Final Project for Spatial Data Science

Chen Liang
Dec 5th, 2020

I. Data

NYC Neighborhoods: Demographic information for New York City aggregated to neighborhood tabulation area (NTA).
Source: American Community Survey 2008-2012, US Census Bureau.
https://geodacenter.github.io/data-and-lab/NYC-Nhood-ACS-2008-12/

II. Research Questions and Hypotheses

I am looking to see how the length of commute time varies across census tracts in New York city. Commuters often reside in places with cheaper house prices and even areas specifically for dormitory communities (I will refer to these regions as commute belts) and travel daily to work in core city centers (I will refer to these regions as commute cores).

- Research Question #1: Does transportation infrastructure matter to the spatial distribution of commute cores in New York City? If so, how?
- Research Question #2: Is a region's level of wealth inequality related to its residents' average commute time? If so, is there any heterogeneity in the relationship?
- Research Question #3: Why do workers in some regions more likely to commute a long time to work? Is that because the neighboring regions have less jobs? In other words, are the places with more long-distance commuters surrounded by places with higher unemployment rates?

III. Variables and Calculation

- Carto ID primary key of dataset (***cartodb_id),*** used as regional ID.
- Median household income in 2012 Inflation Adjusted Dollars (***medianinco***)
- Gini Index of Income Inequality (***gini***)
- Unemployment rate (***UEMPRATE)***
- More than 90 min commute to work for Workers 16 Years and Over Who Did Not Work at Home (***comm90plus),*** and similarly: comm_less5, comm_5_14, comm_15_29, comm_30_44, comm_45_59, comm_60_89.
- Total number of workers included (***totalcomm***) = comm_less5 + comm_5_14 + comm_15_29 + comm_30_44 + comm_45_59 + comm_60_89 + comm90plus
- Percentage of workers who need more than 45 mins to commute (***comm45plus***) = (comm_45_59 + comm_60_89 + comm90plus) / totalcomm
- Similarity, ***comm30less*** = (comm_less5 + comm_5_14 + comm_15_29) / totalcomm

IV. Commute Cores & Transportation Infrastructure

Figure 1 is a box map which shows the spatial distribution of the percentages of workers who need more than 45 mins to commute. We can observe that first, there seem to be several commute clusters, but almost all commute cores locate close to the East River, suggesting the importance of water transportation. The only two exceptions are first, Staten Island, which has its own commute cores close to the relatively more populated East and North Shores, and second, Floyd Bennett Field. While it appears to be a commute core in the far south of the city along the shore of Jamaica Bay, it is more like an outlier. Floyd Bennett Field is an airfield and part of the Gateway National Recreation Area. That is, it is an unpopulated public green space, and its few residents tend to work specifically for the airfield and don't have to commute a lot.

But the impact of transport infrastructure is not limited to that of water (and maybe air) transportation. On the southern side of the city, commuters with longer travel time tend to cluster around major highways. That is, the access to highways seems to enable or maybe even encourage residents in the surrounding regions to look for faraway jobs. Or it is also possible that many people who work in expensive regions choose to avoid higher living expenses and rents by living close to highways and commuting for a longer time. On the opposite, residents in places with similar socioeconomic characteristics but no direct access to highways may just tend to work locally and instead have shorter commute time.

Secondly, with a reference to **Figure 2**, a box map for median household income, we can observe that commute cores tend to be richer regions. For example, residents in Manhattan enjoy very short commute time and residents in the neighboring regions commute for a longer time to work in Manhattan. In fact, the southeastern sides of the city look almost like a giant commute belt surrounding the prosper East River. As we can see from **Figure 3**, residents in poorer regions (regions with lower median household income) do tend to have higher commute time. But this relationship is not fully linear: for example, with similar median income, regions with higher unemployment rate will have more residents who need to commute further for jobs.

V. Spatial Heterogeneity and Wealth Inequality

Commute time is often considered a proxy for the accessibility of transport infrastructure. Median household income and Gini coefficient offer two measurements of regional wealth distribution. Median income tells us approximately how rich a place is in general. It is close to average income, but is less sensitive to the existence of outliers, such as a few billionaires whose income can significantly increase the whole region's average income. Gini coefficient, on the other hand, tells us how much of a place's wealth is dominated by a small group of people. A Gini of 0.49 (approximately the 75th percentile in the **Figure 4's** Box Map) means that 1% of all the region's population owns 50% of all wealth.

The given variables allow us to estimate whether a region's distribution of wealth is associated with the development of its transport infrastructure. Theoretically, it is hard to predict whether the relationship is negative or positive. On the one hand, we can hypothesize that in places where wealth is distributed more evenly, local governments would try to offer more public transportation to average workers. On the other hand, places with very high level of wealth inequality--such as lower Manhattan--often have a longer history of economic prosperity and thus may have more developed public transportation.

The general observation from **Figure 4** is that regions with higher Gini coefficients tend to be the places where people commute less. But is that true statistically, and how does this pattern vary across different regions?

Here, I use ***comm30less*** instead of ***comm45plus*** because commuting for over 45 minutes almost certainly means that the worker commutes to census tracts faraway. The percentage of workers who commute for less than 30 minutes, instead, better indicates how they move within their own census tracts or only to nearby areas. **Figure 6** shows four scatter plots for Gini coefficients and commute time and I used the quantile map in **Figure 5** for brushing.

As shown in **Figure 6**, in general, places with a higher level of wealth inequality tend to be the places where residents commute less, and this relationship is statistically significant. But the R-squared value is only 0.046, suggesting that our linear model cannot explain much of this relationship. Our brushing, however, yields an interesting observation. In places where median income is low, Gini coefficients have nothing to do with commute time. But as median income increases, the positive relationship between Gini coefficients and commute time becomes more and more significant and the R-squared value also increases to 0.258.

There are many ways to interpret the observations, but without other supporting evidences, it is hard to draw a concrete conclusion. My tentative explanations are, first, in poor regions with a high level of wealth inequality, a small group of commuters--the relatively rich, white-collar population--travel a long distance every day to work. The other residents, however, work locally and earn only minimal income, but enjoy very short commute time. In poor regions with low Gini, while the long-distance rich commuters are less, the limited number of commuters may suggest worse transport infrastructure. That is, local workers may also have to travel a longer time to work. This heterogeneity means that the overall commute time is relatively fixed.

Second, for New York city, the top 1/3 of the regions with the highest median income tend to be the regions which have been historically rich and developed. A higher Gini coefficient here indicates more about the number of millionaires and billionaires residing in the place than about the difference in wealth between local dishwashers and a few white-collar workers in a poor commute belt. That is, a rich region with a higher Gini is likely to be a region closer to financial hubs where the public transportation (e.g., subways) is easily accessible and the residents who can afford living here won't bother to work somewhere else.

VI. Univariate Spatial Autocorrelation and Weights

I used Moran's I and three different spatial weights to explore the spatial autocorrelation of commute time (***Comm45plus).*** The three weights are 1) first-order Rook contiguity, 2) K-Nearest neighbors (kNN) with k=5, and 3) distance weight with a Euclidean bandwidth of 14000. The connectivity graphs are shown in **Figure 7** and the weight characteristics are shown in **Table 1**.

First, the characteristics of spatial weights tell us that our areal units (census tracts) are very small and distribute very densely in the map. A census tract even has 27 rook neighbors which it shares a fragment of boundaries with. This suggests that for future research, we might need to test how the patterns of commute time are sensitive to our scale of areal unit, especially given how easy it is to commute outside of a census tract. KNN weights distribute with a similar pattern. The matrix better constrains the number of neighbors for a few outlier areal units, such as the Central Park, but ignores many close neighbors simply because their centroids are far away.

The spatial weights based on distance, on the other hand, reflect about 10-20 mins of drive at lower Manhattan. With the uncertainty of water transportation, morning traffic, and different road conditions, it is hard to estimate an exact length of commute. Nevertheless, this weight matrix provides a context for us to understand how distance may affect the clustering pattern of commute belts and cores. This weight also reminds us how small our areal unit is: on average, one can travel from one census tract to about 180 other census tracts within 20 minutes (an estimation based on Euclidean bandwidth of 14000). This means a small local cluster is very easily to be formed, but this cluster may just indicate that a homogeneous area is divided into too many little pieces. Thus, we need to focus on larger clusters instead.

As shown in **Figure 8a and 8b**, the clustering patterns of commute time identified by rook and kNN weights are very similar. But it is interesting how the kNN method omits the high-high cluster on the Staten Island detected by the rook method. This is probably because the census tracts spread so loosely there that the two high-high cores on the Staten Island become one of each other's five nearest neighbors and the difference is cancelled out. The kNN method also shows less high-high regions on the east side, where census tracts distribute so densely that the selection of five neighbors is almost arbitrary. The kNN method, however, detects a high-high cluster on the very northeastern side of the city, which the rook method does not detect. As shown in **Figure 7's** connectivity graphs, this is because unlike the rook method, two close census tracts separated by water can still be neighbors with the kNN method.

The distance weight matrix, on the other hand, shows that while there might be a few other smaller urban areas, there is actually only one major commute core in the city: Manhattan. This pattern corresponds to our original observation that the southeastern side of the city is basically a giant commute belt surrounding the East River. This pattern is very obvious even with 999 permutations and 0.001 significance threshold. With a much higher z-value compared to other weight matrices, the distance weight matrix shows that the spatial autocorrelation effect exists in many orders of neighbors. In fact, given that one can easily travel from one census tract to another within 10 minutes, the clustering pattern of the percentage of workers who commute more than 45 minutes might be even stronger when we consider higher orders of neighbors.

VII. Bivariate Spatial Autocorrelation

Lastly, I examine the hypothesis that places with more long-distance commuters tend to be surrounded by places with higher unemployment rates. I use both rook and distance weight matrices to test this hypothesis.

As shown in **Figure 9a**, there is a statistically significant (z-value = 22.6142 with 999 permutations) and positive (Moran's I = 0.2407) spatial correlation between comm45plus and unemployment rates. That is, regions with higher percentages of workers who need to commute more than 45 minutes are very likely to be surrounded by regions with higher unemployment rates. The low-low clusters include the general Manhattan area, the Little Neck and Glen Oaks neighborhoods on the very east end of the city, the south Brooklyn, and the central-east side of the Staten Island. The low-low clusters include the Bronx county north to Manhattan, the east Brooklyn (with neighborhoods such as Brownsville and East New York), and the Jamaica neighborhoods in the borough of Queens.

The cluster map based on the distance matrix displays similar clustering patterns. Again, as shown in **Figure 9b**, there is a statistically significant (z-value = 69.5714 with 999 permutations) and positive (Moran's I = 0.1498) spatial autocorrelation between comm45plus and unemployment rates. The slope looks flatter, mainly because with this relatively large distance threshold, even the richest places have some poor neighbors and vise versa. But the correlation itself is more salient with a higher z-value. It suggests that if we expand the rook contiguity to more orders of neighbors, we may detect even more salient spatial autocorrelations.

But the fact that spatial patterns become more obvious with more orders of neighbors triggers a few warnings to our data interpretation. First, as mentioned before, our areal units might be too small for us to fully examine questions about commute time. A census tract only covers a few blocks and often cannot be differentiated from its neighbors with similar socio-economic conditions. Also, even the smallest level of local governments spans over multiple census tracts. Thus, an individual census tract does not tend to have enough agency to significantly influence its neighbors by carrying out social policies, building infrastructures, or encouraging specific business or industry.

Second, the seemingly obvious spatial autocorrelations might be less about spillover effects (causation) but more about certain similar characters that coexist in the region (correlation). For example, there are a few census tracts in the southeastern Brooklyn borough, which have both high percentage of workers commuting over 45 minutes and high lagged unemployment rate. But this might be because they are surrounded by a couple of neighborhoods which held the highest poverty and crime rates in New York City. Therefore, it is hard to tell whether residents in these places have to commute further because their surrounding census tracts do not provide enough job opportunities, or because this whole region has a miserable financial condition that makes it hard to boost local job markets, improve road conditions, or provide public transportation.

But either way, because our subject of interest, New York City, is not a big region (only five boroughs), census tracks still present a nice detailed picture of the local heterogeneity in the city and allow us to better depict the boundaries of each spatial cluster.

VIII. Conclusion

As a conclusion, it seems that both water transportation and highways matter to the spatial distribution of commute cores in New York City. More convenient water transportation helps accumulate wealth along the river shores and residents who can afford living in these places won’t bother to commute to work somewhere else. More convenient highways create commute belts in poorer neighborhoods where residents avoid intimidating rents and living expenses by commuting longer to work in richer neighborhoods.

Second, in richer neighborhoods, the positive relationship between wealth equality and commute time is more significant. That is, we need to consider spatial heterogeneity when interpreting this linear model, because a high Gini coefficient in a poor or a rich neighborhood can mean very different social realities.

Finally, it seems that regions with more workers who commute over 45 minutes tend to be surrounded by regions with less employment rates. But it is still too early to conclude based on these spatial clusters that these workers commute longer *because* they cannot find closer jobs. Instead, both longer commute time and high unemployment rates may be the consequence of the miserable economic situation of the whole region.

Table 1: Characteristics of different spatial weights

	Min	Max	Mean	Median	%Non-Zero
Rook	0	27	5.21	5	0.24%
kNN	5	5	5.00	5	0.23%
Distance	5	304	178.22	186	6.10%

Table 2: Global Moran’s I Statistics (999 permutations)

	Pseudo P	**z-value**	**Moran's I**
Rook	0.001	47.5787	0.667
kNN	0.001	50.5971	0.668
Distance	0.001	172.7597	0.486

Figure 1. Box Map for comm45plus, with and without basemap (World Street Map)

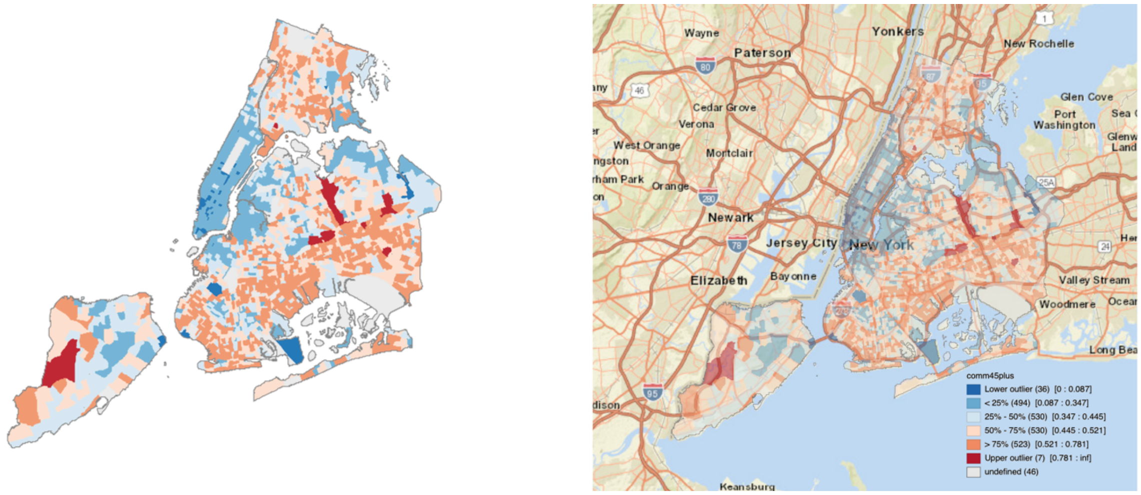


Figure 2: Box Map for Median Income

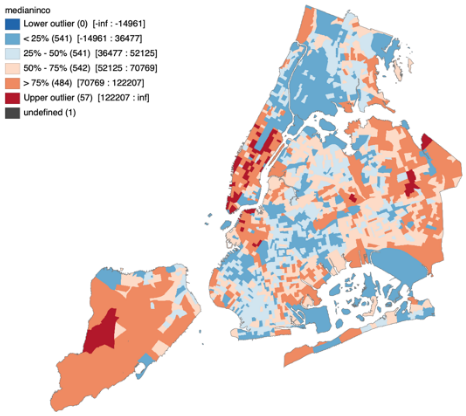


Figure 3. Bubble Chart: median income vs commute time, size & color = unemployment rate

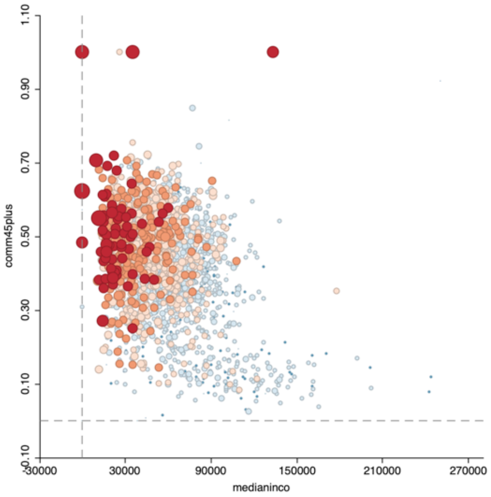


Figure 4. Box Map for Gini

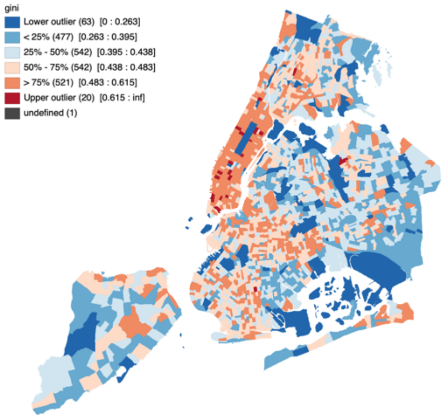
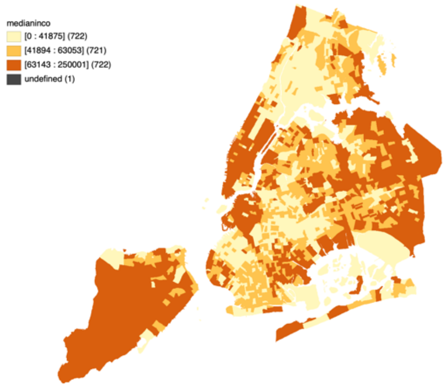
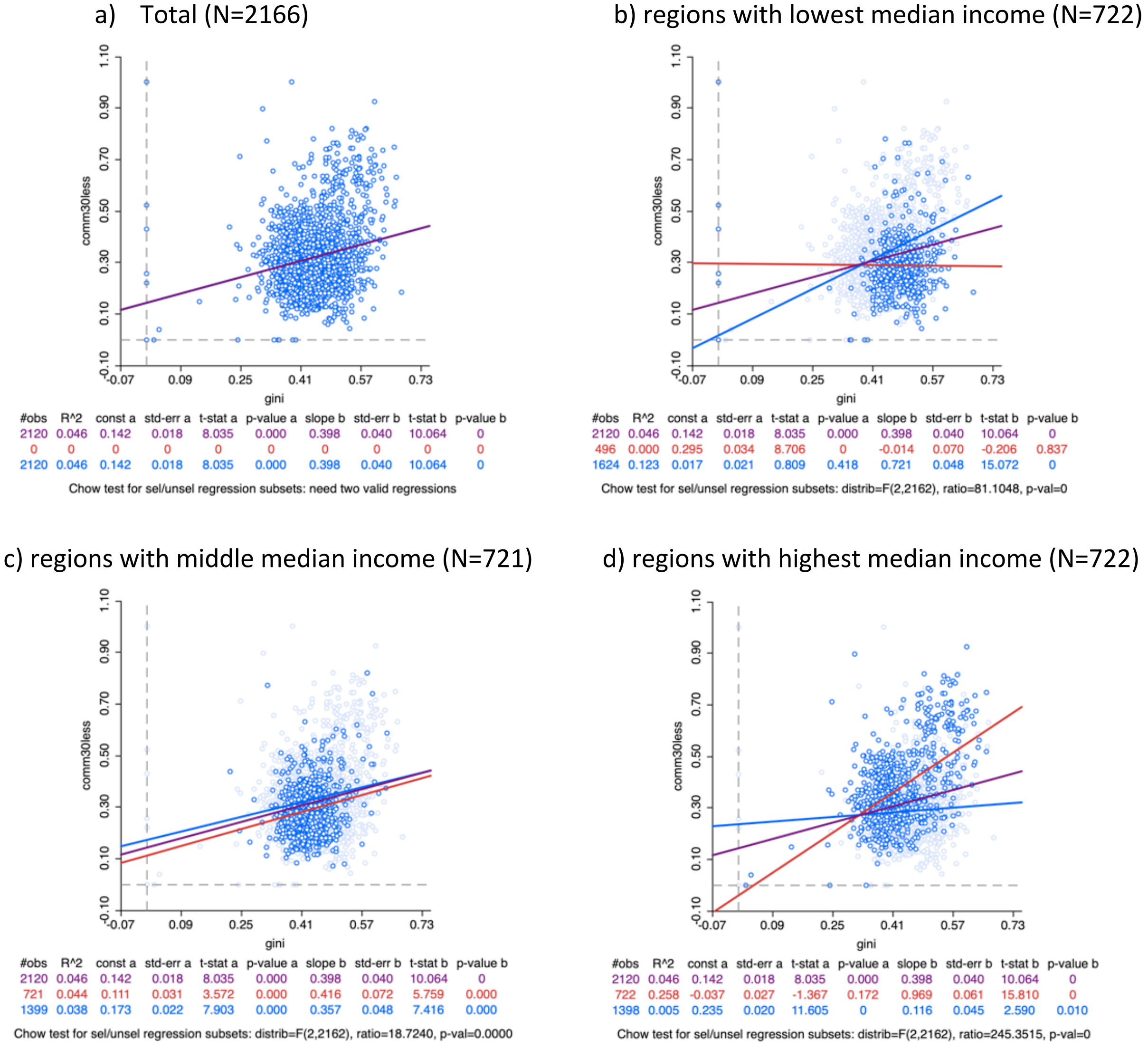


Figure 5. Quantile Map for Median Income



`

Figure 6. Scatter Plots for Gini vs Comm30less



a) Total (N=2166)

```
#obs   R^2    const a  std-err a  t-stat a  p-value a  slope b  std-err b  t-stat b  p-value b
2120   0.046  0.142    0.018      8.035     0.000      0.398    0.040      10.064    0
   0   0      0        0          0         0          0        0          0         0
2120   0.046  0.142    0.018      8.035     0.000      0.398    0.040      10.064    0

Chow test for sel/unsel regression subsets: need two valid regressions
```

b) regions with lowest median income (N=722)

```
#obs   R^2    const a  std-err a  t-stat a  p-value a  slope b  std-err b  t-stat b  p-value b
2120   0.046  0.142    0.018      8.035     0.000      0.398    0.040      10.064    0
 496   0.000  0.295    0.034      8.706     0          -0.014   0.070      -0.206    0.837
1624   0.123  0.017    0.021      0.809     0.418      0.721    0.048      15.072    0

Chow test for sel/unsel regression subsets: distrib=F(2,2162), ratio=81.1048, p-val=0
```

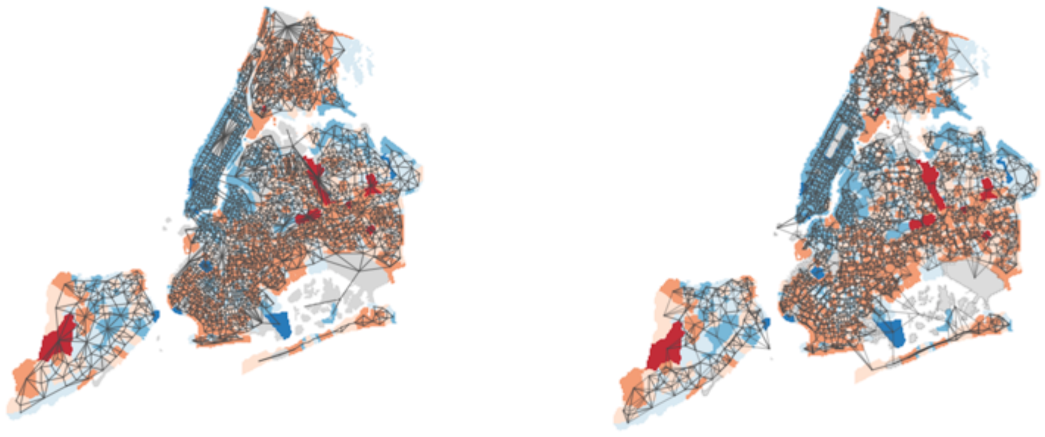
c) regions with middle median income (N=721)

```
#obs   R^2    const a  std-err a  t-stat a  p-value a  slope b  std-err b  t-stat b  p-value b
2120   0.046  0.142    0.018      8.035     0.000      0.398    0.040      10.064    0
 721   0.044  0.111    0.031      3.572     0.000      0.416    0.072      5.759     0.000
1399   0.038  0.173    0.022      7.903     0.000      0.357    0.048      7.416     0.000

Chow test for sel/unsel regression subsets: distrib=F(2,2162), ratio=18.7240, p-val=0.0000
```

d) regions with highest median income (N=722)

```
#obs   R^2    const a  std-err a  t-stat a  p-value a  slope b  std-err b  t-stat b  p-value b
2120   0.046  0.142    0.018      8.035     0.000      0.398    0.040      10.064    0
 722   0.258  -0.037   0.027      -1.367    0.172      0.969    0.061      15.810    0
1398   0.005  0.235    0.020      11.605    0          0.116    0.045      2.590     0.010

Chow test for sel/unsel regression subsets: distrib=F(2,2162), ratio=245.3515, p-val=0
```

Figure 7. Connectivity Graphs for different weights (theme is a Box Map for Comm45plus)



a) Rook (order = 1)

b) KNN with 5 Neighbors

`

c) Distance with 14000 Euclidean Distance Threshold

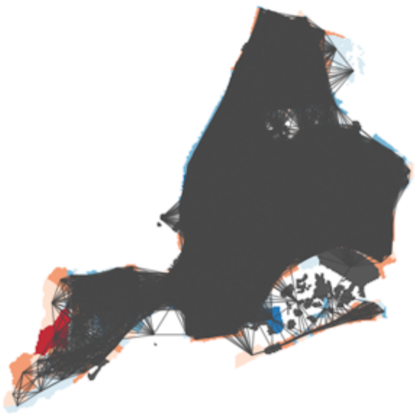


Figure 8a. LISA Cluster Map for Comm45plus, weighted by Rook (0.05 significance)

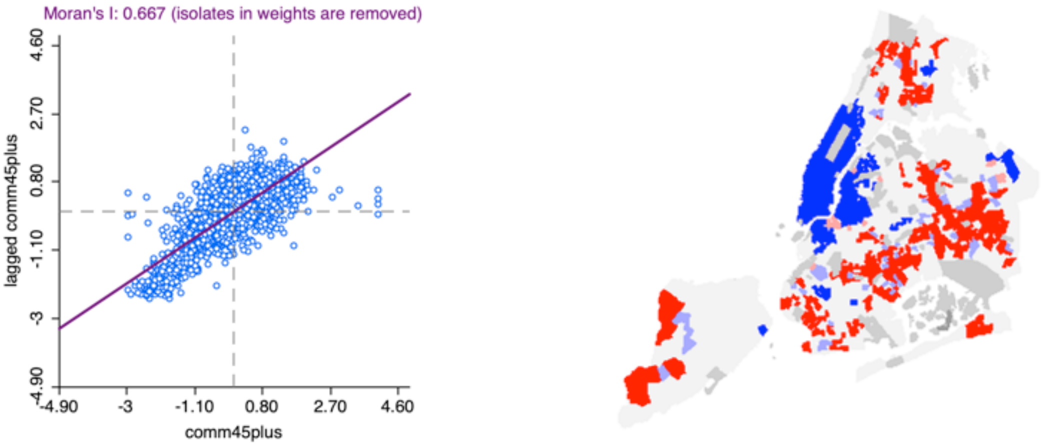
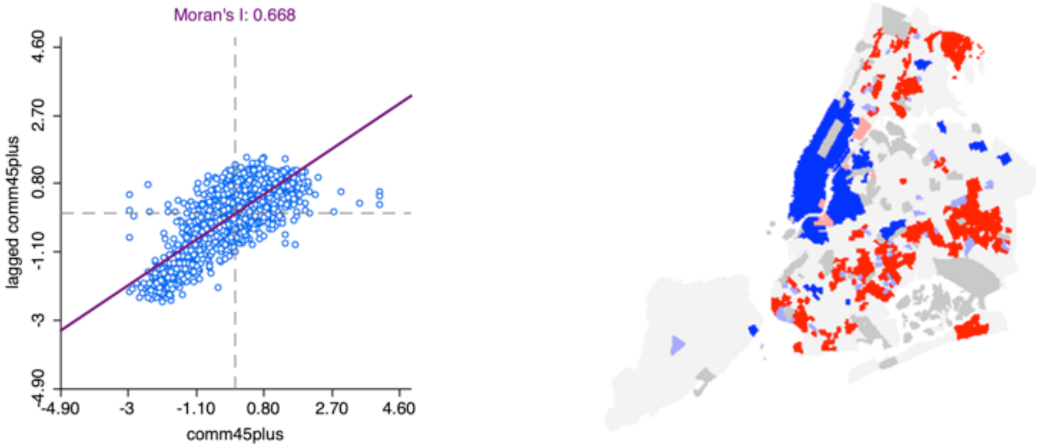
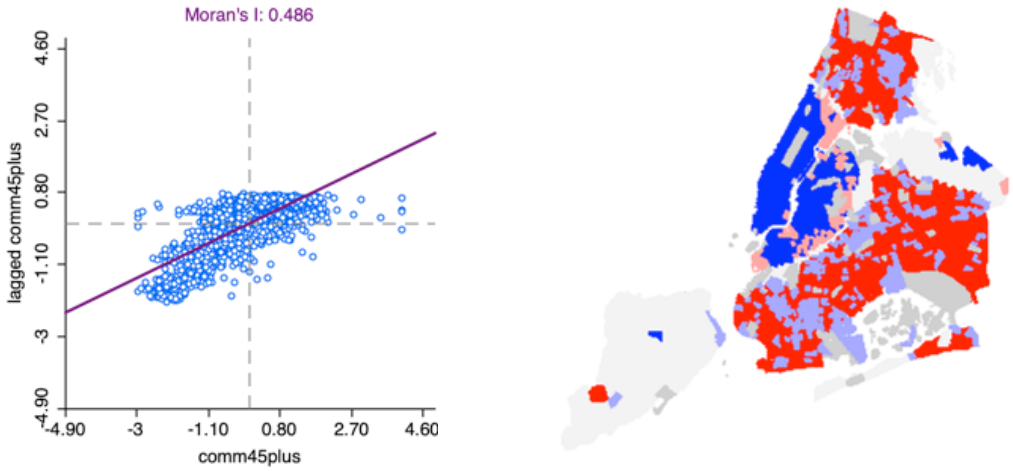


Figure 8b. LISA Cluster Map for Comm45plus, weighted by kNN (0.05 significance)



`

Figure 8c. LISA Cluster Map for Comm45plus, weighted by distance (0.05 significance)



0.01 Significance & 0.001 Significance

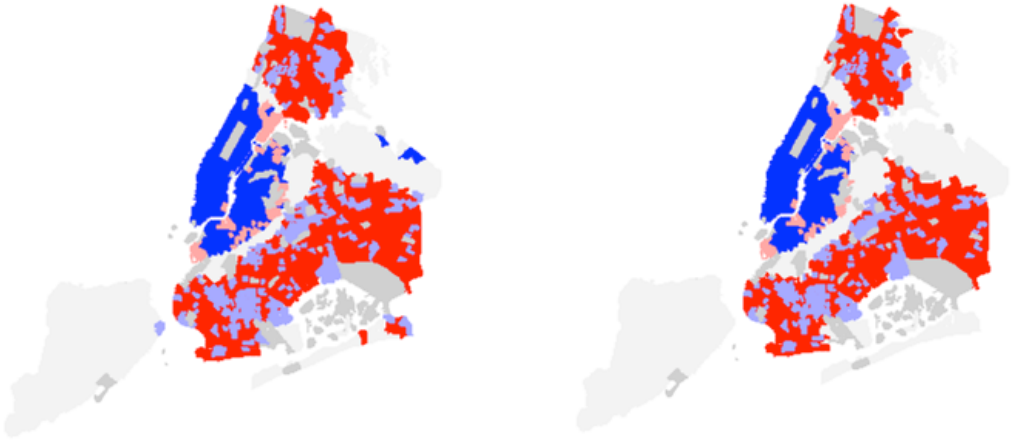
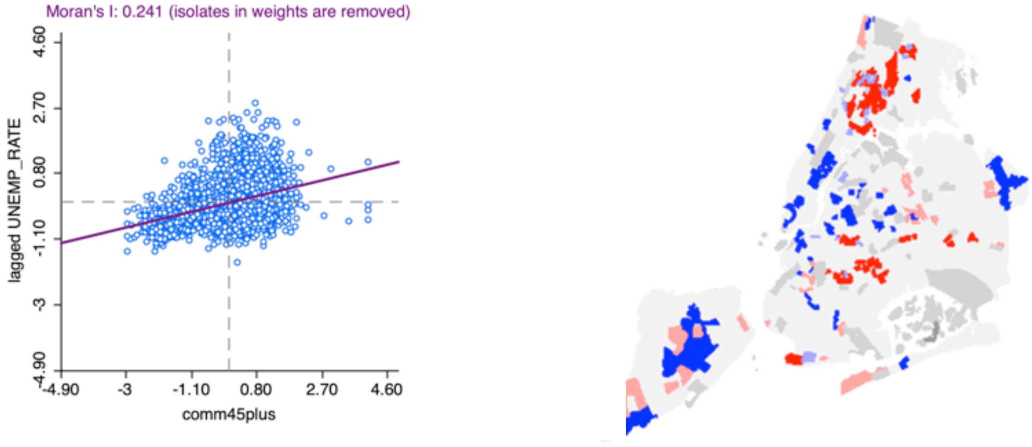
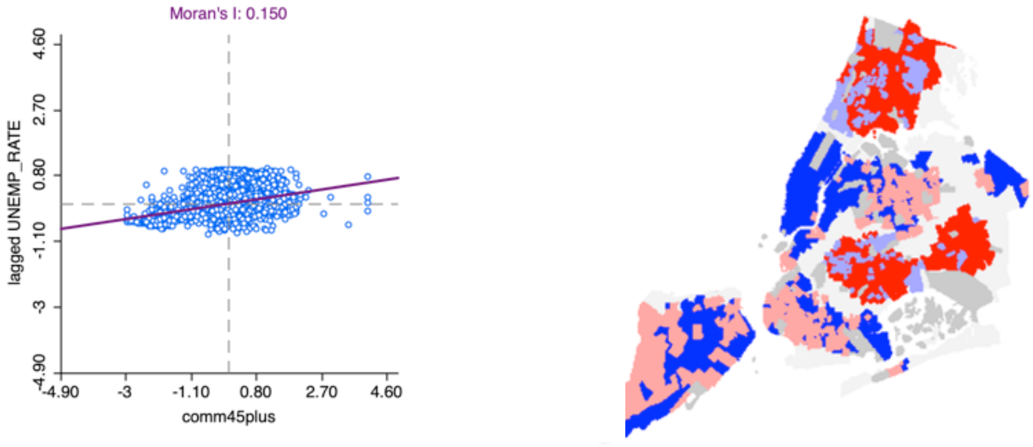


Figure 9a. comm45plus vs lagged unemp_rate, Rook, (0.01 significance, 999 permutations)



`

Figure 9b. comm45plus vs lagged unemp_rate, Distance, (0.01 significance, 999 permutations)



`