Assignment 4 - Putting it all together

Yongfei Lu

1. Introduction

Since 2019 Airbnb has launched the Superhost program, which certificates those mostly dedicated to providing outstanding hospitality service as superhosts. Curious about the great success Airbnb has gained in the City of Chicago, the researcher wants to explore the spatial association among 4 important variables, sphost_ds (superhost density in each community), price (nightly room rent), black_pct (black population proportion in each community) and occup_rate (airbnb room occupancy rate in each community).

2. Initial Analysis

From the point layer (Figure 1) of superhost in the city of Chicago, we can see that most of them reside in the northeast of the city, while there are much less in the southwest. The distribution pattern is further verified by the superhost density box map (Figure 2), which filters out the effect of uneven population distribution across the city (communities with larger population can have more superhosts). The averages chart for sphost_ds with map brushing (Figure 3) also consolidates our observation, showing that the respective superhost density are 2.84 for the 19 selected communities in the northeast, and 0.41 for the remaining unselected 60 ones.

As for the distribution of nightly room price, the box map for price (Figure 4) presents a similar pattern: the 7 upper outlier communities clump around almost exactly where communities with high superhost density reside, while communities with low room price overlap the regions where communities have low superhost density. Nevertheless, the distribution of black population ratio turns out to be an inverse pattern: communities with low black population ratio tend to appear in the northeast and middle west, while those with high ratio appear to be segregated in the south. Regarding the distribution of the occupancy rate (Figure 5), it doesn't seem to have a consistent pattern with the other 4 four variables. Generally, communities in relatively north tend to have higher occupancy rate than those in the south. However, looking in more details, high occupancy rate also appears in some southern communities and low occupancy rate, in some north communities.

To sum up, the interesting distribution patterns of the first 3 variables seem to suggest a potential spatial pattern: communities with high superhost density tend to have high room price and low black population ratio. In the next sections, we would conduct more intensive exploratory data analysis and spatial autocorrelation exploration to find more interesting patterns in detail.

3. Exploratory Data Analysis

First, to develop a big picture of the general pattern for the 4 variables, we put the 4 box maps together and select the observations with extremely high superhost density

(Figure 7), we can see that communities with high superhost density and room price but low black population ratio gather around the northeast of the city, while occupancy rate seem not to have a clear association with the superhost density distribution. The observation is in line with our initial analysis above. Then we plot 3 conditional maps (Figure 8 - 10) to further explore the relationship among the 4 variables. Figure 8 indicates that occupancy rate is not clearly associated with superhost density and black_pct. Even though many low occupancy rate communities reside in the south with low black_pct and superhost density, the high and low occupancy rate communities still distribute across the city with any obvious pattern. Figure 9 and 10 imply that price is positively associated with superhost density, while black population ratio and occupancy rate don't have clear-cut associations with room price. However, Figure 10 still indicates a weak potential pattern that in communities where black_pct is high, extremely high room price is less likely to appear in communities even if their superhsot density is high.

In the bubble for sphost_ds and price (Figure 11) with dark orange as large black_pct and light blue as small black pct, we can see that among communities with small black population proportion, room price increases with superhost density and that communities with large black_pct have a upper bound of room price and superhost density. This means that high black_pct can exert a negative influence on room price and superhost density, which matches our inference from Figure 10. This can be conversely verified by the parallel coordinate plot (Figure 12), communities with high superhost density and price only corresponds to low black population proportion.

At last, we explore the bivariate relationship in more depth. From the scatter matrix (Figure 13), two linear fits are conspicuous: (1) sphost_ds and black_pct; (2) sphost_ds and price. Figure 14 presents more details about (1). Generally, black_pct is negatively associated with sphost_ds, with a very significant negative slope of -8.818. By selecting the lower black_pct bandth (less than 30 percent), we see that there is a significantly positive association between sphost_ds and black_pct, meaning that when black population ratio is low, communities with higher superhost density tend to have slightly higher black population ratio and that when black_pct is relatively high, communities with higher superhost density tend to have lower black population ratio. The p-value of the Chow test statistic is 0, meaning the difference between the selected and unselected part is significant. The observation consolidates our previous inference in Figure 10 to 13. In Figure 15, we can see that in general, one unit increase in superhost density can lead to 31.500 dollars of increase in the average room price in that community, and that the positive effect is significant with a p-value of 0.000. Besides, we also find that when superhost density in the community exceeds about 3, a significantly different effect becomes dominant: one unit increase in superhost density leads to about 20.97 dollars of decrease in the community's room price. However, the negative effect is not significant with a p-value of 0.561. The result indicates that the superhost accreditation gives the hosts in the communities with higher superhost density more market power to set higher room price, through which they can probably earn more (the hypothesis

still needs to be checked). Such a simple exploration confirms the signaling theory first proposed by Michael Spence1. The superhost status probably serves as a positive signal of the service quality in the market and successfully boost the Airbnb business in the City of Chicago.

From the intensive data exploration above, we cannot identify a clear association of occupancy rate with other variables in question. Therefore, we'll just choose the 3 variables, sphost_ds, price, and black_pct, to explore the spatial autocorrelation in the next section with the purpose of finding more interesting spatial associations.

4. Spatial Autocorrelation Analysis

In this section, we'll explore the spatial autocorrelation for the community superhost density and possible explanatory variables, including black_pct and price. We'd like to see whether there are any local clusters or spatial outliers.

First of all, we construct 3 spatial weights for each of the 2 most interesting variables, sphost_ds and black_pct. As required, we prepare 2 geographies. One is the original Chicago community polygon layer, and the other, the centroid point layer (Figure 16) derived from the community map. Each of the 3 spatial weights (queen contiguity, distance-band, and k-nearest neighbor weights) will be computed in each of the 2 geographies.

Using alpha_id as the id variable, we respectively compute the queen contiguity weights, great-circle-distance-band weights (using arc distance as the distance metric due to simple latitude-longitude projection) and KNN weights (setting 6 as the number of neighbors).

Figure 17 and 18 give the property information for the 6 weights. As for the commarea_q (queen contiguity weight), it's symmetric weight; the number of neighbors range from 1 to 9, meaning that there is no neighborless community; the mean and median neighbors are 5.12 and 5.00 respectively; and only 6.65% of the cells in the matrix is non-zero, demonstrating the sparsity of the weights. For the commarea_d (distance-band weight), it's also symmetric, but has much higher density of non-zeros in the weight matrix (20.81%). This is because the threshold value of the distance is specified to ensure that each point has at least one neighbor, thus increasing the number of neighbors when the points are highly unevenly distributed (some are very sparse, others are dense). Similarly, the maximum, mean and median neighbors increase to 26, 16.03, 17 respectively. In the case of the commarea_k6 (KNN weight), it's asymmetric and much more stable than the other 2 weights; min, max, mean and median neighbors are all 6; the non-zero proportion (7.79%) is slightly higher than that of the queen contiguity weight but far lower than that of the distance-band weight.

Figure 19 shows the connectivity histogram for the 3 weights. As expected, the connectivity histogram for the distance-band weights has the largest range and its

distribution is the most uneven among the 3 spatial weights. The distribution of queen contiguity weight neighbor is closest to a normal distribution, and the KNN weight connectivity histogram has 77 observations in the 6-to-7 interval.

The connectivity graphs for the 3 types of spatial weights are shown in Figure 20 & 21. The graphs further verify the property of the 3 weights: the one for distance-band weights is the densest, then the KNN weights, followed by the queen contiguity weights.

The analysis for the centroid geography is almost the same as the above, with just some negligible differences in certain statistics.

To explore the local spatial autocorrelation in the maps, we will employ the Univariate and Bivariate Local Moran, Univariate and Multivariate Local Geary as well as Multivariate Quantile LISA analysis.

The cluster map (Figure 22) augments the significant locations with an indication of the type of spatial association, based on the location of the value and its spatial lag in the Moran scatter plot. Here, we analyze the superhost density cluster map derived from the queen contiguity weights since the patterns are consistent and stable across the 6 maps. By the default setting, 999 permutations are performed and the p-value is 0.05. In this example, three out of four categories are represented, with red for the high-high clusters (15 in our example), dark blue for the low-low clusters (19 locations), light blue for the low-high spatial outliers (1 location), and light red for the high-low spatial outliers (0 location). The pattern is that communities with high superhost density tend to have neighboring communities with also high superhost density more than spatial randomness. On the other hand, low superhost density communities also tend to be surrounded by low density communities more so than would be case randomly, but only a very few communities with low superhost density are surrounded by high density neighbors more. Besides, high-high spatial clusters tend to appear in the northeastern of the city, while low-low clusters tend to gather around the southern and southwestern of the city. The low-high cluster is in the Humboldt Park community. By further exploration, we will find some other interesting patterns to explain such phenomenon.

The cluster map for black population proportion per community (Figure 23) also has three out of four categories represented. The high-high clusters appear in northern and middle part of the city, while the low-low clusters appear in the southern of the city. This means that black communities tend to be surrounded by black ones more so than would be the case randomly, vice versa. Such a pattern indicates a conspicuous race segregation in the city: black people mainly gather around the southern part. Not coincidentally, the black_pct high-high clusters take place where superhost_ds low-low clusters appear. This further indicates that black people are less likely to be superhosts. Therefore, if superhost density somehow represents greater chance to earn more via Airbnb rental, black people seem to have less economic gains from the new economy. However, a causal relationship still remains to be confirmed, and such an inference still

needs more rigorous verification.

The Figure 24, the Bivariate Local Moran cluster map for sphost_ds and black_pct, confirms our observation in Figure 22 & 23. There is strong low-high cluster in the southern. However, the interpretation of the bivariate Local Moran cluster map warrants some caution because this could either be attributed to black population proportion in surrounding locations impacting the central location's superhost density or central location's black population ratio affecting its own superhost density. A bit counterintuitively, there are also strong low-low and high-low clusters in the northern, which means that both low and high superhost density communities in the north side can be surrounded by low black density communities more so than would be randomly. The slope of the linear fit from the Moran scatter plot (Figure 25) is -0.213, implying a negative spatial autocorrelation. Also, the significance map (Figure 26) with commarea_q as the spatial weight demonstrates that the 10 locations are securely significant with p = 0.001.

Next, we further explore the patterns by drawing the Local Geary cluster map (Figure 27 & 28 using queen contiguity and KNN weights respectively) for each individual variable. The result confirms a match among the 3 individual variables in both geographic and attribute space. Here we set permutations = 99999 and p = 0.05, and find that strong high-high clusters for sphost_ds and price correspond to the low-low clusters for black_pct in the northeast of the city and that a strong low-low cluster for sphost_ds and price corresponds to a negative cluster for black_pct in the southwest. Interestingly, we also find a common high-high cluster for the 3 variables in the middle west of the city. Besides, for sphost_ds, there is a cluster being classified as other positive than low-low and high-high.

The big picture is summarized in the co-location map in Figure 29, which is constructed by first turning the significant locations in each significance map into an indicator variable for the selection, and then using these three variables in a co-location map. We can easily find the overlap among the significant locations distributed in the 3 Local Geary Univariate cluster map. Here 5 locations are identified as significant for all 3 variables. Moreover, 10 locations are discerned as commonly insignificant and 62 locations are only significant for less than 3 of the variables.

Then the Multivariate Local Geary functionality is employed. In our example, a p-value of 0.001 also corresponds roughly to the FDR. From the significance and cluster maps (Figure 30), we illustrate the effect of tightening the p-value from 0.05 to 0.001. The number of significant location is reduced from 59 to 24. The significant locations cover all the 4 positive clusters identified in the previous co-location map.

However, the results of the Multivariate Local Geary significance or cluster map should be treated with caution because the multivariate statistic involves various tradeoffs. The significance map and a 3D data cube (Figure 31) are drawn to illustrate the complexity:

observations in question that are very close in multi-attribute space still have some of their neighbors much further removed. Also, the PCP (Figure 32) can further illustrate the tradeoffs: In this instance, while the distances between the paths are small, there is much less of a close match. Then we select 3 highly significant neighboring locations in the right-hand panel, and we find that the neighbors track each other better and meets our expectation in the multivariate attribute space.

At last, we conduct the Multivariate Quantile LISA by setting sphost_ds, price and black_pct as the 3 variables, as shown in Figure 33. By the default setting (999 permutations and p = 0.05), we see that 5 out of the 77 locations are identified as the cores of actual clusters. Then we set the significance level as the False Discovery Rate (0.0166667) and do 99999 permutations. The result (Figure 34) still remains the same as before. Finally, we construct the co-location map (Figure 35) using the QT_sp, QT_pr and QT_blk derived from the Multivariate Quantile LISA process. However, it turns out that none of the locations in the significance map can be recognized as the spatial clusters of co-locations. This is in line with our economic intuition, because while high superhost density is associated with high room price, the high value of these 2 variables is usually associated with low black population ratio in the communities. Therefore, we plot the co-location maps (Figure 36) for the fifth quantile of sphost_ds and price along with the 2nd and 3rd quantile of black_pct. And it turns out all the 5 significant locations in the significance map have corresponding significant locations in the co-location map, indicating the special spatial association among the 3 variables.

The highlighted 6 communities gather around the northeast of the city. High superhost density, high room price and low or medium black population ratio tend to happen simultaneously in these locations. However, our research cannot identify any causality, which demands other rigorous social science methods. It's also worth mentioning that the quantile LISA method can cause loss of information due to concentration on extremes of the variable distributions.

In conclusion, our analysis indicates a strong spatial association among the superhost density, room price and black population ratio. More specifically, a community with high superhost density is more likely to be surrounded by communities with high average room price and low or medium black population ratio. This implies conspicuous race segregation in the city of Chicago, where black people have inferior economic opportunities in the emerging 'sharing economy'. On the other hand, we cannot see any spatial associations between occupancy rate and the other 3 variables. Besides, regardless of the trivial discrepancy between Multivariate Local Geary and Multivariate Quantile LISA, they yield almost the same pattern, demonstrating the robustness of our analysis. Nevertheless, our analysis cannot give a causality inference.

Appendix A Variables

Variable Name	Data Type	Description	Source
alpha_id	integer	Community id	Created by the author
community	string	Community name	Chicago Data Portal: Boundaries - Community Areas (current)
sphost_ds	real	Superhost density in each community: the number of superhosts per 1000 people in the community	Inside Airbnb: detailed listing data of August 2019
black_pct	real	Black population percentage: the ratio of black popuplation over the total population in the community	CSV table: Chicago Community Area (CCA) CDS data
occup_rate	real	Occupancy rate: mean average occupancy rate of all the Airbnb hosts in the community	Inside Airbnb: detailed listing data of August 2019
price	real	average room price per night in the community	Inside Airbnb: detailed listing data of August 2019

Appendix B Graph

Figure 1: Airbnb Superhost Location Point Map

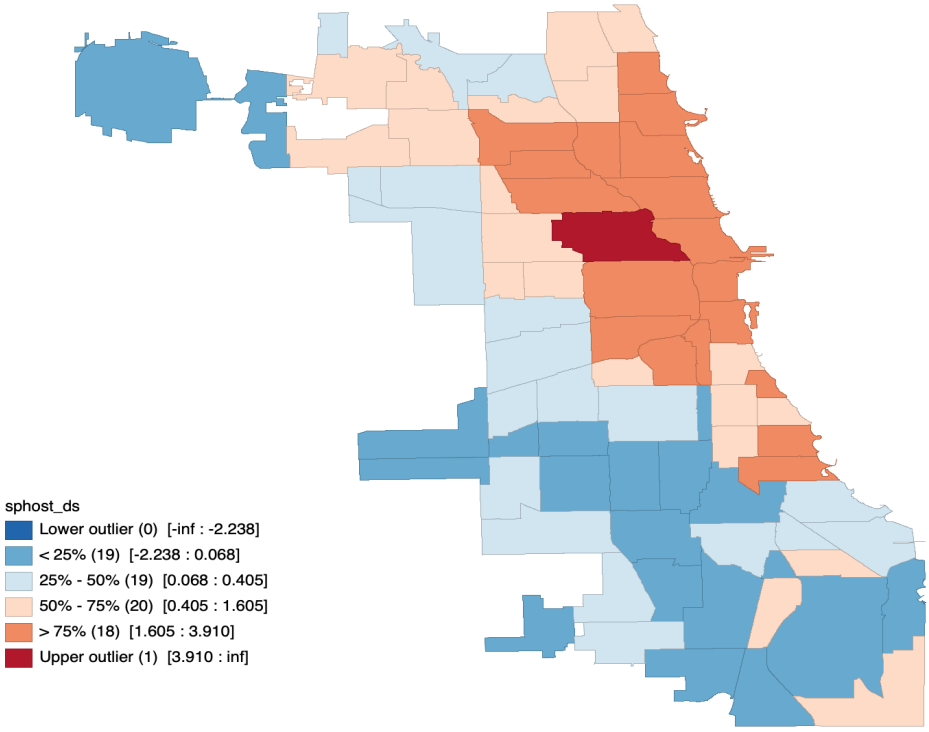

Figure 2: Box Map for Airbnb Community Superhost Density

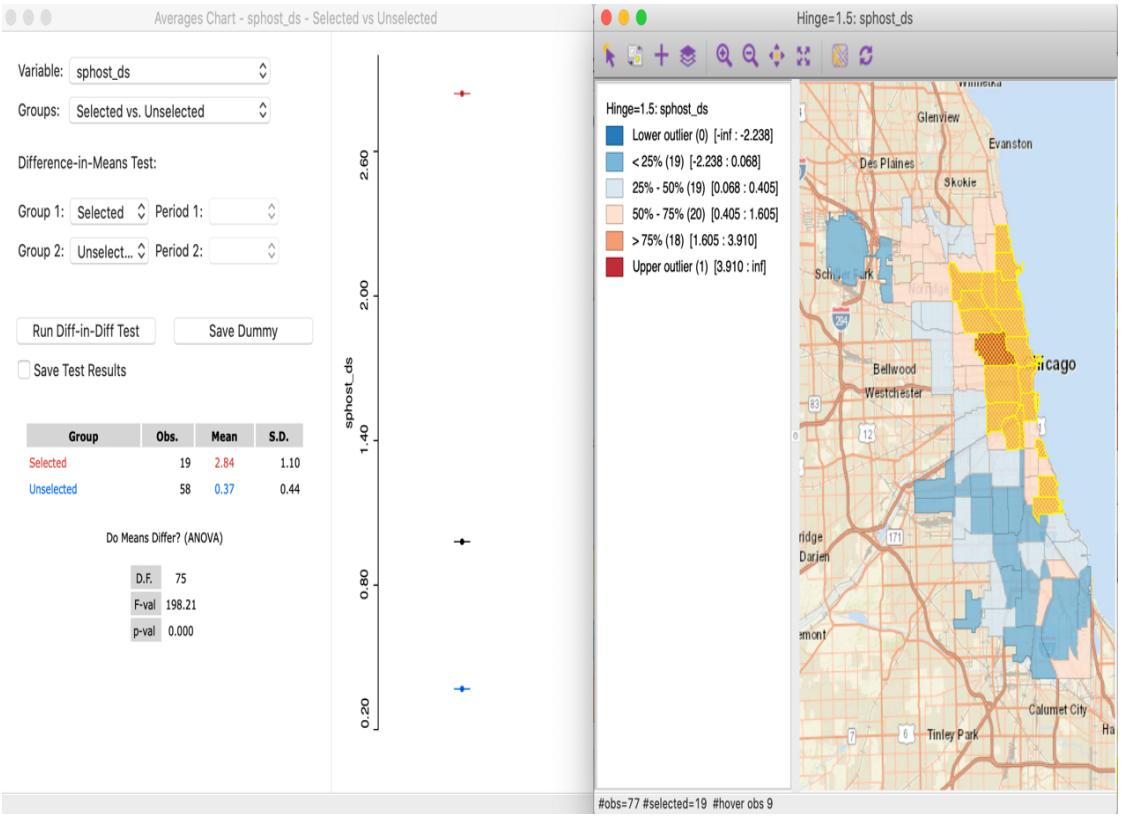

Figure 3: Averages Chart for Airbnb Community Superhost Density

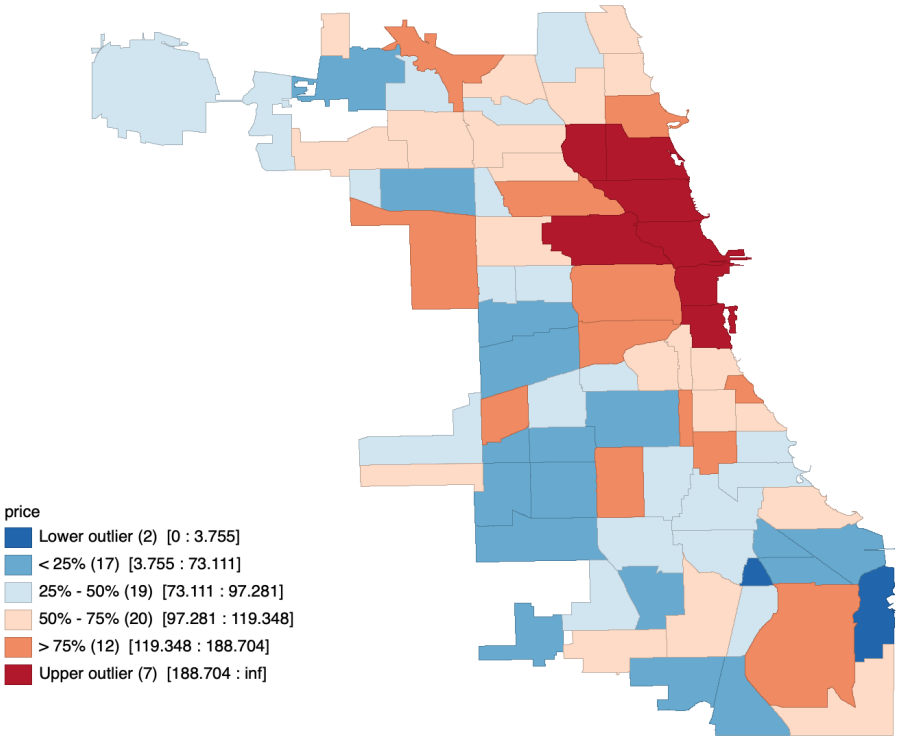

Figure 4: Box Map for Airbnb Community Room Price

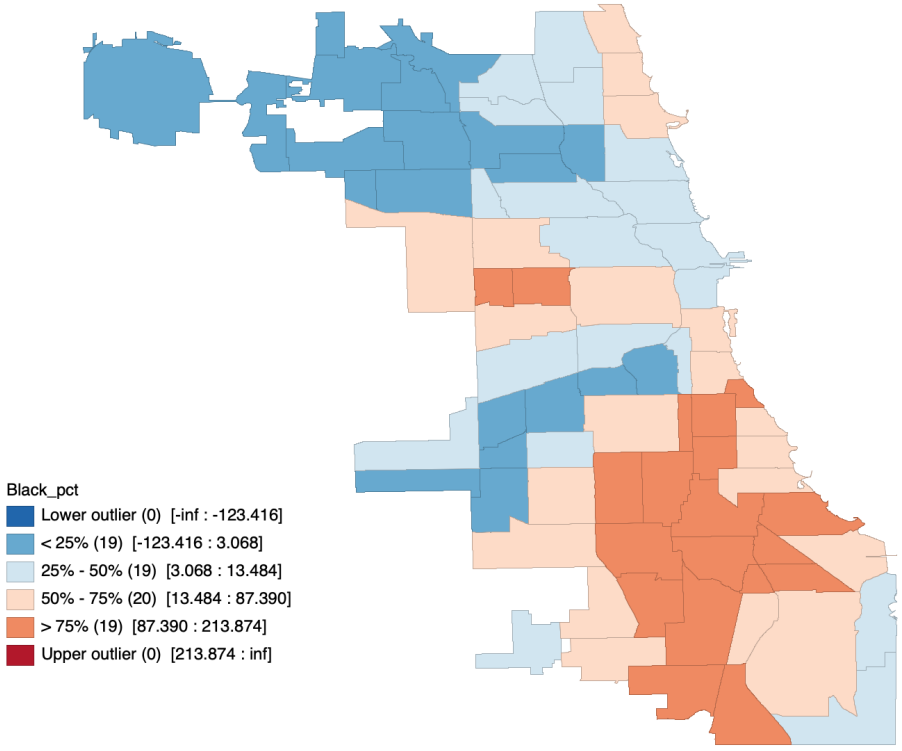

Figure 5: Box Map for Airbnb Community Black Population Ratio

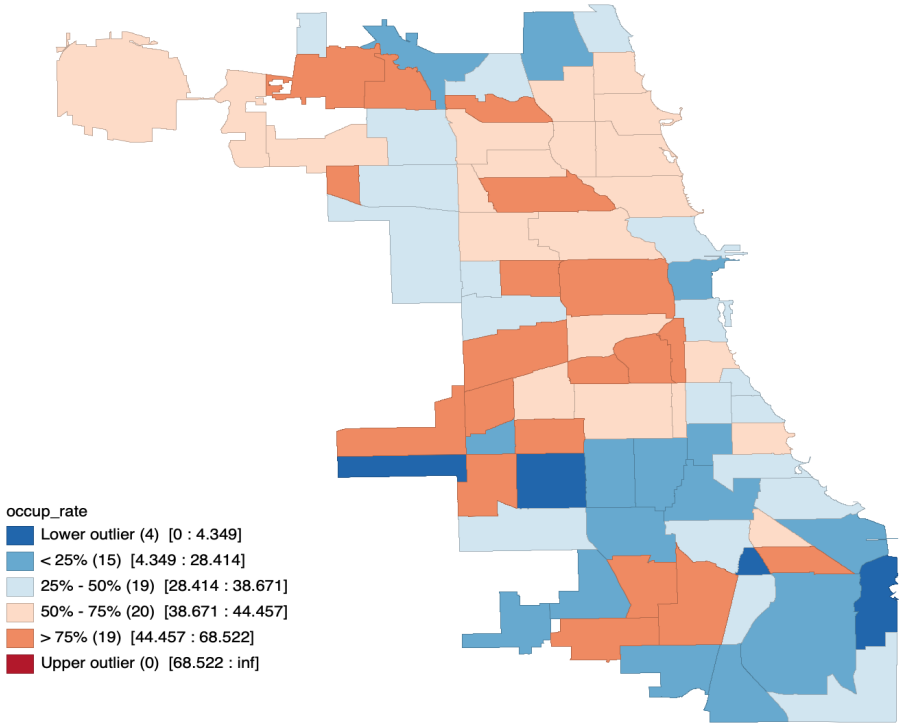

Figure 6: Box Map for Airbnb Community Occupancy Rate

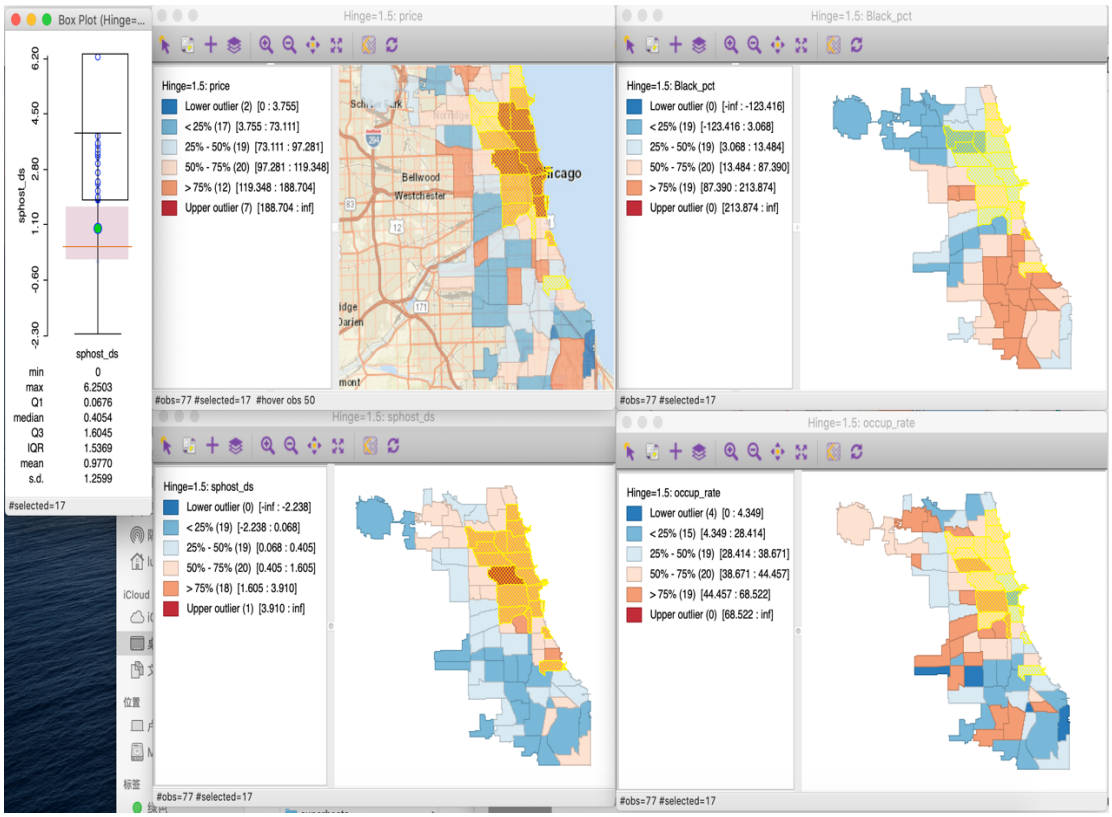

Figure 7: All Box Maps for Airbnb with Box Plot for sphost_ds

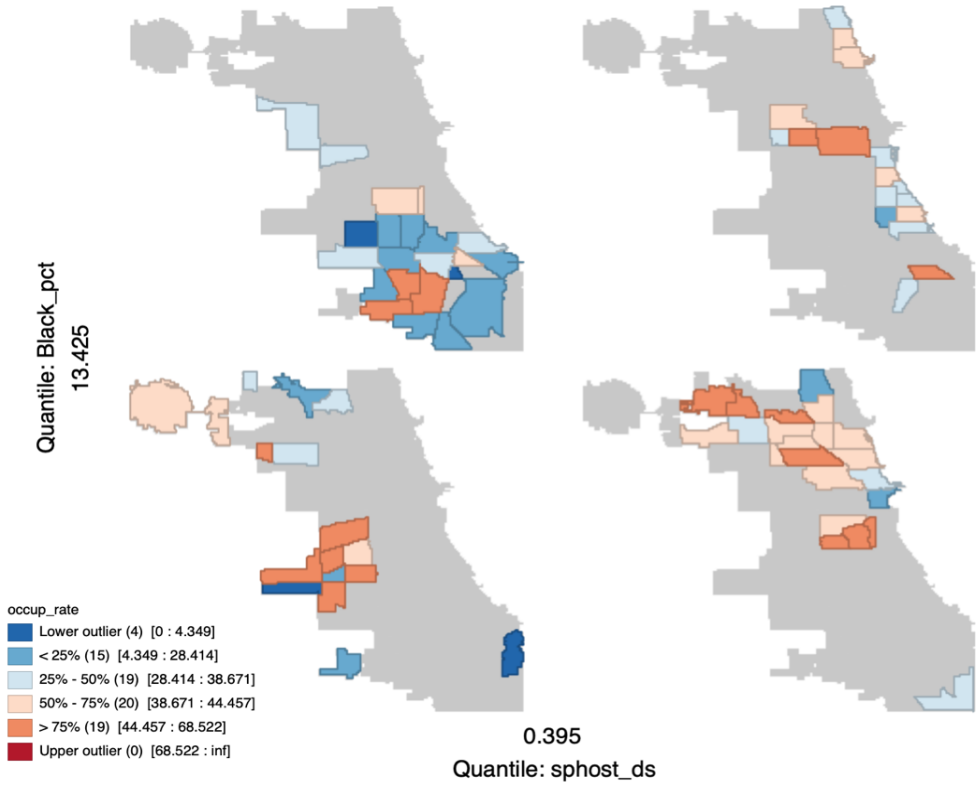

Figure 8: Conditional Map on black_pct And sphost_ds

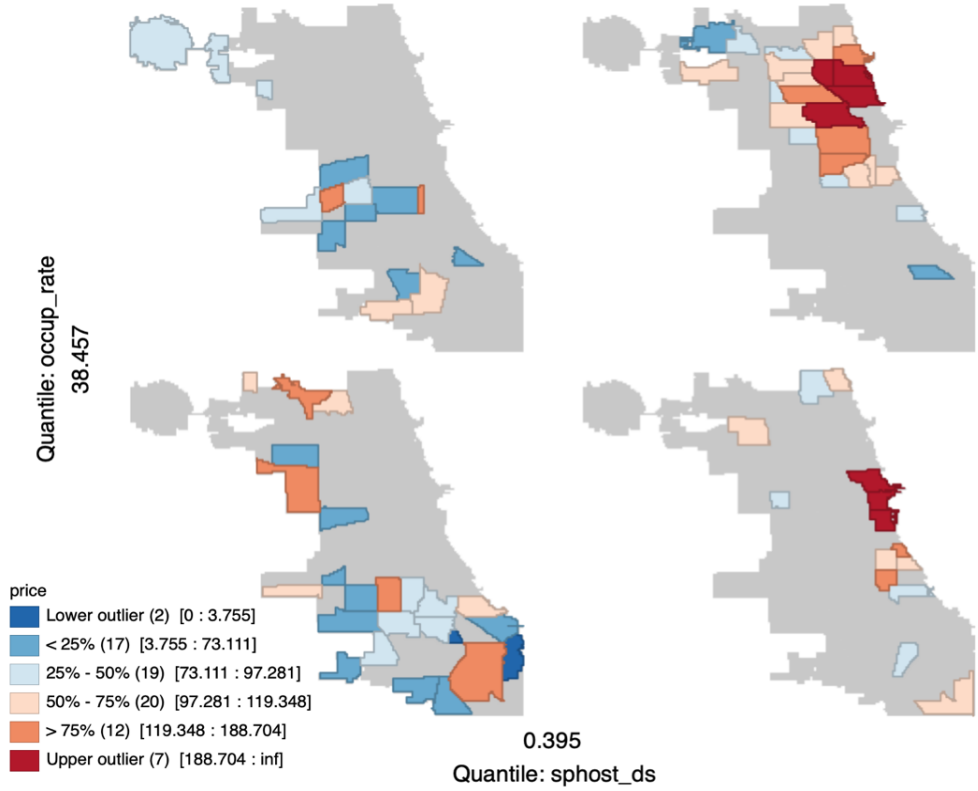

Figure 9: Conditional Map on occup_rate And sphost_ds

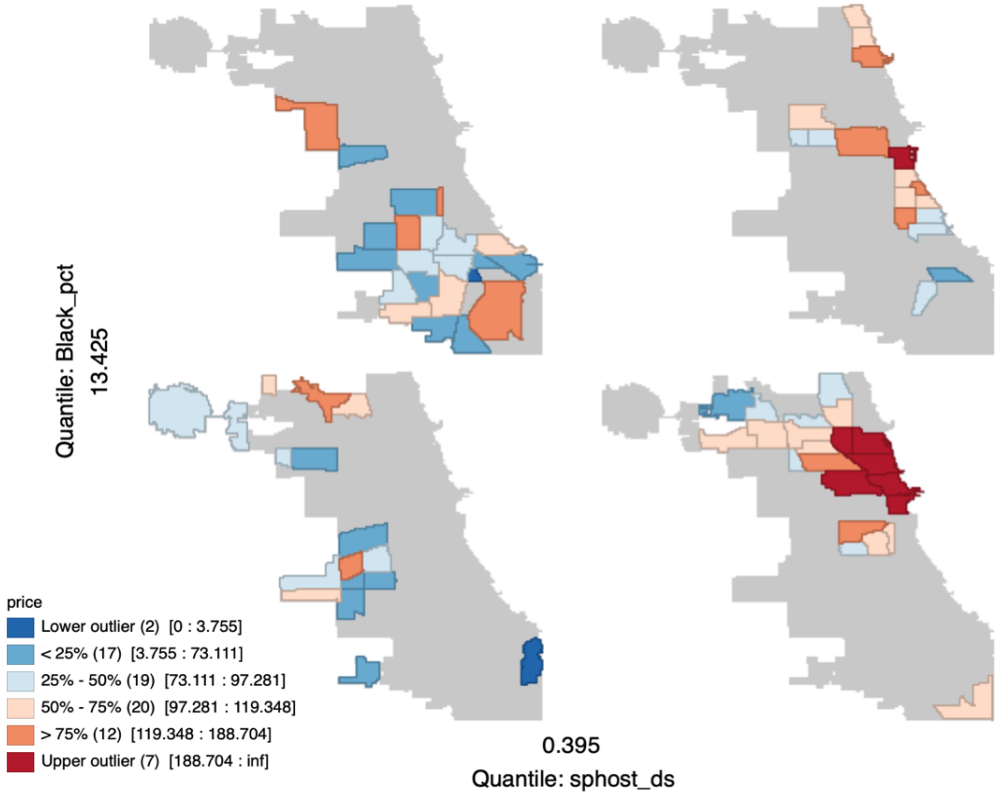

Figure 10: Conditional Map on black_pct And sphost_ds

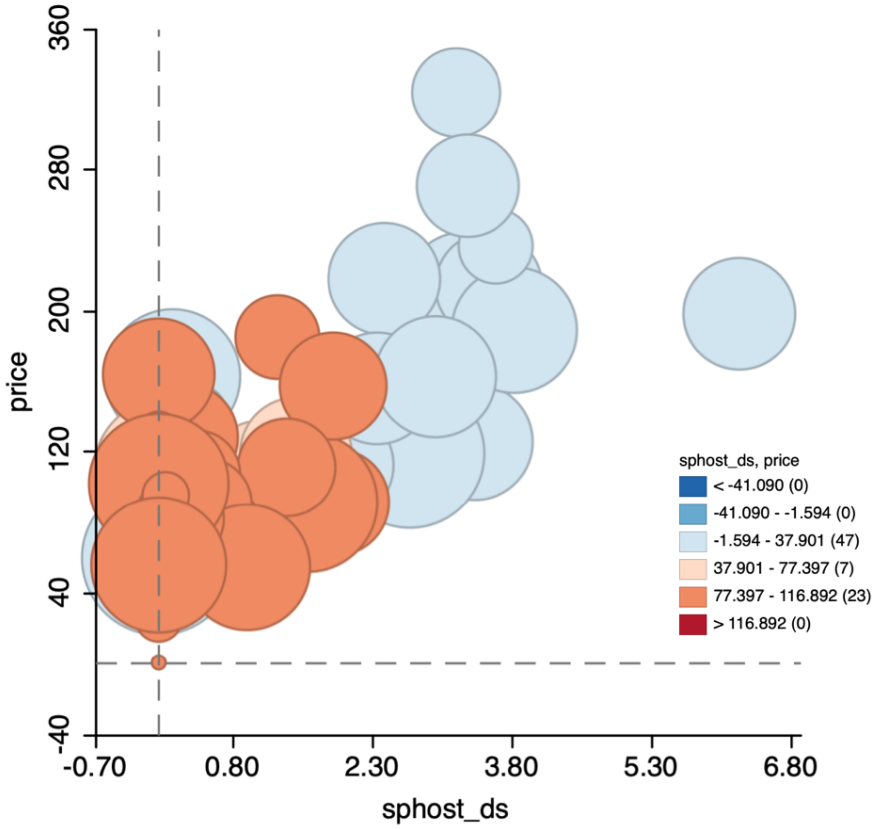

Figure 11: Bubble Chart for price And sphost_ds

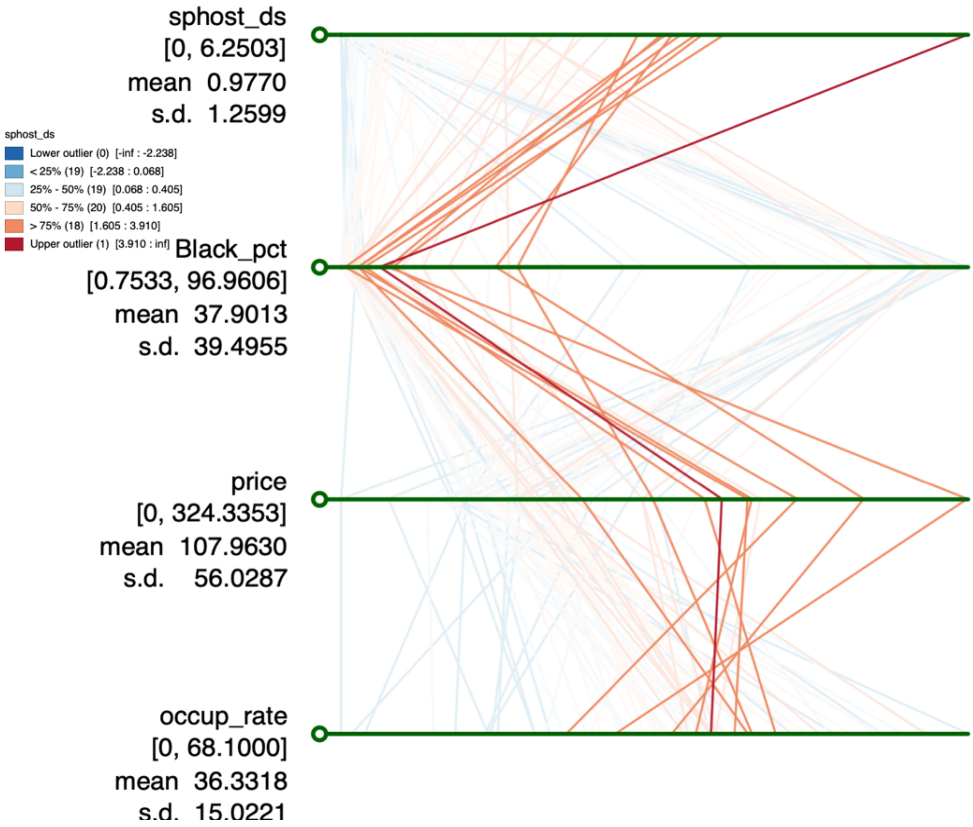

Figure 12: Parallel Coordinate Plot for All 4 Variables

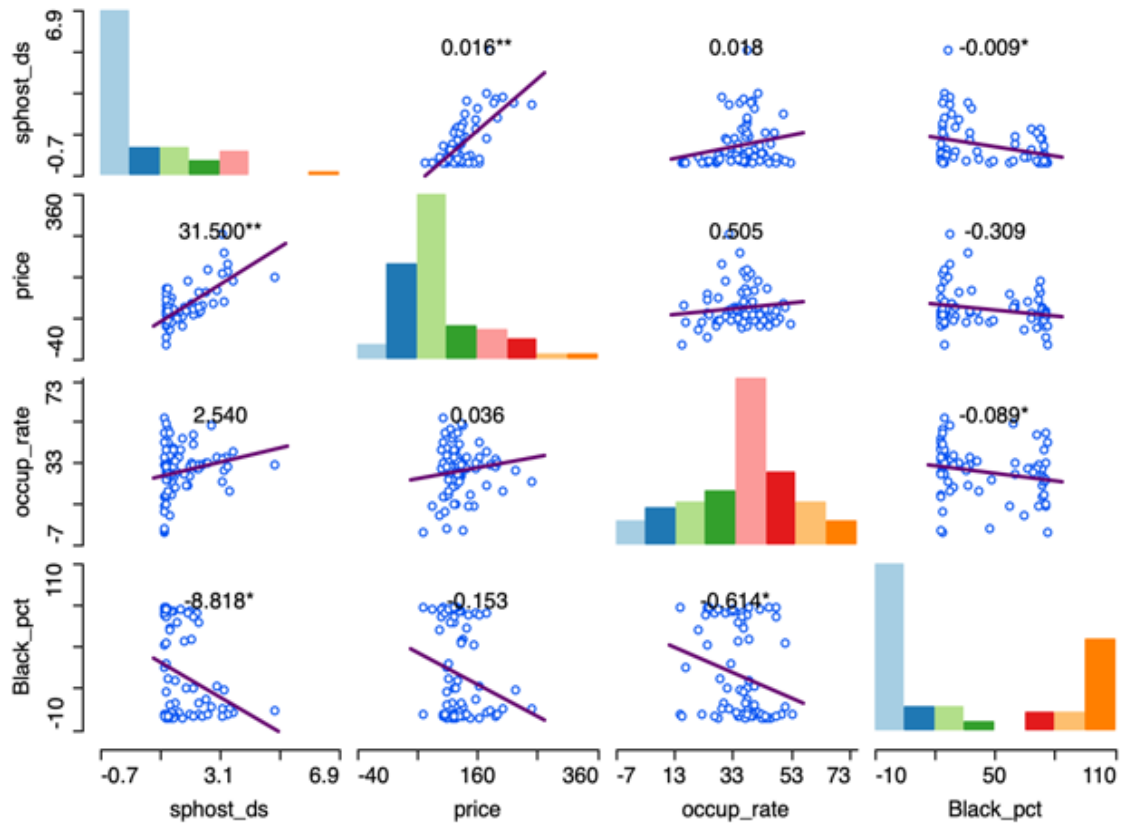

Figure 13: Scatter Plot Matrix

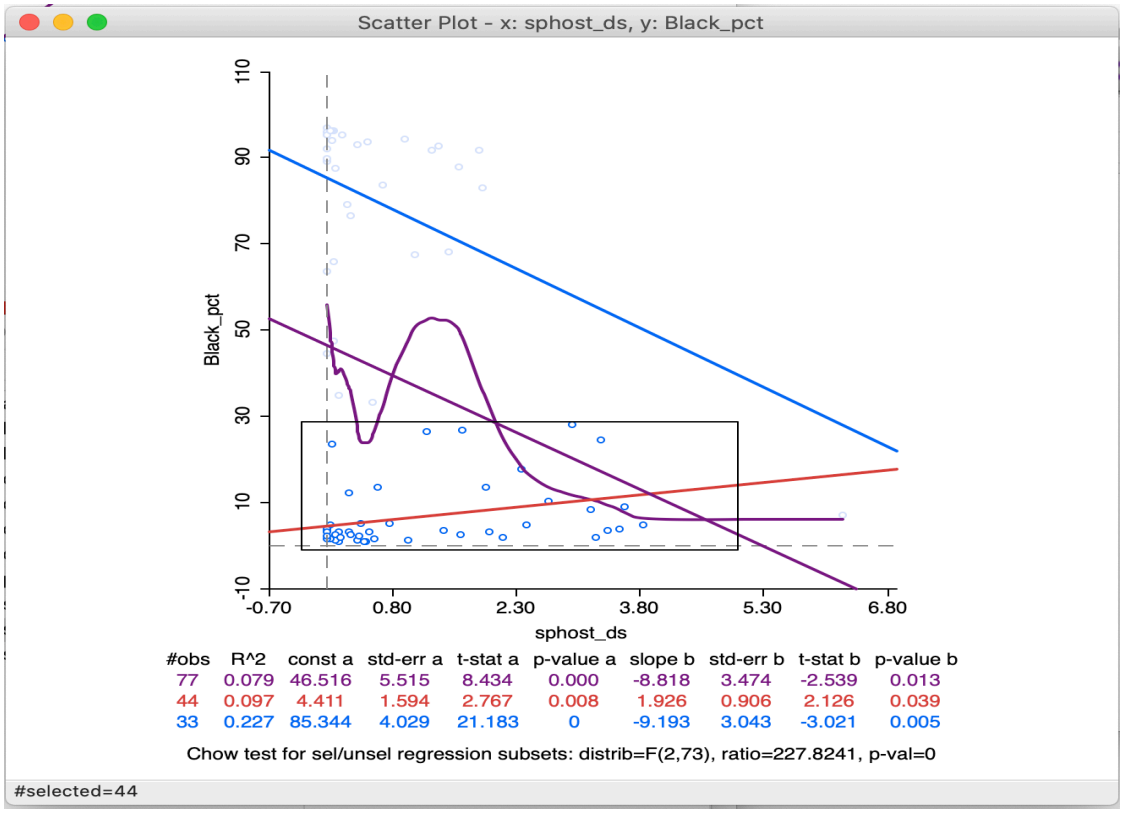

Figure 14: Scatter Plot for black_pct

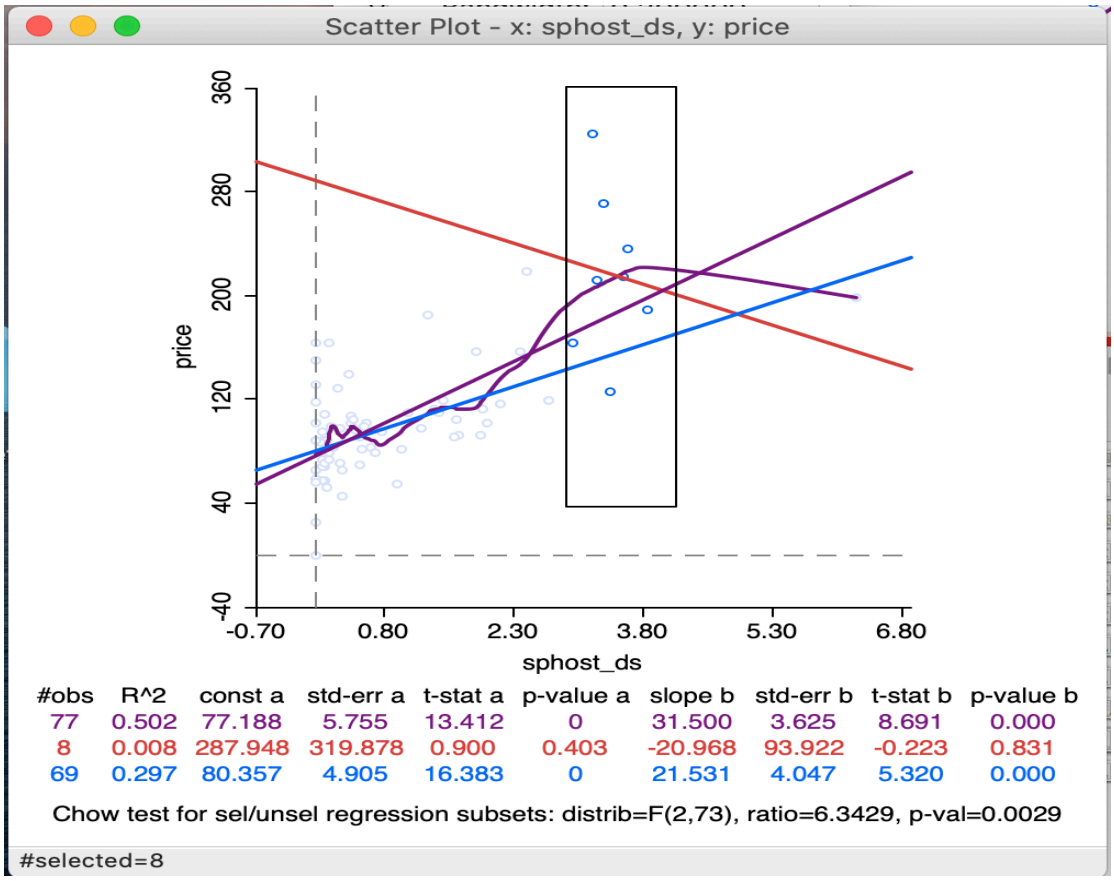

Figure 15: Scatter Plot for price

Figure 16: Centroid Layer for Community Polygon

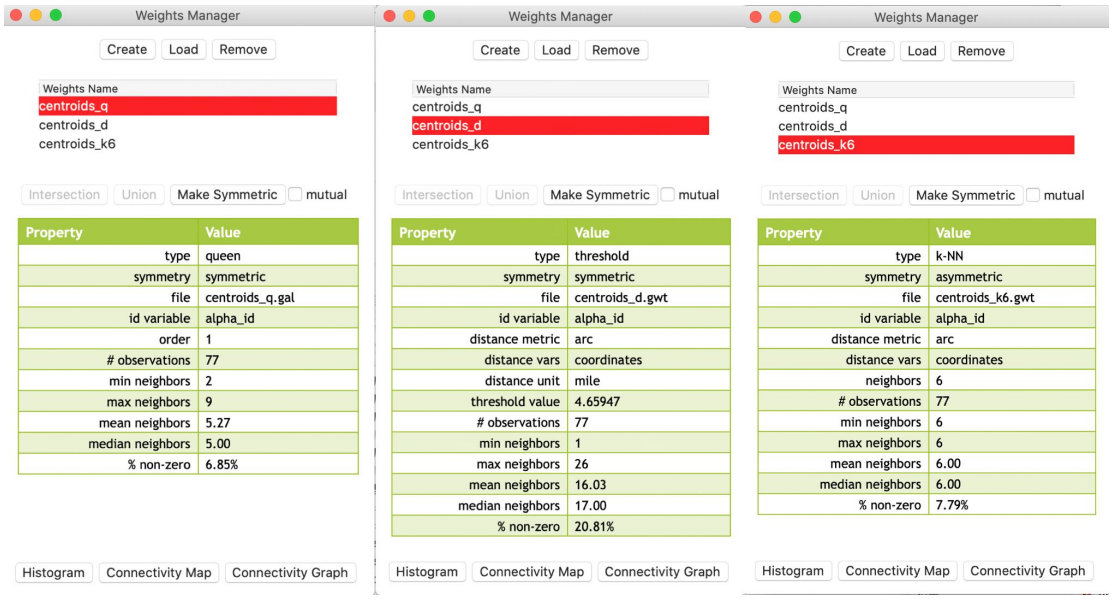

Figure 17: Spatial Weights for Polygon Layer

Figure 18: Spatial Weights for Centroid Layer

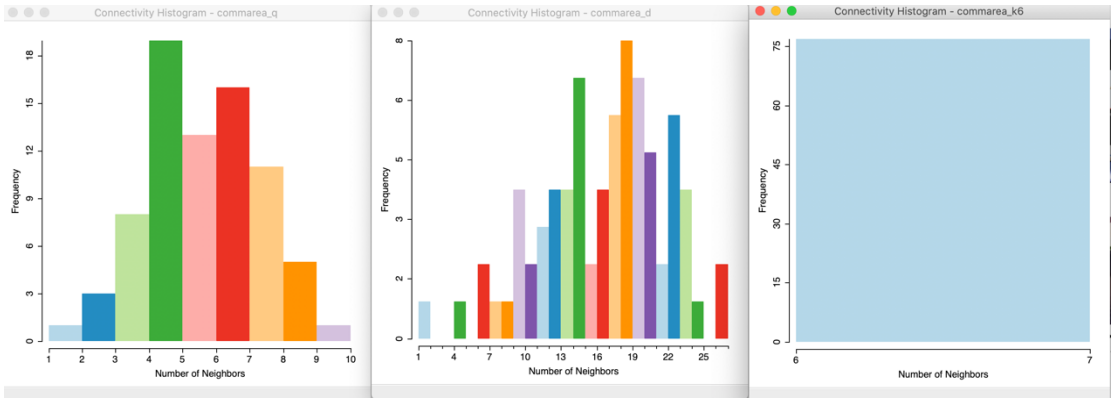

Figure 19: Connectivity Histograms

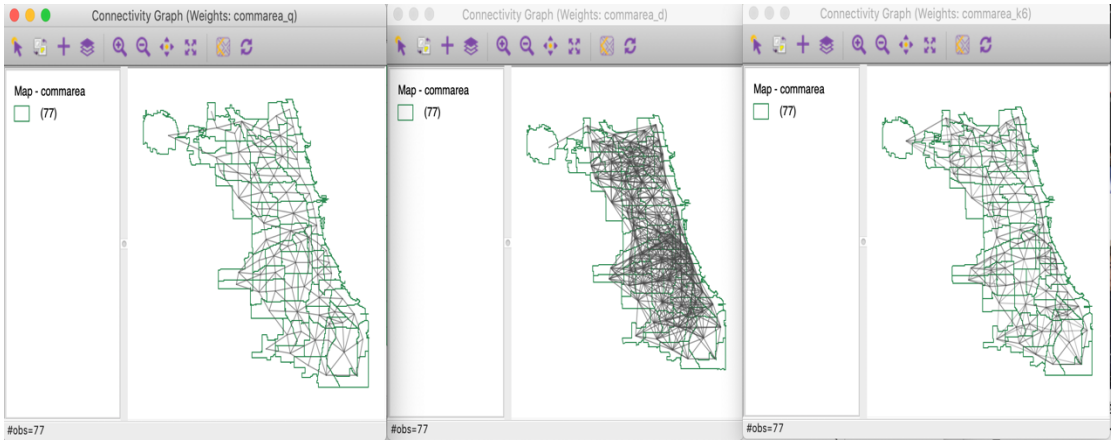

Figure 20: Connectivity Graphs for Polygon Layer

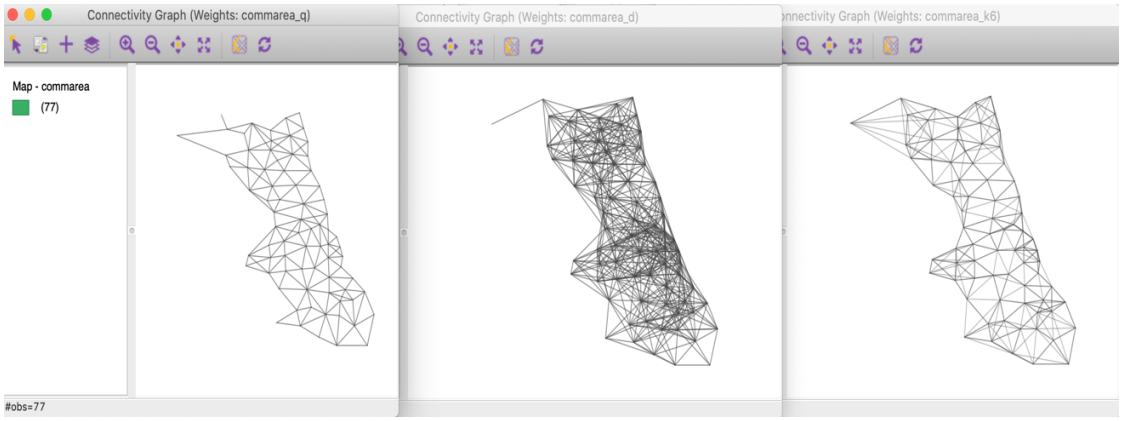

Figure 21: Connectivity Graph for Controid Layer

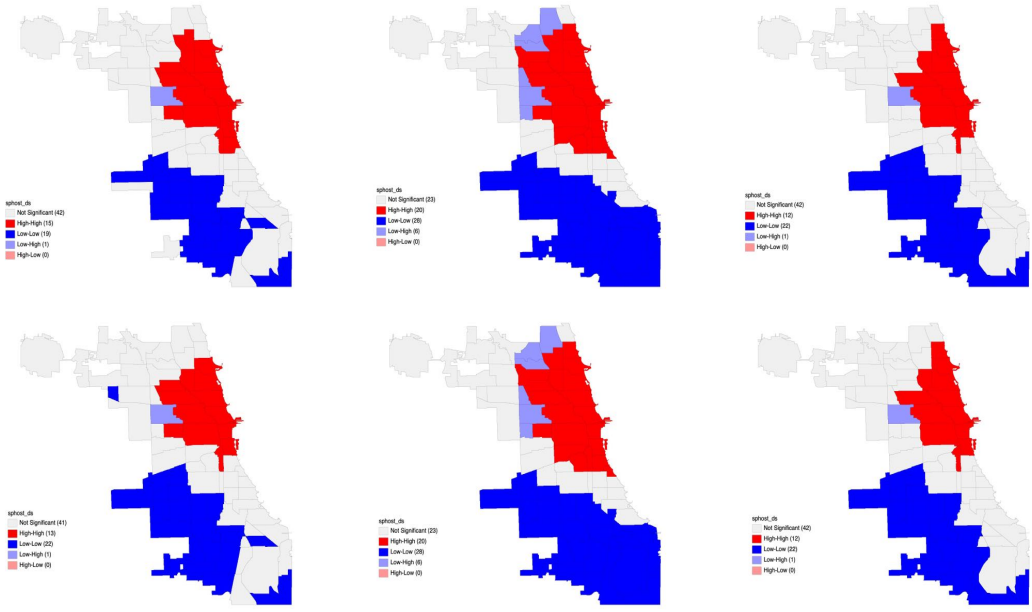

Figure 22: Cluster Map of Local Moran's I for sphost_ds

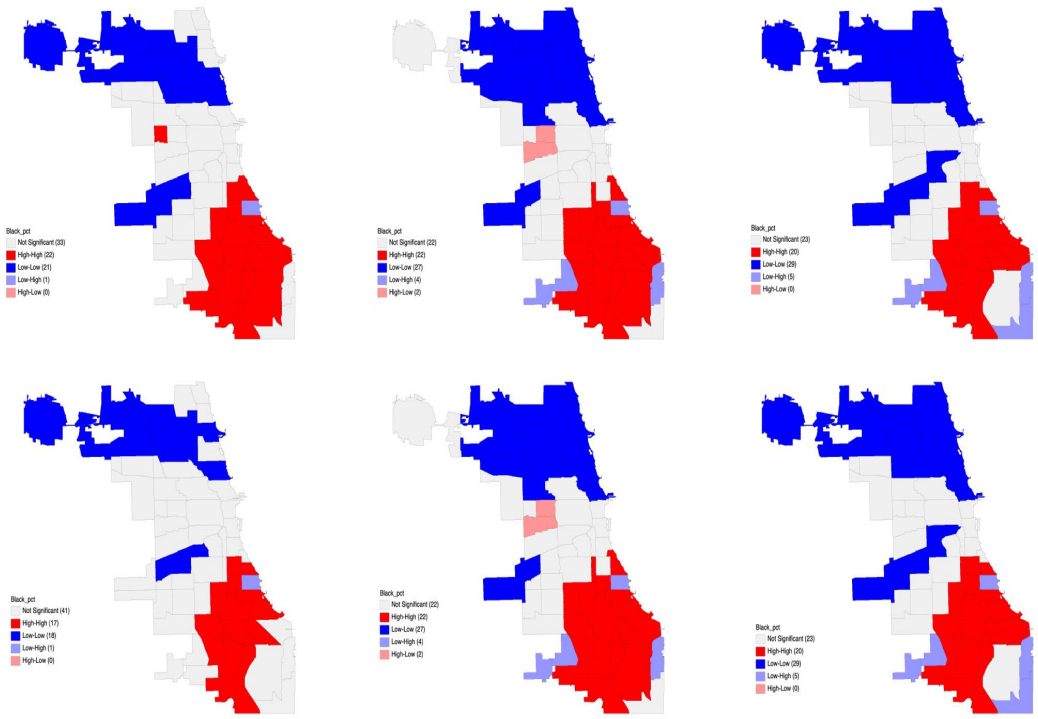

Figure 23: Cluster Map of Local Moran's I for black_pct

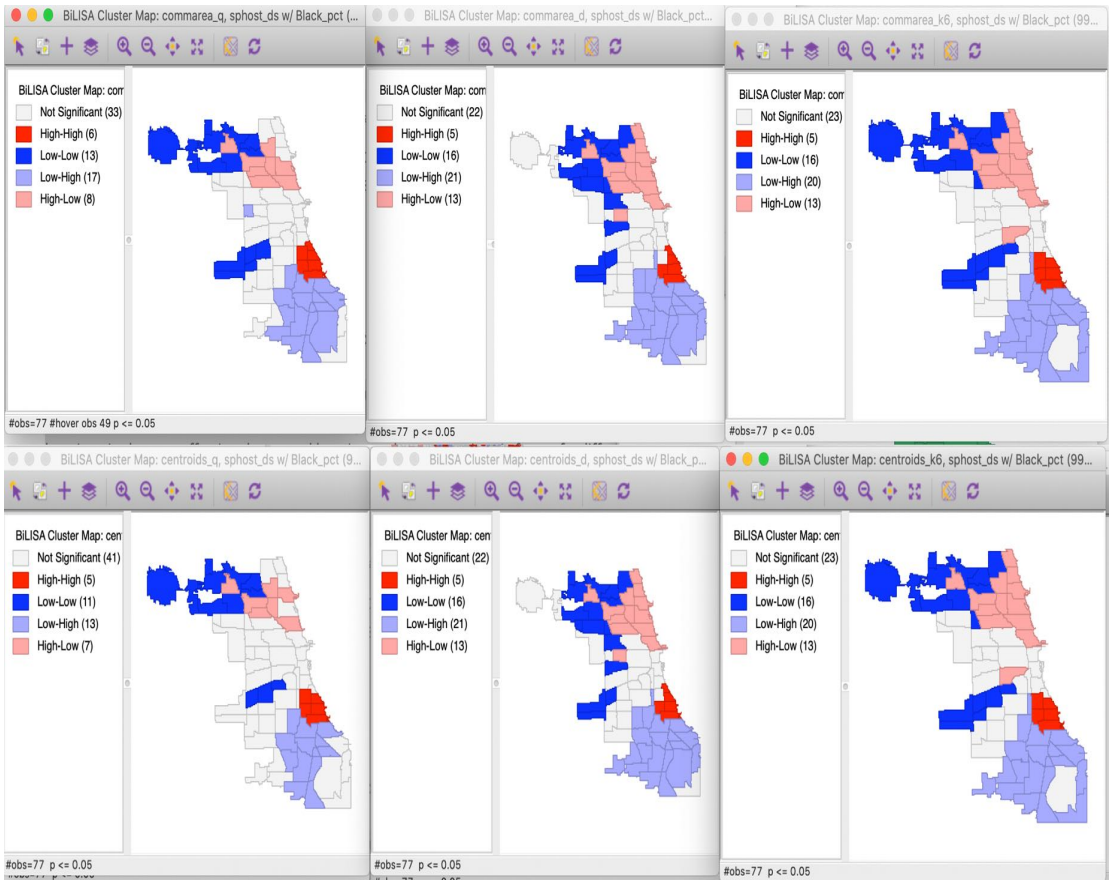

Figure 24: Bivariate Local Moran Cluster Map

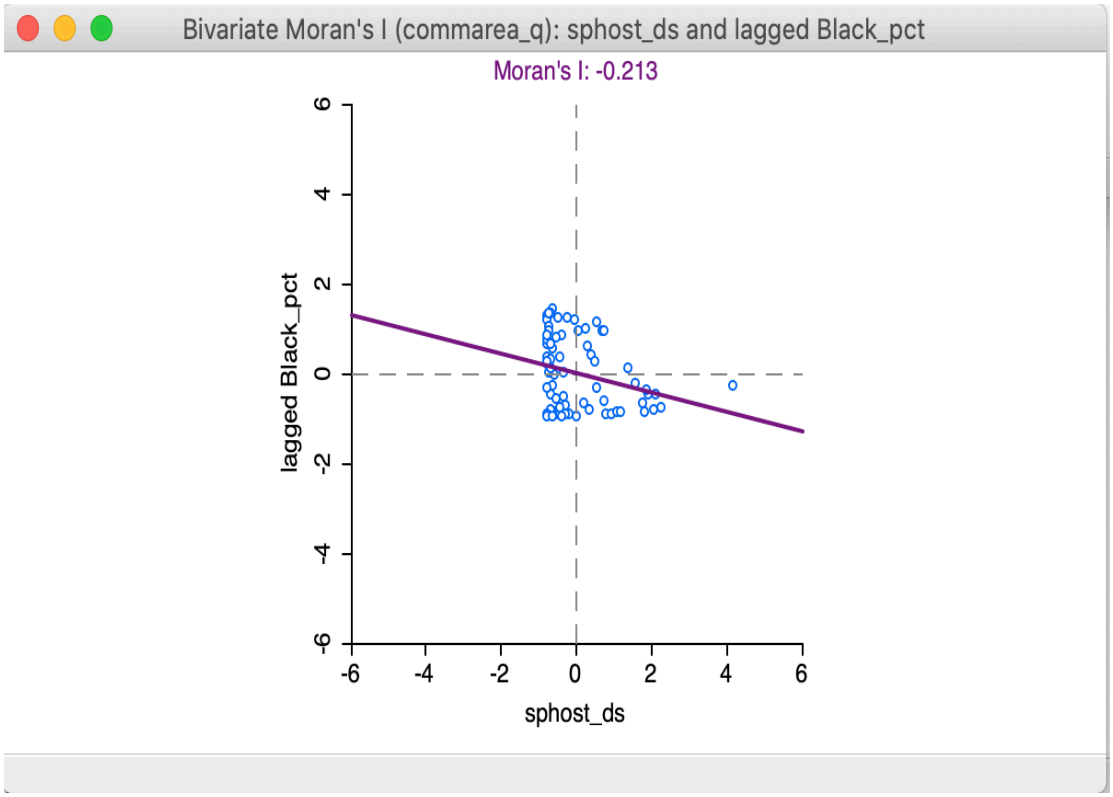

Figure 25: Scatter Plot for Bivariate Moran's I

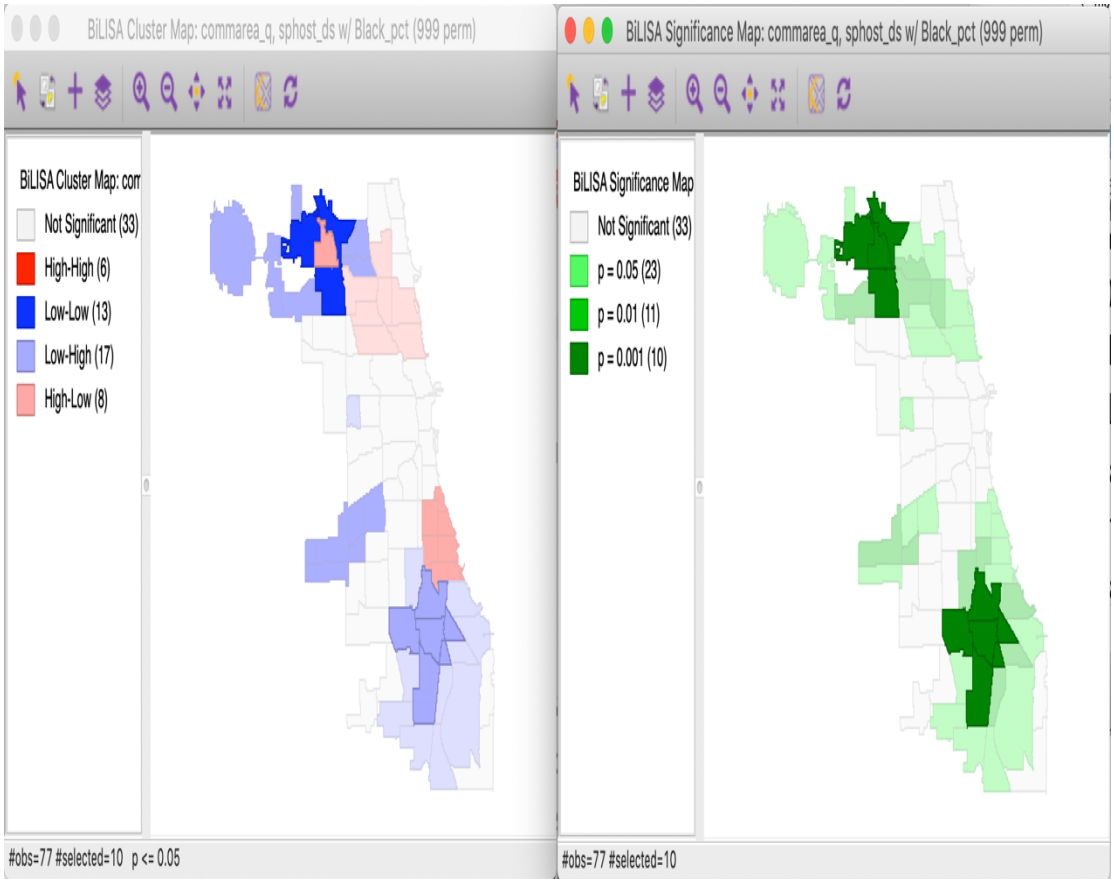

Figure 26: Bivariate Local Moran Cluster Map & Significance Map

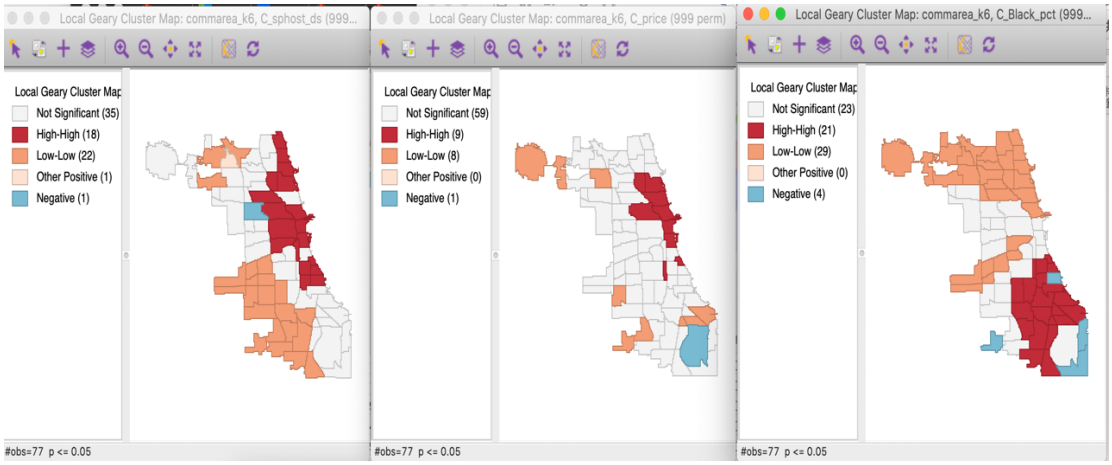

Figure 27: Univariate Local Geary Cluster Map (KNN)

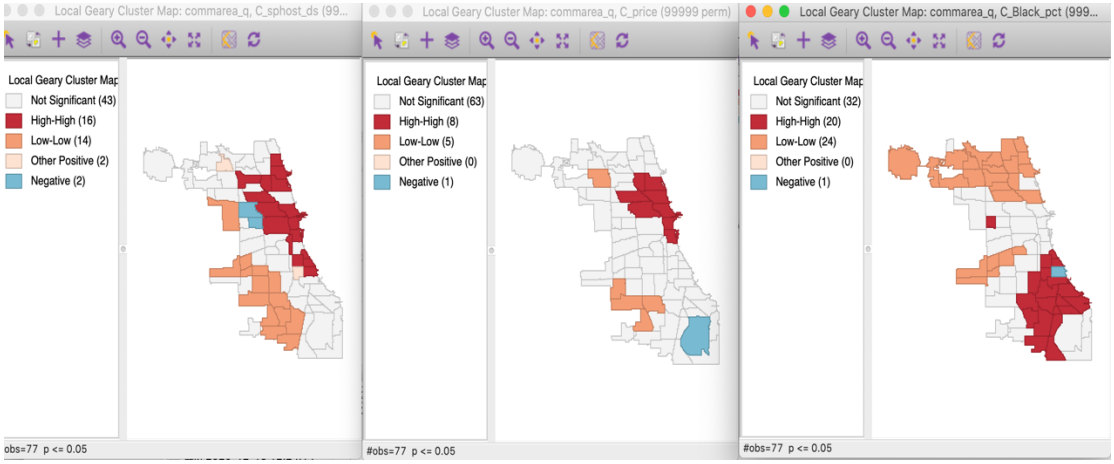

Figure 28: Univariate Local Geary Cluster Map (Queen)

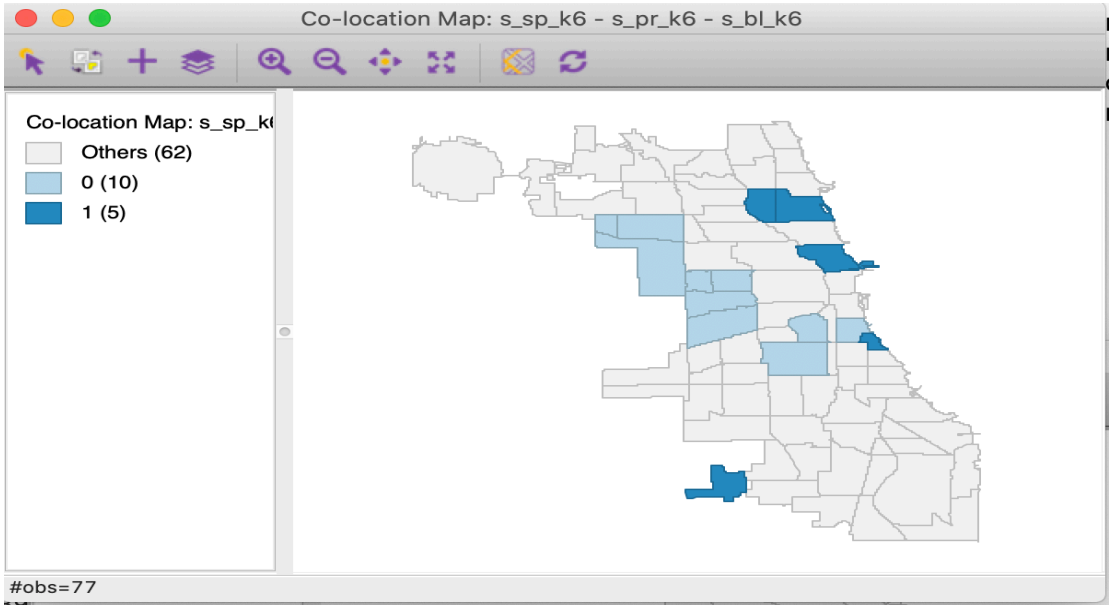

Figure 29: Co-location Map (KNN)

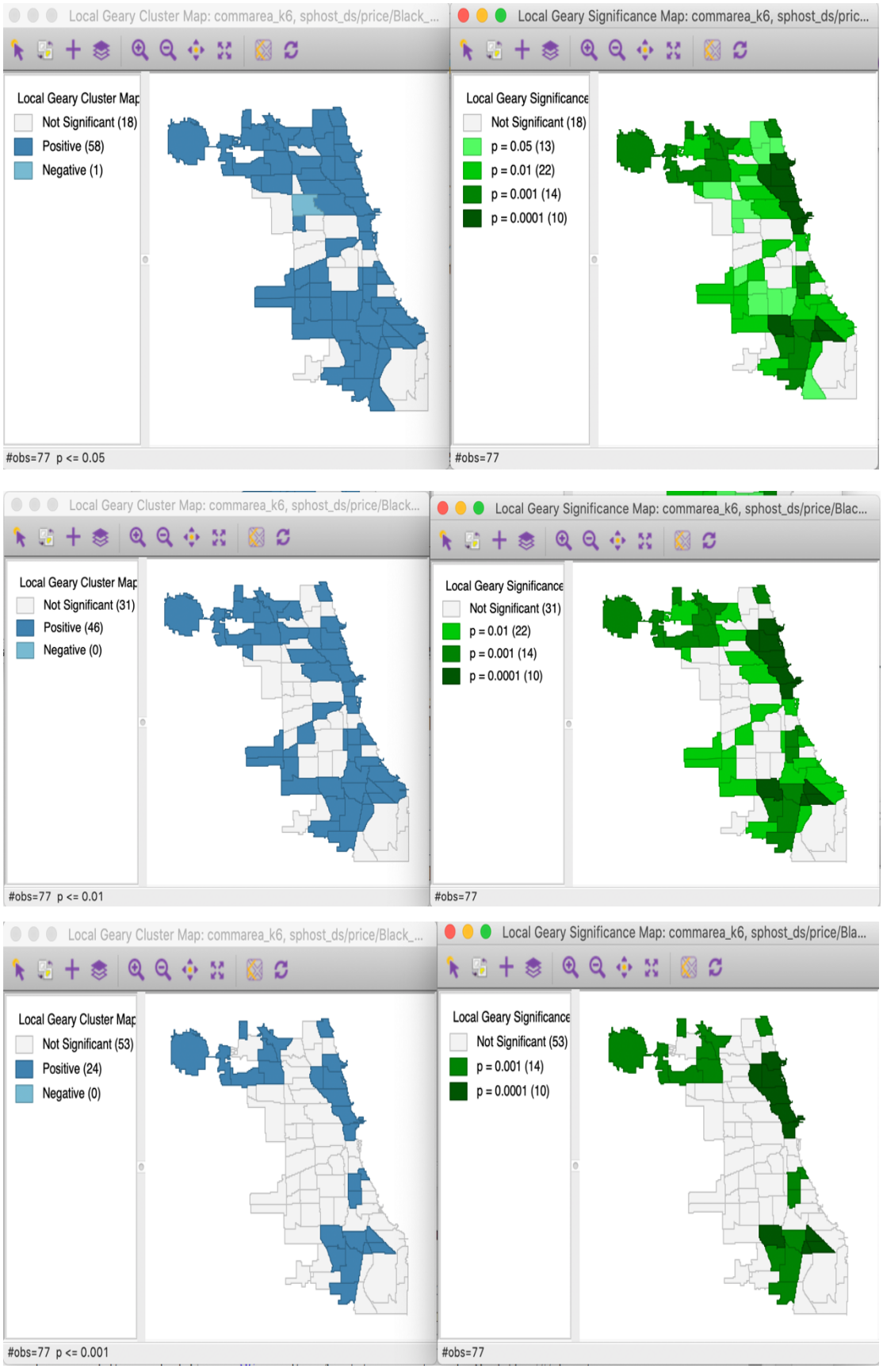

Figure 30: Multivariate Local Geary (KNN)

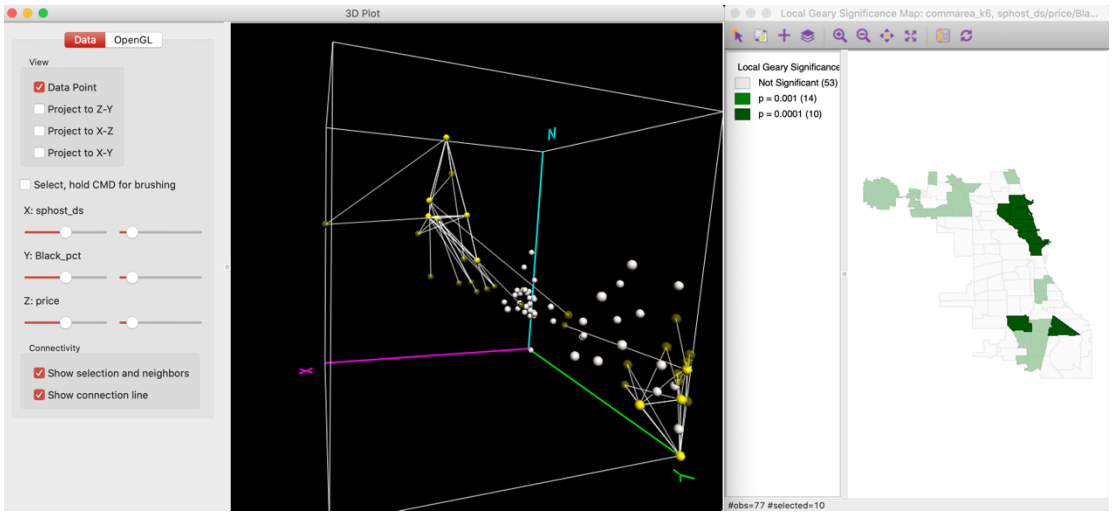

Figure 31: 3D Plot

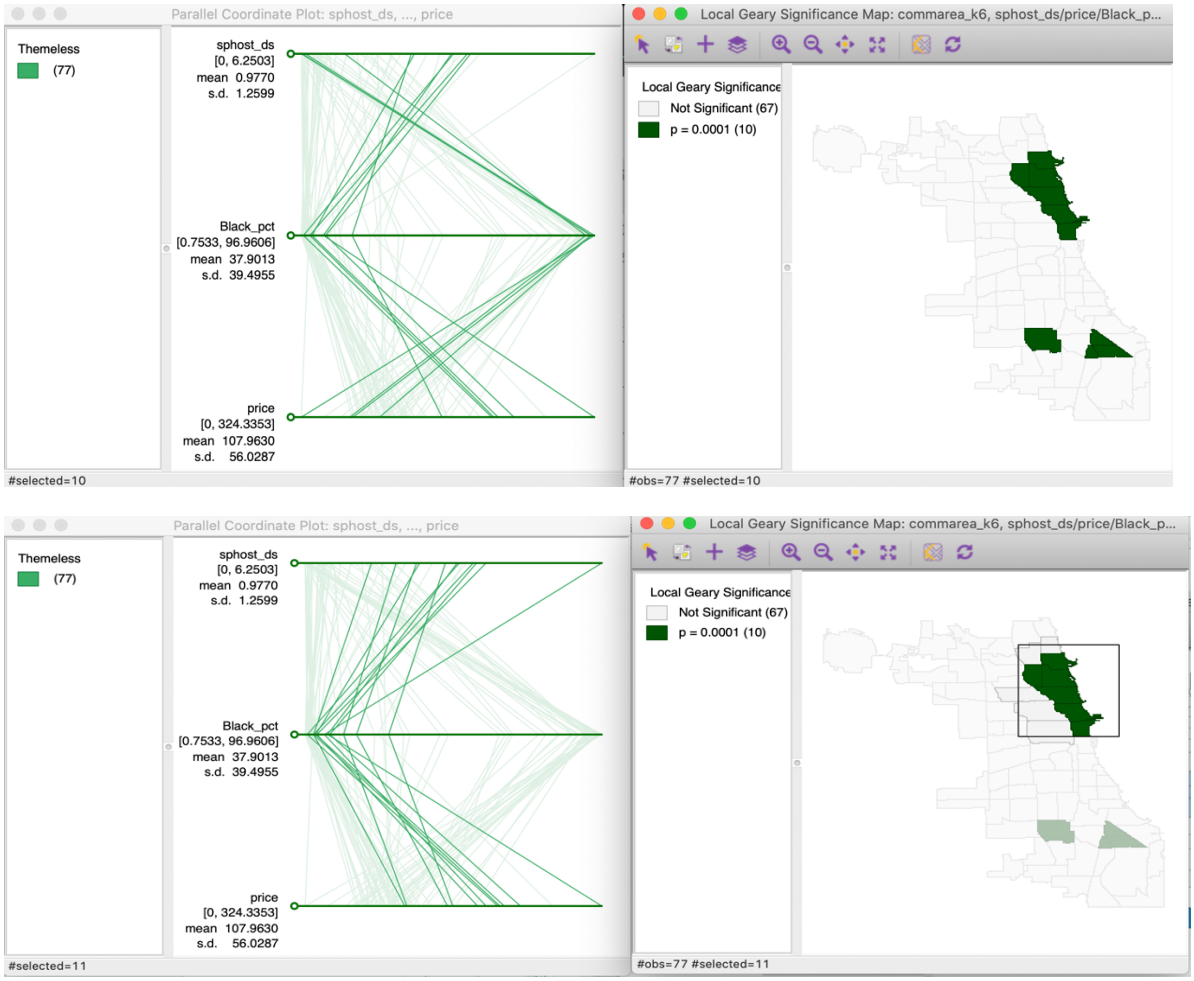

Figure 32: Parallel Coordinate Plot (1)

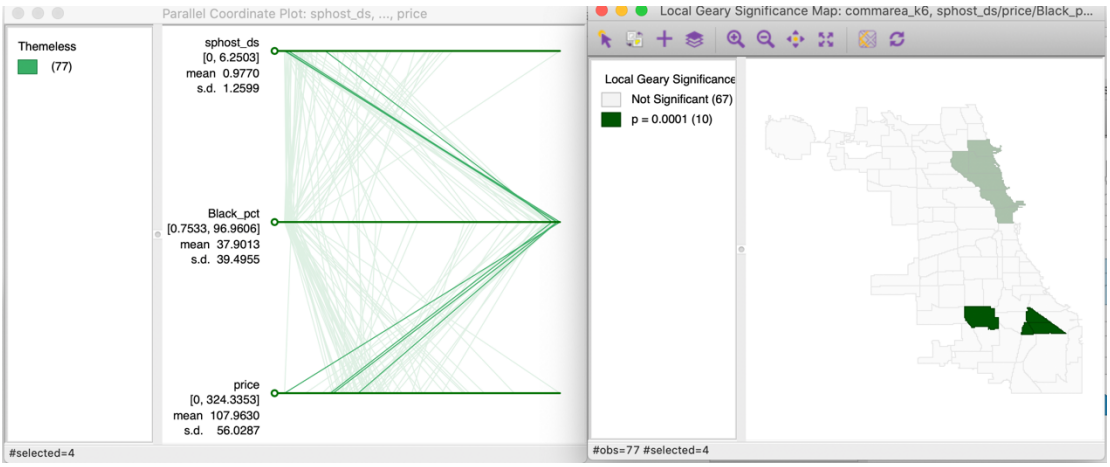

Figure 32: Parallel Coordinate Plot (2)

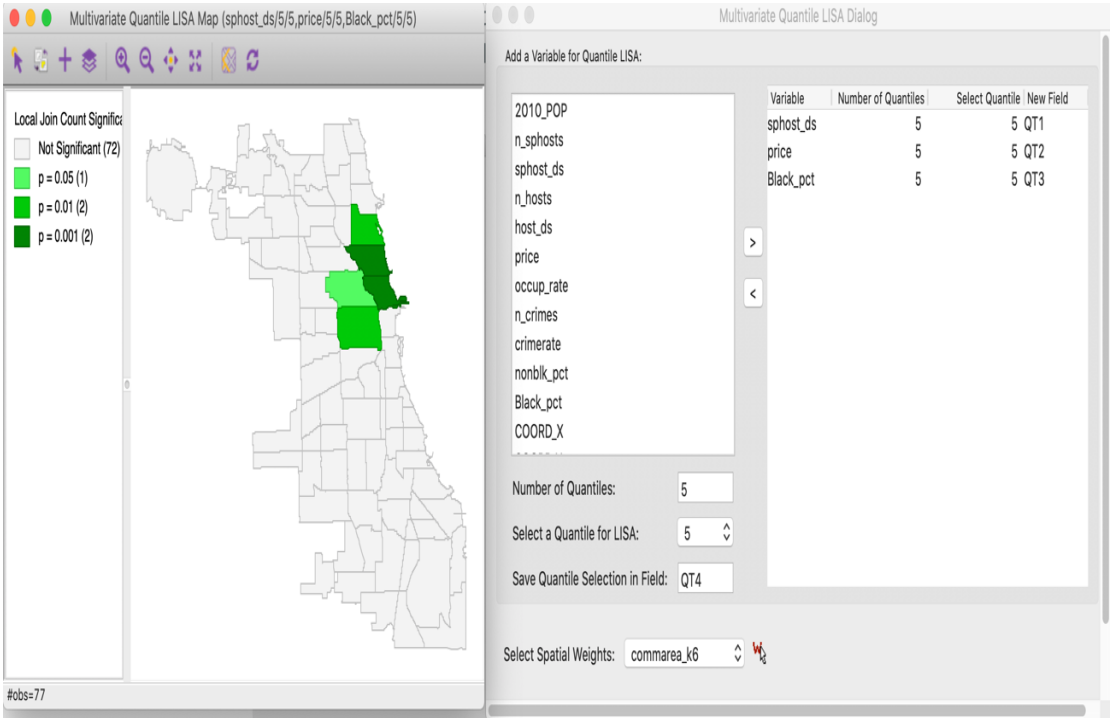

Figure 33: Mutivariate Quantile LISA

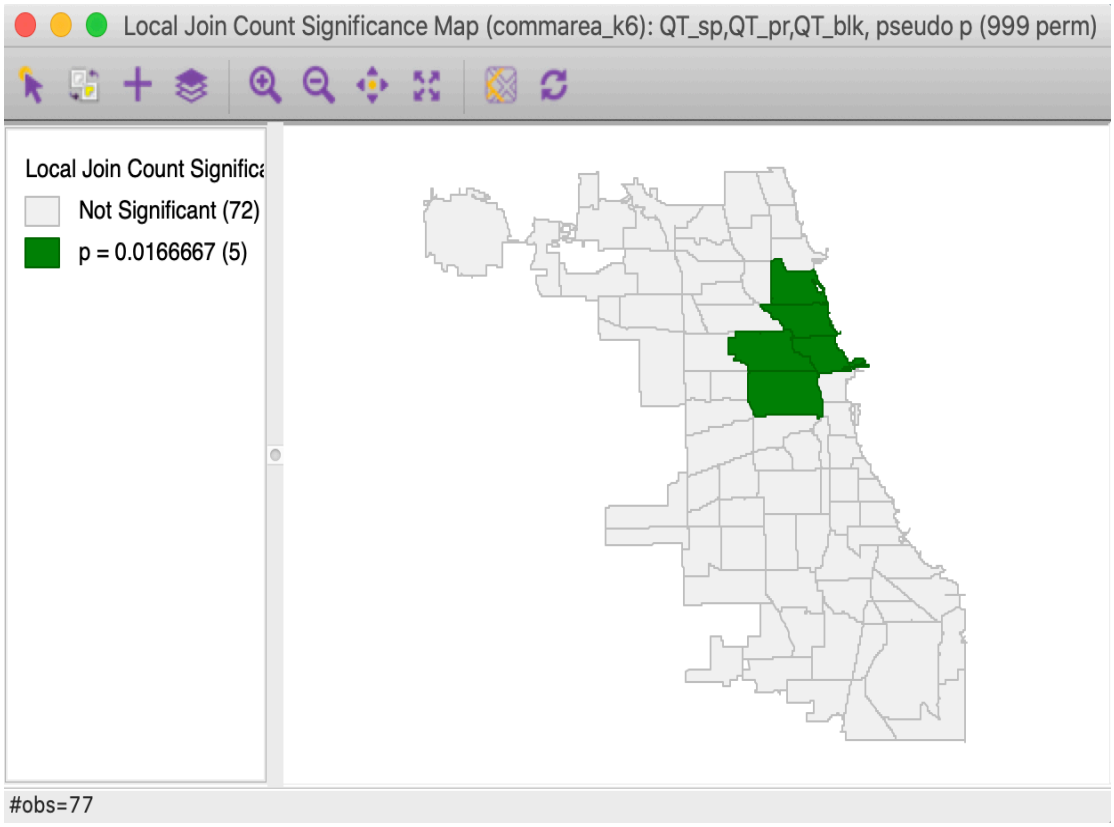

Figure 34: Mutivariate Quantile LISA (with FDR)

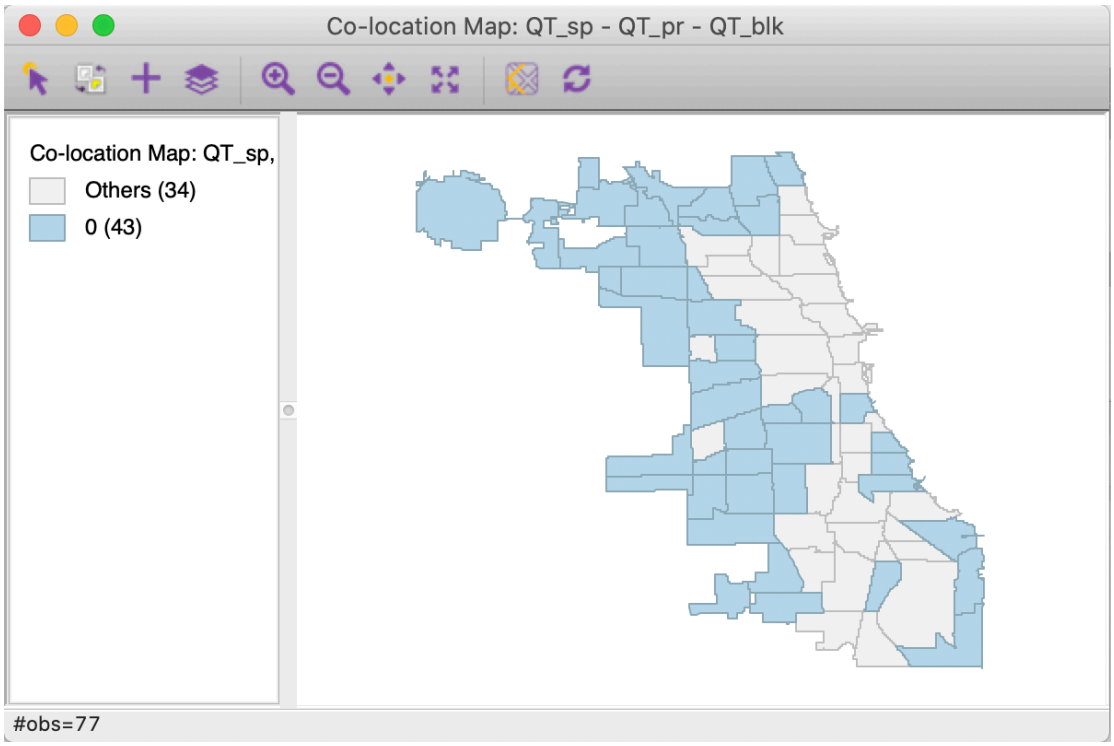

Figure 35: Co-location Map (All 5th Quantile)

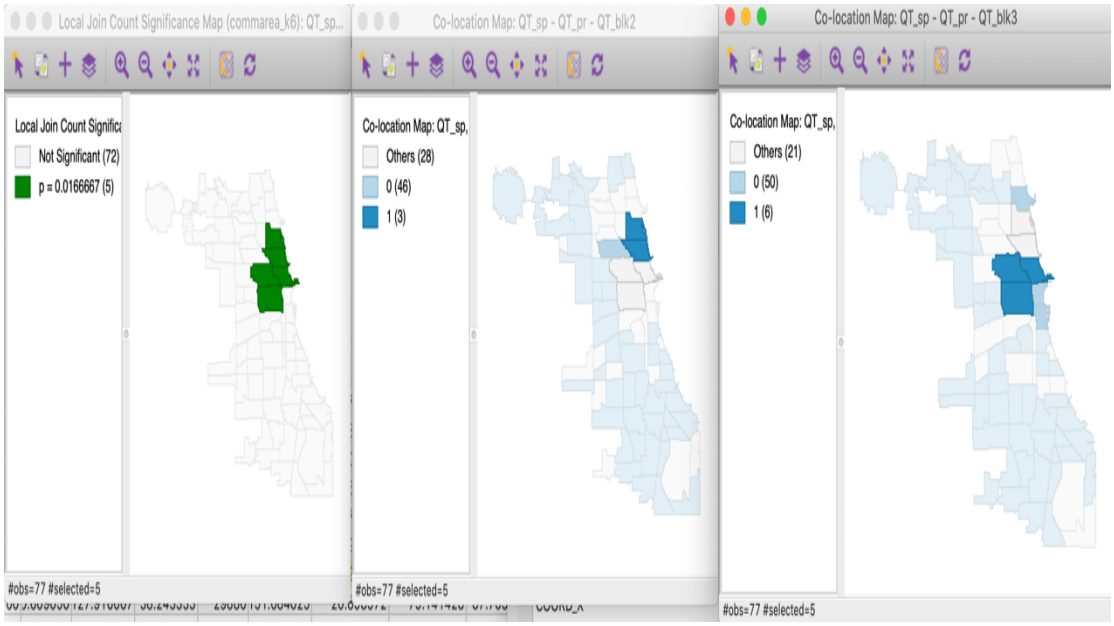

Figure 36: Co-location Map (2nd and 3rd Quantile of black_pct)