to use a *local* measure of spatial autocorrelation. Local measures consider each single observation in a dataset and operate on them, as opposed to on the overall data, as *global* measures do. Because of that, they are not good a summarizing a map, but they allow to obtain further insight.

At the core of these method is a classification of the observations in a dataset into four groups derived from the Moran Plot: high values surrounded by high values (HH), low values nearby other low values (LL), high values among low values (HL), and viceversa (LH). Each of these groups are typically called "quadrants". An illustration of where each of these groups fall into the Moran Plot can be seen below:

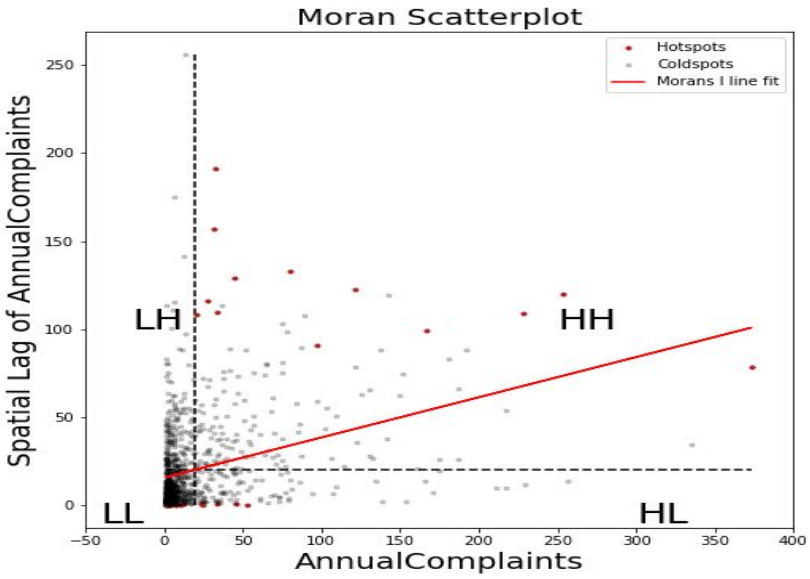


FIGURE 9: Classification of the Moran Plot into 4 regions: High-High, High-Low,Low-Low, Low-High

To know whether each of the locations is a *statistically significant* cluster of a given kind, we again need to compare it with what we would expect if the data were allocated in a completely random way. After all, by definition, every observation will be of one kind of another, based on the comparison above. However, what we are interested in is whether the strength with which the values are concentrated is unusually high.

The core idea is to identify cases in which the comparison between the value of an observation and the average of its neighbors is either more similar (HH, LL) or dissimilar (HL, LH) than we would expect from pure chance. The mechanism to do this is similar to the one in the global Moran's I, but applied in this case to each observation, resulting then in as many statistics as original observations.

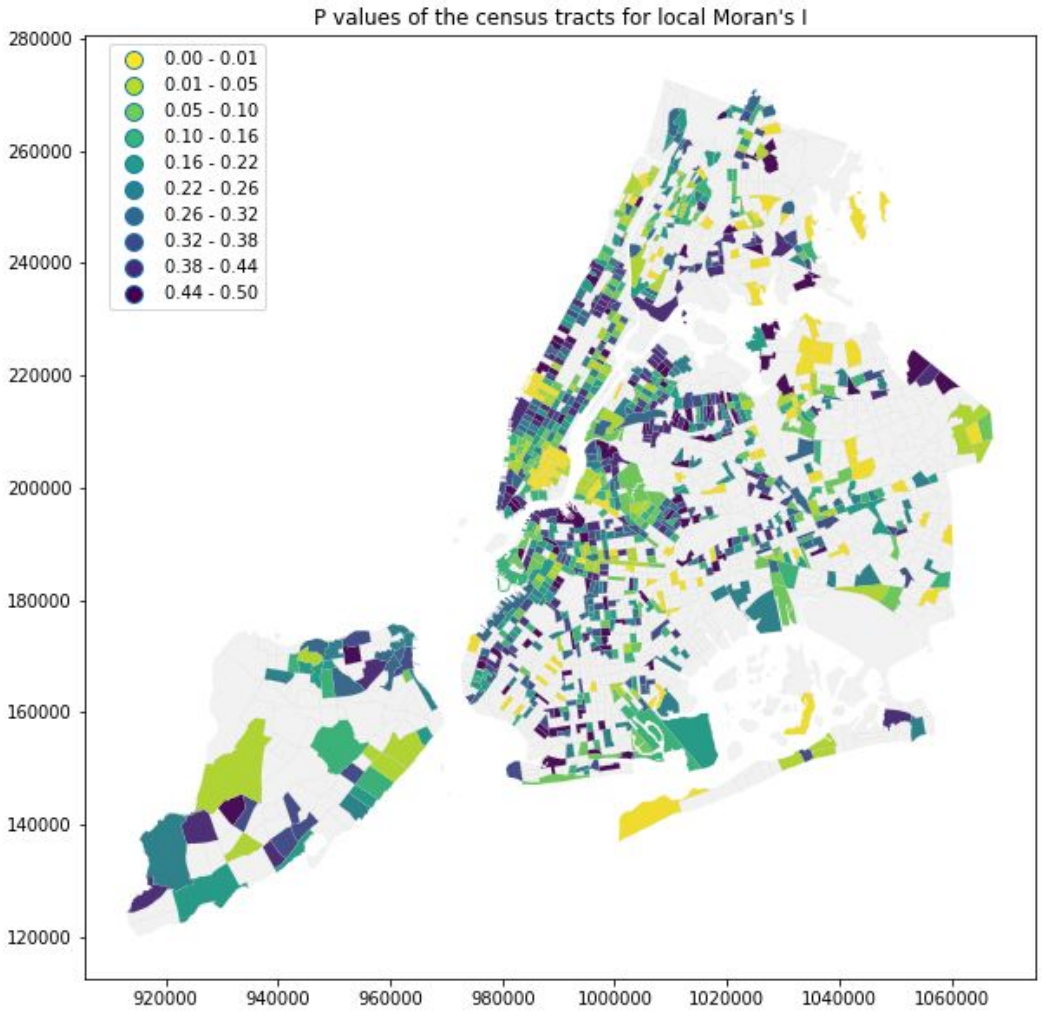


FIGURE 10: The above plot shows the p values calculated for the census tracts with the local moran's I. The most significant p values are displayed in yellow.

Thus, LISAs focus more on the spatial instability. They provide an insight into those areas which mark a departure from the general trend.

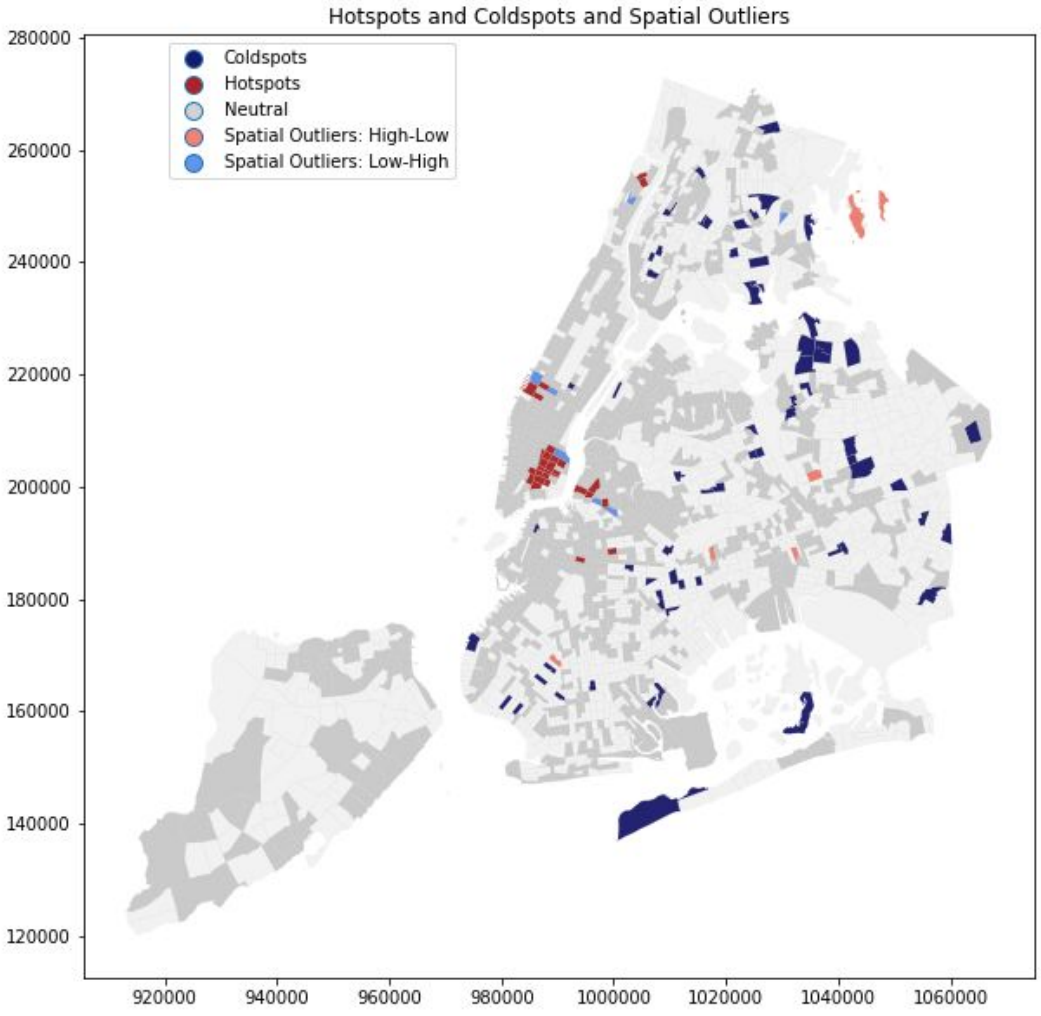


FIgure 11: The plot shows the hotspots ,coldspots and the other spatial outliers(high-low and low-high regions). If compared to the p values plot above, one can observe that the clusters are almost always formed in the areas of significant p-value.

Thus we are able to see how the hotspots are concentrated in Manhattan and Brooklyn which is not surprising since Manhattan's night life is famed around the world. Coldspots are distributed all over New York City. The low-high regions tend to occur near the hotspots whereas the high-low regions almost always occur in isolation.

GETIS AND ORD'S G

The General G statistic of overall spatial association is given as:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \ \forall j \neq i \tag{1}$$

where x_i and x_j are attribute values for features i and j, and $w_{i,j}$ is the spatial weight between feature i and j. n is the number of features in the dataset and $\forall j \neq i$ indicates that features i and j cannot be the same feature.

The z_G-score for the statistic is computed as:

$$z_G = \frac{G - \mathrm{E}[G]}{\sqrt{\mathrm{V}[G]}} \tag{2}$$

where:

$$\mathrm{E}[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}}{n(n-1)}, \ \forall j \neq i \tag{3}$$

$$\mathrm{V}[G] = \mathrm{E}[G^2] - \mathrm{E}[G]^2 \tag{4}$$

In order to calculate the Getis and Ord's G statistic, we first set a distance band to specify the radius of neighborhood. This distance is calculated in such a manner that all polygons have at least one neighbor.

The weight matrix so generated is then converted into binary format which is then used to calculate the global Getis and Ord's G statistic.

```
In [786]: g = G(y, wt)
          g.G , g.p_sim, g.EG

Out[786]: (0.12462429626626316, 0.001, 0.063984639111029404)
```

FIGURE 12: Code snippet showing the Getis and Ord's global G statistic ,its p-value and the expected value of the G-statistic under the assumption of normality.

The output of the code snippet tells us that that the Getis and Ord's G statistic is 0.13 with a p-value of 0.001 and the expected G statistic value of 0.06 under the assumption of normality. To conclude, the global Getis and Ord's G statistic implies that we can reject the null hypothesis that the complaint counts are distributed randomly across the different census tracts. Thus, there exists a spatial correlation between them.

LOCAL GETIS AND ORD'S G AND G* STATISTIC

Getis and Ord's G can be localized in two forms: G_i and G_i^*.

$$G_i(d) = \frac{\sum_j w_{i,j}(d) y_j - W_i \bar{y}(i)}{s(i)\{[(n-1)S_{1i} - W_i^2]/(n-2)\}^{(1/2)}}, j \neq i$$

$$G_i^*(d) = \frac{\sum_j w_{i,j}(d) y_j - W_i^* \bar{y}}{s\{[(nS_{1i}^*) - (W_i^*)^2]/(n-1)\}^{(1/2)}}, j = i$$

where we have $W_i = \sum_{j \neq i} w_{i,j}(d)$, $\bar{y}(i) = \frac{\sum_j y_j}{(n-1)}$, $s^2(i) = \frac{\sum_j y_j^2}{(n-1)} - [\bar{y}(i)]^2$, $W_i^* = W_i + wi, i$, $S_{1i} = \sum_j w_{i,j}^2 (j \neq i)$, and $S_{1i}^* = \sum_j w_{i,j}^2 (\forall j)$, $\bar{y}$ and s^2 denote the usual sample mean and variance of y.

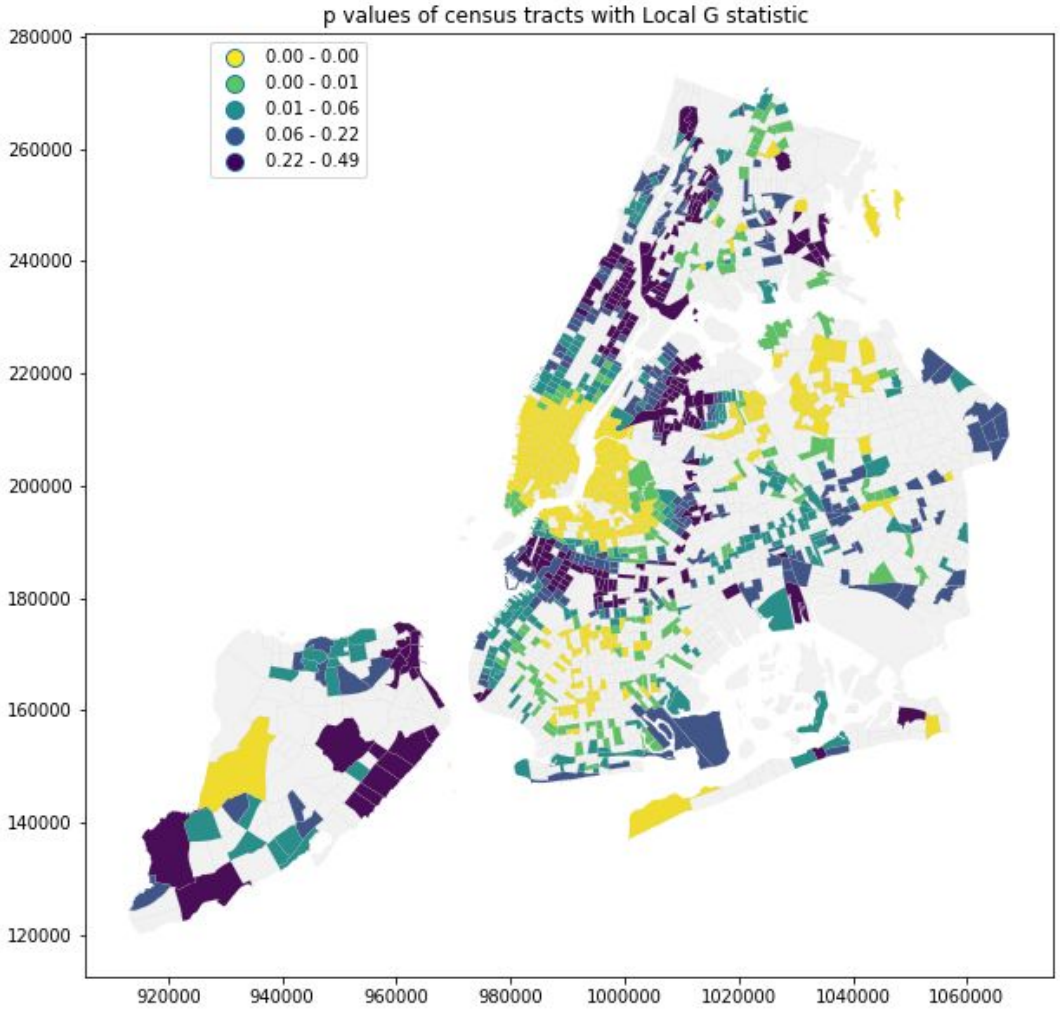


FIGURE 13: The above plot shows the p-values of the census tracts calculated using the local G statistic. The census tracts having significant p-values (0.01 - 0.05) are shown in lighter colors.

This local autocorrelation measure is once again used to identify the location of clusters. Unlike the local Moran's I , it does not detect the spatial outliers. It only detects the hotspots and coldspots at the given level of significance.

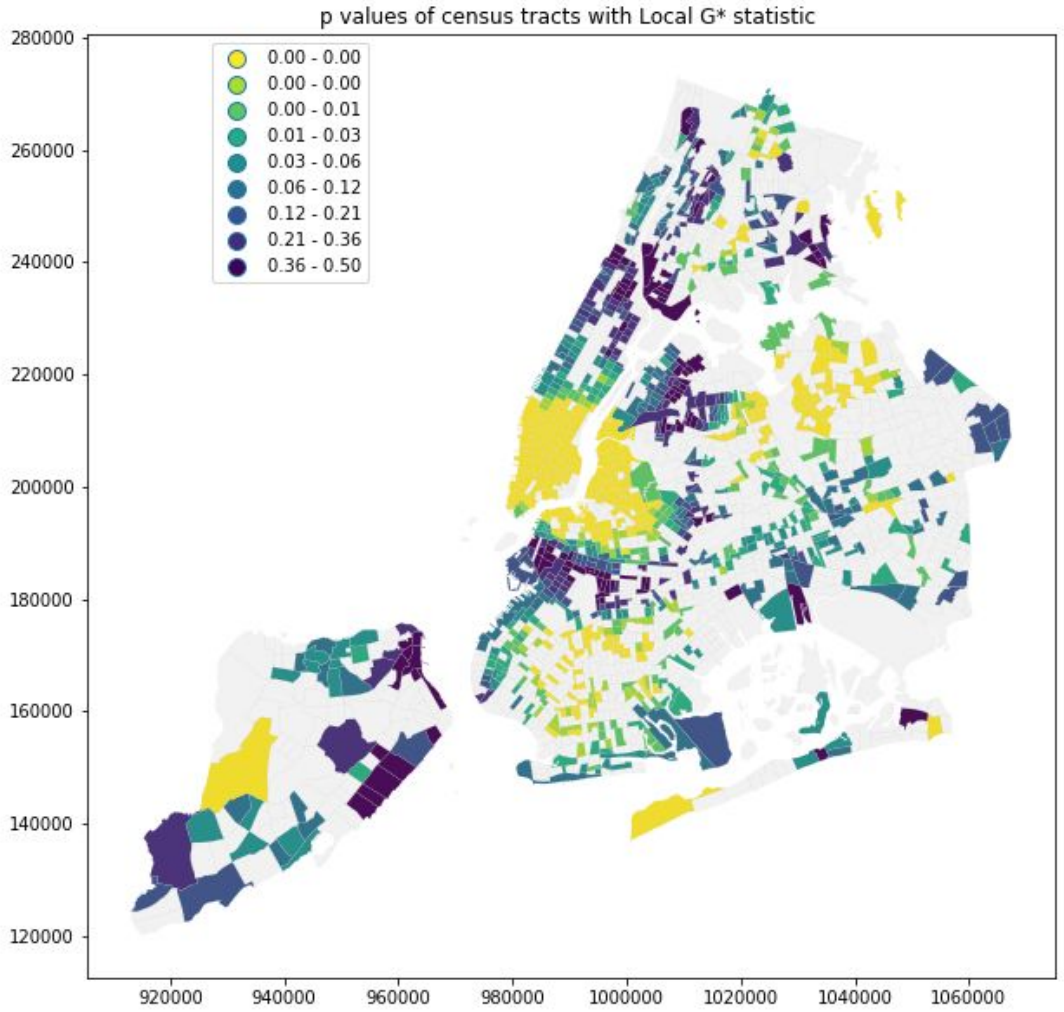


FIGURE 14: The above plot shows the p-values of the census tracts calculated using the local G* statistic. The census tracts having significant p-values (0.01 - 0.05) are shown in lighter colors.

A high z-score and a low p-value for a feature indicates a significant hotspot. A low negative z-score and a small p-value indicates a significant cold spot. The higher (or lower) the z-score, the more intense the clustering. A z-score near 0 means no spatial clustering.

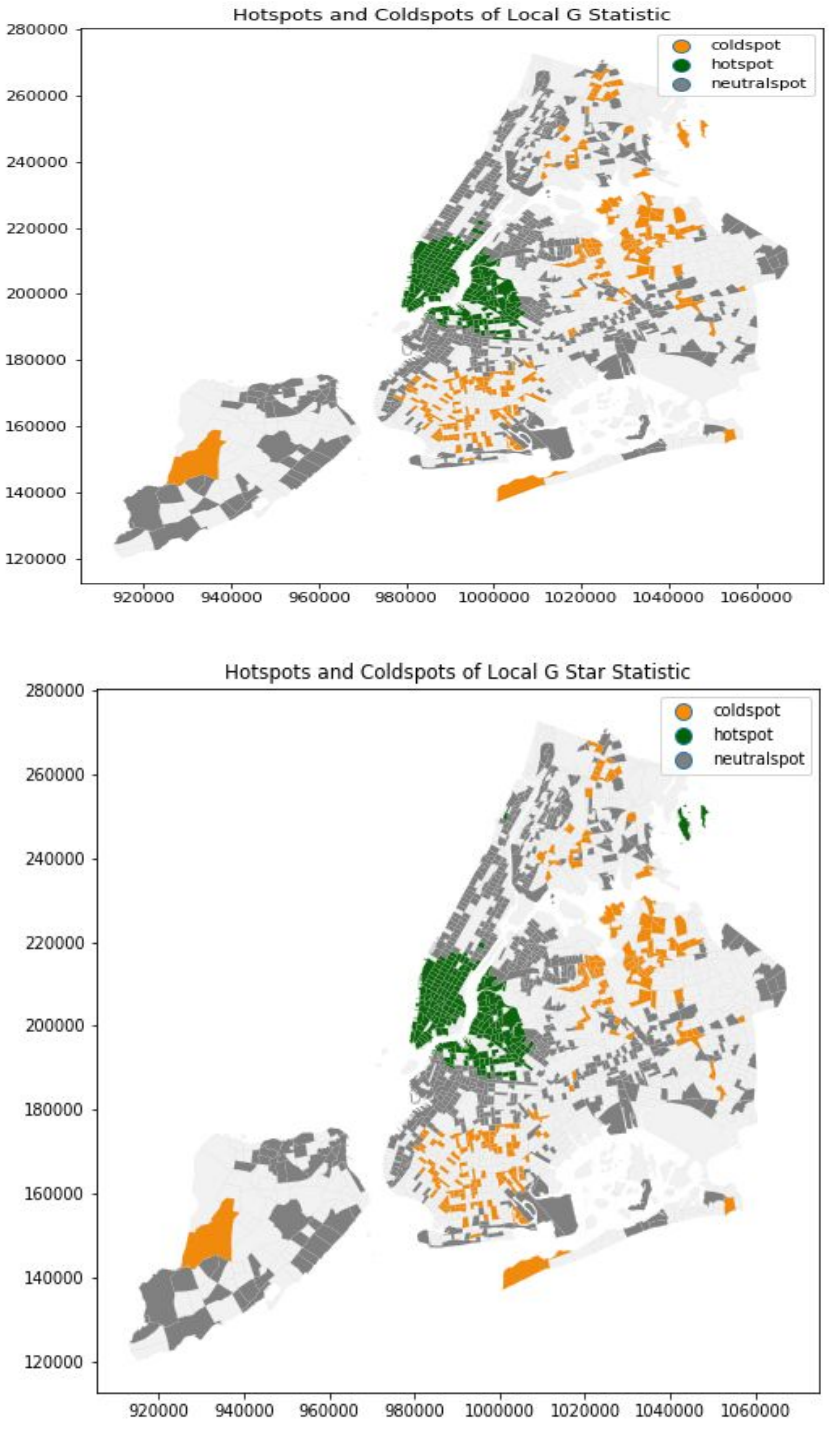


FIGURE 15: The two plots show the hotspots and the cold spots clusters as identified by the Getis and Ord's local G and local G* statistic respectively.

WIth the exception of the census tracts in the northeastern part of New York which has been classified as hotspot by one and as a cold spot by the other, there appear to be no visible differences in the clusters identified.

CONCLUSION

Spatial analysis techniques were successfully applied to identify the hubs of local activity in New York City at night time. While the global Moran's I statistic and the Global Getis and Ord's G gave us a quantitative value to estimate the degree of spatial autocorrelation, the Local Indicators of Spatial Association (LISA) were able to identify the locations of the clusters as shown below.

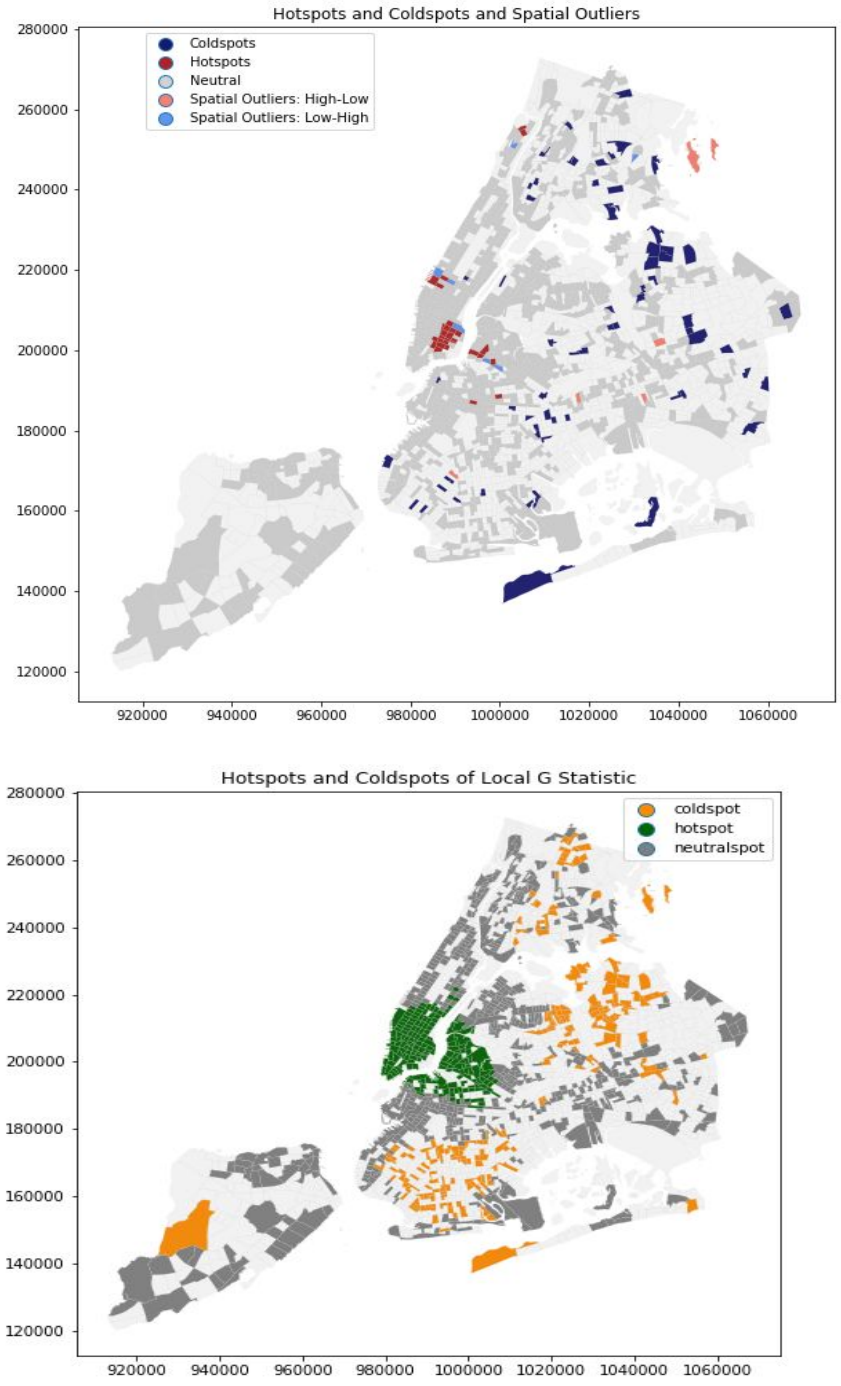


FIGURE 17: Local Moran's I fig(i) and local G statistic fig(ii) successfully isolated the clusters and identified the hubs of activity in census tracts with respect to bars restaurants and pubs.

REFERENCES

1. https://pysal.readthedocs.io/en/v1.11.0/users/tutorials/autocorrelation.html#getis-and-ord-s-g
2. https://pysal.readthedocs.io/en/latest/users/tutorials/autocorrelation.html
3. https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html#getis-ord-statistics
4. http://darribas.org/gds_scipy16/ipynb_md/04_esda.html
5. https://pysal.readthedocs.io/en/latest/library/weights/Distance.html
6. https://github.com/vvt221/python-geospatial-2018/blob/master/GeospatialAnalysis_CitiBike.ipynb
7. https://data.cityofnewyork.us/Social-Services/311-Noise-Complaints/p5f6-bkga