The Spatial Distribution of Farmers Markets in Philadelphia

1. Introduction

Farmers markets constitute a dominant component in a city and bring residents benefits in multiple aspects. They not only provide people with fresher, seasonal, healthier and a better variety of foods at affordable price and reachable distance, but also provide people a place to meet neighbors, enjoy an outdoor leisure walk while getting needed groceries. The access to farmers markets ensures people get fresh and healthy food timely and conveniently, thus maintaining a healthy and comfortable life. However, because of the unbalanced economic development, not all neighborhoods, especially parts of southern, northern, and northeast Philadelphia have access to farmers markets, which deprive local residents of the numerous benefits offered by farm markets. In order to provide policy implications and suggestions for local governments on this problem, this study examines the spatial distribution of farmers markets in Philadelphia to identify whether farmers markets are clustered by conducting nearest neighbor and K-function analysis in ArcGIS.

2. Methods

2.1 Quadrat Method

In this study, we will test whether the point pattern of farm markets is random or not. The null hypothesis states that the point pattern of farm markets is random, while the alternative hypothesis states that the point pattern of farm markets is clustered. A study region is subdivided into a lattice containing a bunch of square cells in equal size, known as quadrats. We define that a point process is completely spatially random (CSR) when two conditions are satisfied. First, the probability of a point lands into any quadrat is directly proportional to the area of this quadrat. The point is equally likely to land in any quadrat if the sizes of quadrata are equal, Second, where one-point lands does not affect where any other points land. In other words, point i should locate independently of point j.

Quadrat method is a measure of point density. It first splits the study region into several cells, then summarizes the number of points in each cell, calculates the variance divided by the mean of how many points each cell has, known as VMR. The VMR value decides the point density pattern. There are obviously a number of limitations in the quadrat method, since it highly depends on the size of quadrat without the consideration of the distances between each points and how these points are arranged in space. Because of the different cell size it chooses, the quadrat method is likely to miscalculate the same point density pattern into different patterns, or

different point density patterns are miscalculated into the same. Considering these limitations, the quadrat method is not generally used in practice and in this assignment, we will conduct nearest neighbor analysis and k-function analysis to identify the point density pattern.

2.2 Nearest Neighbor Analysis Method

The Nearest Neighbor Analysis examines the distance between each point and its closest point, then from a CSR pattern, compares these to expected values for a random sample of points ("Nearest Neighbor Analysis" 2021). It is measured with nearest neighbor index (NNI), an index based on the distance between points with the following formula:

$$NNI = \frac{Observed\ Average\ Distance}{Expected\ Average\ Distance\ (when\ pattern\ is\ random)} = \frac{\overline{D}_O}{\overline{D}_E} \quad (1)$$

Where $\overline{D}_O = \frac{\sum_{i=1}^{n} D_i}{n}$, which denotes the average of all observed average distance between each point and its nearest neighbor, while $\overline{D}_E = \frac{0.5}{\sqrt{n/A}}$, which denotes the expected average distance between each point and its nearest neighbor given that point pattern was random. n is the number of features and A is the area of the rectangle around the points.

We determine whether we have a significant clustering or dispersion pattern according to the value of NNI. If the observed average distance is close to the expected average distance given a random pattern, then the NNI is close to 1, which means that we have a random pattern. If the observed average distance is close to 0, suggesting that all the points are located at almost the same spot, then NNI is close to 0, suggesting that we have a clustered pattern. If the observed average distance is much greater than the expected average distance given the random pattern, then NNI is close to 2, which indicates a dispersed pattern.

We will conduct a hypothesis test to determine whether there is a significant clustering or dispersion. The null hypothesis is that the observed point pattern is random, where it isn't significantly different from the expected point patterns. The alternative hypothesis states that the observed point pattern is not random, where it is either significant clustering or dispersion.

The test statistic has a z (standard normal) distribution with the following equation:

$$z = \frac{\overline{D}_O - \overline{D}_E}{SE_{\overline{D}_O}} = \frac{\frac{\sum_{i=1}^{n} D_i}{n} - \frac{0.5}{\sqrt{\frac{n}{A}}}}{\frac{0.26136}{\sqrt{\frac{n^2}{A}}}} \quad (2)$$

Where SE is the standard error of $\overline{D}_O$.

According to the standard normal table, we can get the p-value from calculated z. In a two-tailed test, $\overline{D}_E \neq \overline{D}_O$, which fit for the alternative hypothesis. z = |1.96| corresponds to an α-value of 0.05. Therefore, if z>1.96 or z<-1.96, we reject H_0 for H_a at α=0.05. Specifically, if z>1.96, we have significant dispersion, which implies that the observed average distance is much greater than the expected average distance, while z<-1.96, we have significant clustering, implying that the observed average distance is much smaller than the expected average distance.

There are still some limitations of Nearest Neighbor Analysis. First, it only considers the average distance to only the nearest neighbor. Second, it greatly depends on the area of the study region. If this area is calculated with the minimum enclosing rectangle, then it would be strongly affected by outliers. Third, it does not consider the fact that both clustering and dispersion may be presented at different scales. For example, there are several hospitals clustering in the center city of Philadelphia, which has an irregular boundary. Other than specifying the shape of the boundary, the nearest neighbor tool uses a minimum enclosing rectangle boundary to cover all points, so the results still show a random pattern, even though those hospitals are obviously clustered in a city scale.

2.3 K-function Analysis Method

K-function is used to illustrate how clustering or dispersion changes when the neighborhood size changes. It places circles with radius d around every point in a plane, counts the number of other points within each circle, and calculates the average number of other points in all circles of radius d. Then, we divide the average count of events by the overall point density in the plane to get the K-function at distance d, denoted as K(d). The formula for the K(d) is presented below:

$$K(d) = \frac{(\sum_{i=1}^{n} \#[S \in Cirlce(s_i, \boldsymbol{d})])/n}{\frac{n}{a}} = \\ = \frac{Mean\ numbers\ of\ points\ in\ all\ circles\ of\ radius\ \boldsymbol{d}}{Mean\ point\ density\ in\ entire\ study\ region\ \boldsymbol{a}} \quad (\mathbf{3})$$

Where:

- n indicates the number of points in the dataset;
- $\boldsymbol{a}$, denoted the area of the study region;
- s_i, the center point of a circle;
- $\boldsymbol{d}$, the radius of a circle;

$\boldsymbol{K(d)}$ is the mean number of events we observe within a circle of radius $\boldsymbol{d}$ taking into account the overall point density in the study region. Under CSR, it turns out that $K(d) = \pi d^2$. When $K(d) > \pi d^2$, it implies clustering at scale d; When $K(d) < \pi d^2$, it implies dispersion at scale d.

Some statistical software packages adopt an L(d) function, rather than the K(d) function for the point pattern analysis. The formula for L(d) is as below where π=3.14:

$$L(d) = \sqrt{\frac{K(d)}{\pi} - d} \tag{4}$$

- Under CSR, L(d) = 0;
- When L(d) > 0, there is clustering at the scale of d;
- When L(d) < 0, there is no clustering at the scale of d.

However, ArcGIS uses a slightly different L(d) function as below:

$$L(d) = \sqrt{\frac{K(d)}{\pi}} \tag{5}$$

- Under CSR, L(d) = d;
- When L(d) > d, clustering;
- When L(d) < d, dispersion.

Based on Ripley's K-function, the Multi-Distance Spatial Cluster Analysis tool is another way to analyze the spatial pattern of incident point data. It summarizes spatial dependence (feature clustering or feature dispersion) over a range of distances. Despite the ambiguity of the default value set by the ArcGIS, we should consider the maximum distance that pairwise distance between two points in our point pattern divided by 2. The following function is a good way to calculate the beginning and incremental distance:

$$\mathrm{d} = \frac{\frac{1}{2} \cdot maximum\ pairwise\ distance}{(\#\ of\ distance\ bands)} \tag{6}$$

The testing of the K-function is based on the randomly permuted point patterns. We propose the following hypothesis:

- H_0: At distance d, the pattern is random;
- H_{a1}: At distance d, the pattern is clustered;

- H_{a2}: At distance d, the pattern is uniform.

We generate many point patterns randomly with n points each, then calculate the L(d) values for each of the patterns. From the L(d) values calculated, we find the lowest value of L(d), referred as $L^{-}(d)$ or *Lower Envelope*, and the highest value of L(d), referred as $L^{+}(d)$ or *Higher Envelope*. Then, for each distance d, we compare the observed value of L(d), denoted by$L^{obs}(d)$, to the Lower Envelope and Higher Envelope:

- If $L^{-}(d) < L^{obs}(d) < L^{+}(d)$, then we can't reject H_0 at distance d. That is, we cannot say that at distance d, the pattern is significantly different from what we'd except under CSR.
- If $L^{obs}(d) > L^{+}(d)$, then we can reject H_0 at distance d for H_{a1} – that is, we have significant clustering at scale d.
- If $L^{obs}(d) < L^{-}(d)$, then we can reject H_0 at distance d for H_{a2}: – that is, we have significant dispersion at scale d.

The confidence envelope is calculated by randomly placing feature points in the study area. Each set of random placements is called a "permutation" and the confidence envelope is constructed from random permutations. the values defining the confidence envelope will change from one run to the next. So, we need to set up a seed value of the number of permutations, then repeating analyses will produce consistent results. The number of permutations selected for the Compute Confidence Envelope parameter may be translated to confidence levels: 9 for 90%, 99 for 99%, and 999 for 99.9%. For instance, when we have 999 permutations and Lobs(d) < L-(d) at some distance d, then we can be approximately 99.9% confident that we have significant dispersion at that distance d.

It's common to have some points that appear to be located at the border of a plane, thus circles for the points in the border do not have a full shape. In order to overcome that issue, the ArcGIS introduces the Ripley's Edge Correction Formula for study regions shaped as rectangles, which assign weights to circles around the points based on the percentage of the areas of the circle within the plane. Alternatively, the Simulate Outer Boundary Values mirrors points across the study area boundary to correct for underestimates near edges. Points that are within a distance equal to the maximum distance band of the edge of the study area are mirrored. Thus, it may provide more accurate neighbor estimates based on points mirrored. In this study, we are going to use the Simulate Outer Boundary Values for correction, because Ripley's Edge Correction only works for rectangular study areas in ArcGIS.

The homogeneousness is an issue we need to take into consideration when we are doing the K-function analysis: sometimes point patterns that are clustering is not because it is naturally distributed like that, but because of other factors that are also clustering: for example, social resources are allocated/clustered in places where there are more population. In ArcGIS, we have

a function called “Spatially Balanced Points" that generates a set of sample points based on a raster that shows “inclusion probabilities” by density. First, we convert population values to probabilities, and convert shapefile to raster format with population density applied. Then we create spatially balanced points, permute random point patterns for 9, 99, or 999 times. During the process of calculating L(d) for each patter, we select to permutations under compute confidence envelope, calibrated by selecting “simulating outer boundary values”, and join the output table for each of the permuted point patterns by distance Expected K. Finally, we can find which distance have clustering, random or dispersion pattern in the resulting output tables.

3. Results

3.1 Initial Nearest Neighbor Analysis Result

By conducting the nearest neighbor analysis, setting the minimum enclosing rectangle for Philadelphia, the result is shown in figure 1, where the nearest neighbor ratio (0.9954) is close to 1, the absolute value of z-score (-0.069945) is smaller than 1.96, and p-value (0.9442) is greater than 0.05. Then we failed to reject H_0 that the farm markets are randomly distributed in Philadelphia.

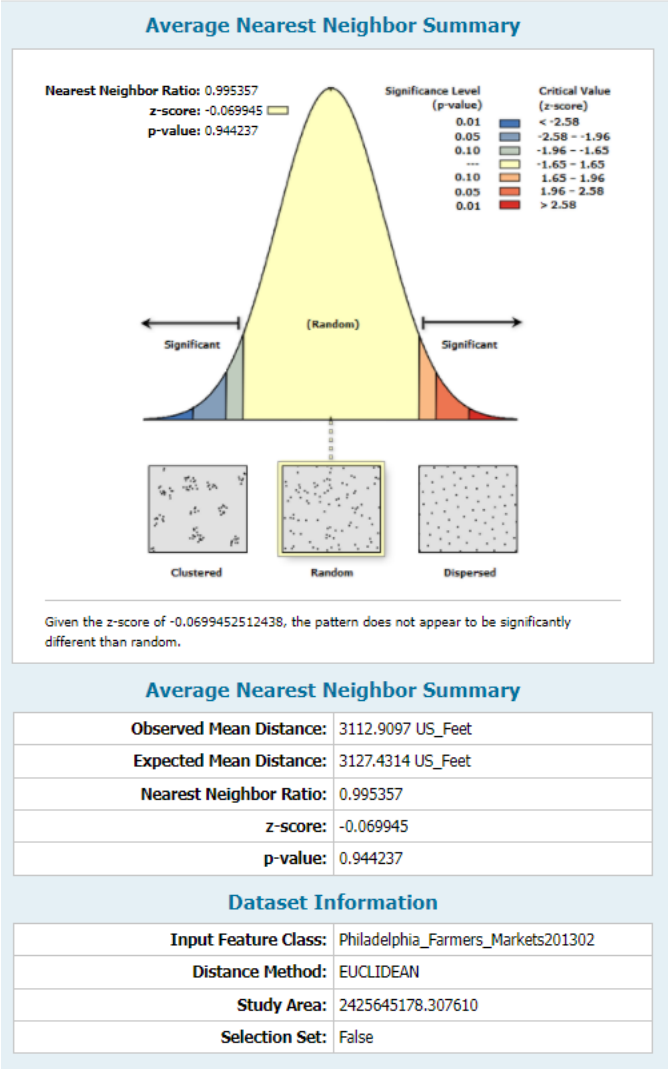

Figure 1. the results of the Nearest Neighbor Analysis

3.2 Second Nearest Neighbor Analysis Result

To solve the common problems generated from nearest neighbor analysis, we re-run the analysis by using the area of Philadelphia rather than the minimum enclosing rectangle as figure 2 shows. The z-score (-3.344634) is much smaller than -1.96 and p-value (0.000824) is much smaller than 0.01. Given the z-score of (-3.344634), there is a less than 1% likelihood that this clustered pattern is the result of random chance. Thus, we reject H_0 for H_a that the farm markets are significantly clustered in Philadelphia.

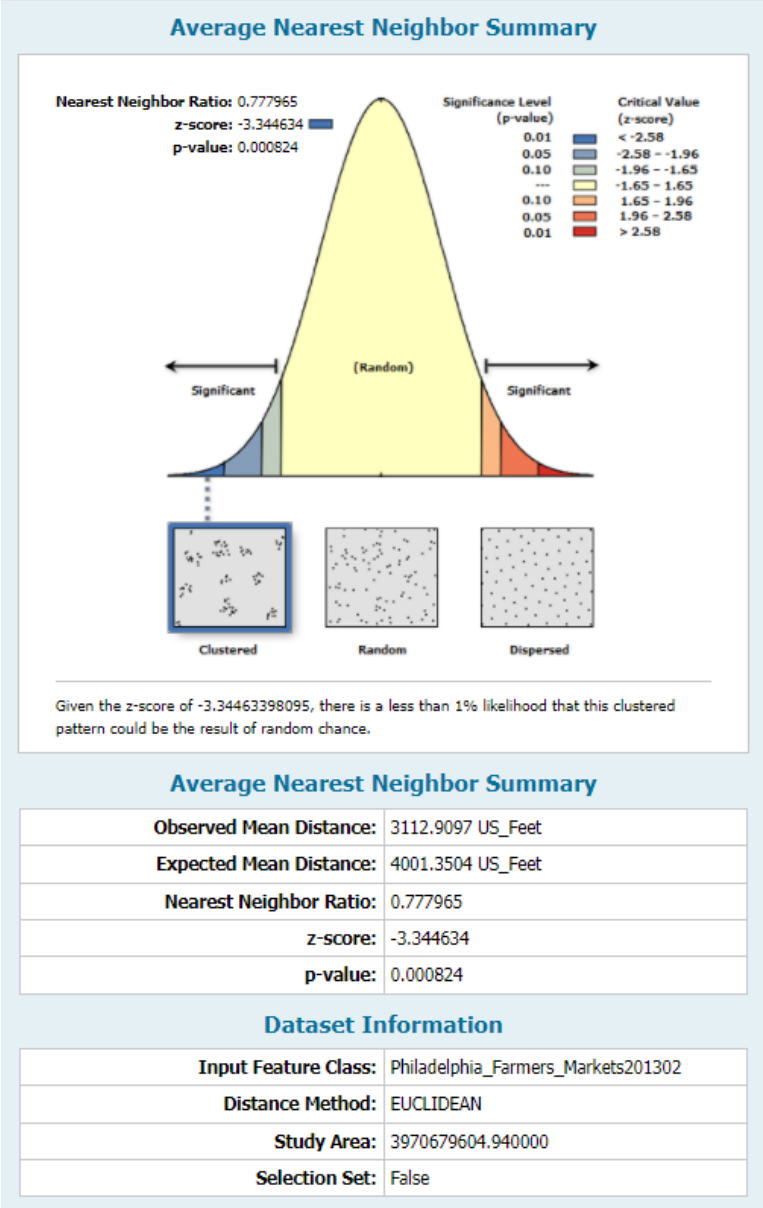

Figure 2. the results of the Nearest Neighbor Analysis, using the area of Philadelphia

3.3 K-function Analysis Result

In the K-function analysis, we set our beginning and incremental distance of 2500 feet under the consideration for it being an appropriate scale when measuring the sizes of urban block groups. By looking at the Table and Figure below, we can see that the Observed K-function is always larger than the Expected K-function for each of the distance values; meanwhile, the Observed K value is always larger than the higher confidence envelope in table 1, and the expected K function is outside the confidence envelope at both small distances and large distances. Thus, we

can reject the null hypothesis and suggesting the spatial clustering for that distance is statistically significant.

By interpreting the clustering, it is not surprising to find out that Northeast, North and South Philadelphia lack farmers markets. However, we argue that such a vacuum is not due to the low population since North Philadelphia has large number of population and South Philadelphia has medium number of populations on figure 4. We think the main reasons are high percentage of poverty and high percentage of parcels being zoned as industrial: the amount of population in those neighborhoods aren't necessarily low. Thus, we may observe different clustering results when we take population factor into analysis but that would not be plausible and applicable for this analysis.

CDP_table

OID	Field1	ExpectedK	ObservedK	DiffK	LwConfEnv	HiConfEnv
0	0	2500	3701.59289	1201.59289	2002.568345	3790.800697
1	0	5000	7691.009666	2691.009666	4287.242692	6691.89508
2	0	7500	11959.653809	4459.653809	7080.148284	8992.996139
3	0	10000	15863.32464	5863.32464	9673.324118	11959.653809
4	0	12500	19372.541598	6872.541598	12385.201216	14601.822449
5	0	15000	22737.45648	7737.45648	14670.322383	17323.47016
6	0	17500	25930.489319	8430.489319	16774.62305	19925.303457
7	0	20000	28846.717239	8846.717239	19242.726366	22501.063247
8	0	22500	31334.48922	8834.48922	21373.762596	24965.262799
9	0	25000	33454.481074	8454.481074	23645.292702	27348.075178

Table 1. K Function Results Output Table

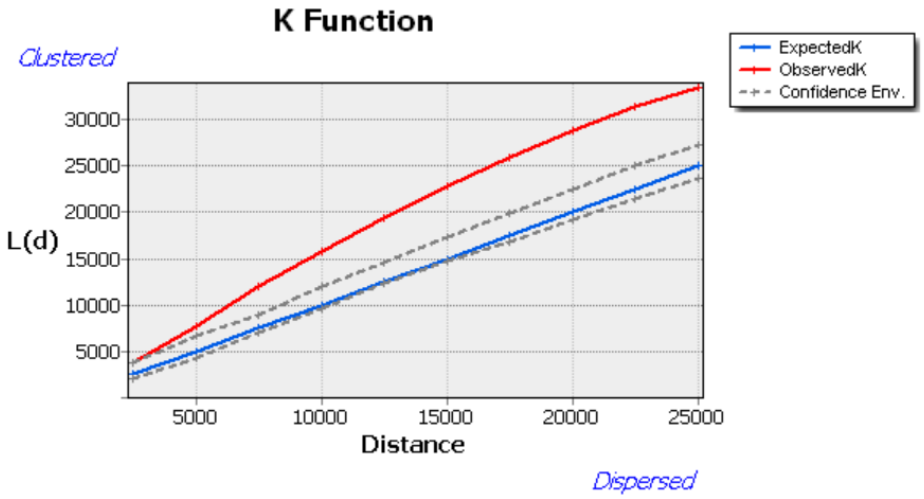

Figure 3. K Function Results Plot

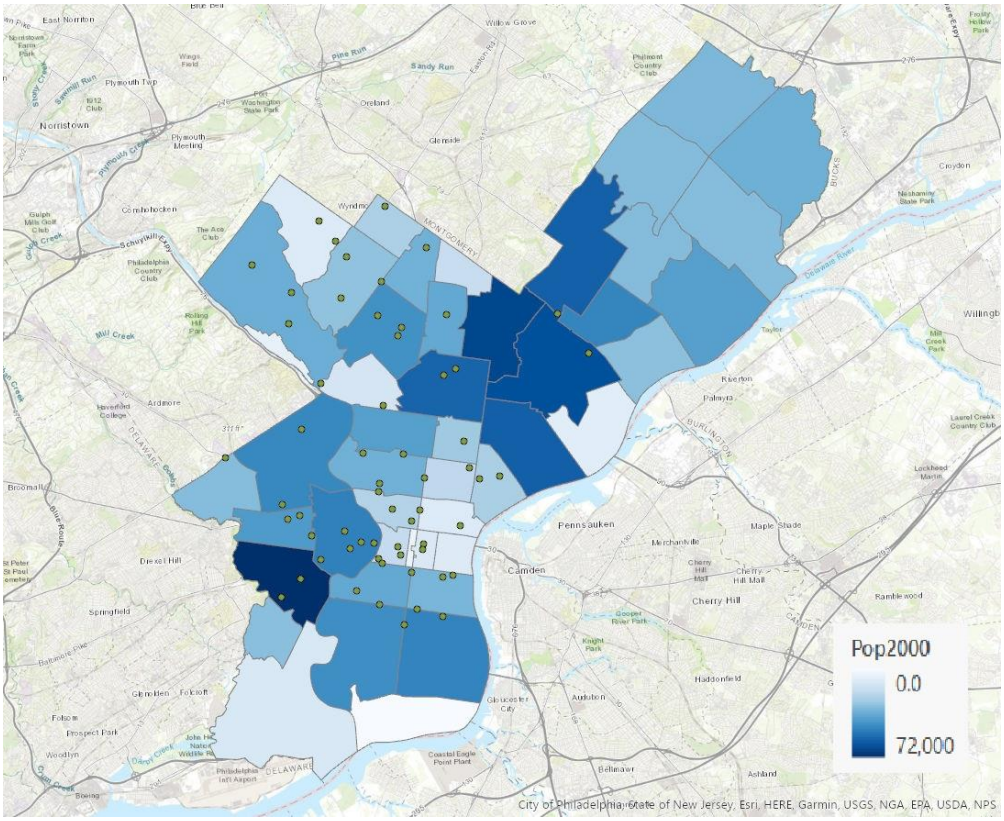

Figure 4. Farmers Markets over Population Distribution Map

4. Discussion

The results obtained from the initial Nearest Neighbor Analysis and K-function Analysis are not consistent with each other since the area boundaries are different. In the initial Nearest Neighbor Analysis, we use the minimum enclosing rectangle, which results in failing to reject H_0 that the farm markets are randomly distributed in Philadelphia. With manually correction of imputing areas of Philadelphia, we re-run the Nearest Neighbor Analysis and reject H_0 for H_a that the farm markets are significantly clustered in Philadelphia. The result from the second Nearest Neighbor Analysis is consistent with K-functions, which is that there is significant spatial clustering of the farm markets in Philadelphia.

They are consistent with my expectations based on the visual examination of the point data. The limitations of Nearest Neighbor Analysis include considering the average distance to only the nearest neighbor, depending on the minimum enclosing rectangle, and neglecting the fact that different scales will result in opposite identifications of clustering and dispersion. Within the minimum enclosing rectangle of the points, it reduces the real study area and results in wrong conclusion in the initial Nearest Neighbor Analysis. As we manually correct the study area to real areas of Philadelphia, the result from second Nearest Neighbor Analysis is consistent with

the right result which we got from the K-functions analysis. Even for K-functions analysis, it has limitation of not taking into consideration of social factors, such as population.

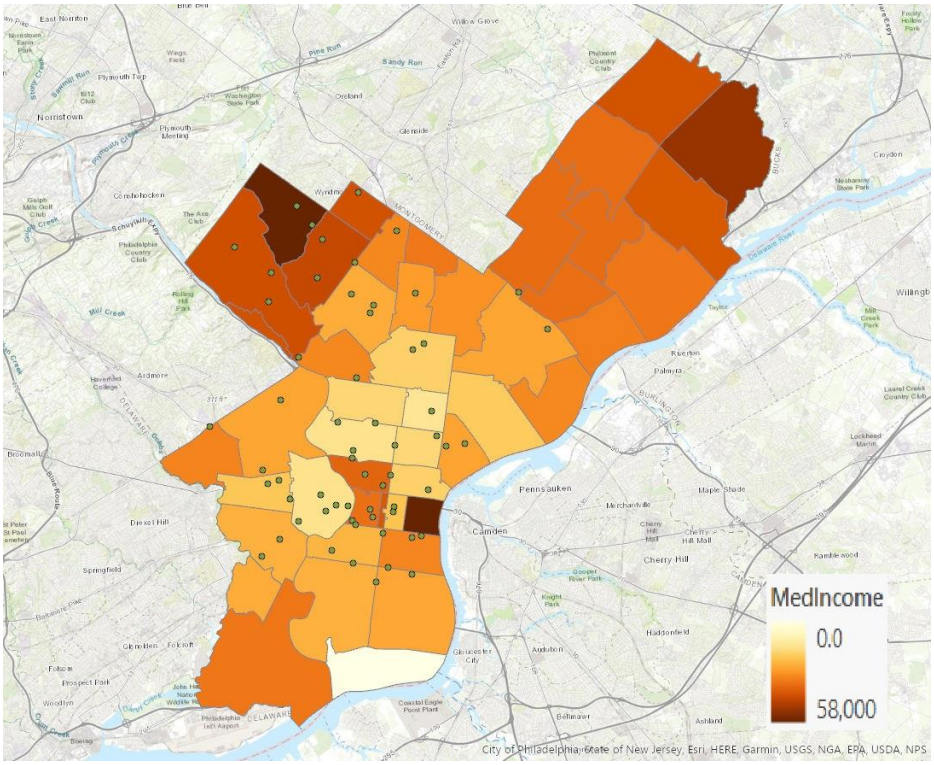

Figure 5. Farmers Markets over Median Household Income Map

According to figure 5, some areas, like North Philadelphia and South Philadelphia, with lower median incomes have fewer farmers markets. However, Northeast Philadelphia and lower west end of Philadelphia, like Eastwick, with medium to high median income also have few or even no farmers markets.

In conclusion, we think farmers markets are clustered based on Nearest Neighbor Analysis and K-functions Analysis. The cluster patterns occurred in some low-income areas, like West Philadelphia, as well as in some areas with medium to high income, such as center city Philadelphia. So, we can't say that farmer's markets are clustered in low-income areas. It doesn't make sense to use the minimum enclosing rectangle as the study area, which neglects the fact that North, Northeast, and South Philadelphia lacks farmers markets and results in wrong result whether it is a cluster or a random pattern. Based on the discussions above, median house income and population are not the only social factors that affect the distribution of farmers markets in Philadelphia. We should also consider other factors, like race, education level, zoning etc. Some races don't care too much if food is organic, fresh, or cultivated and freezing. People with low education may not pay attention to hold healthy eating habit. Neighborhoods near industrial zoning also lack farmer markets, like Northeast and lower west end of Philadelphia.

Another factor we didn't take into consideration is that we only focus on farmers markets within boundary of Philadelphia but neglect the ones outside of the boundary but close to the neighborhoods in boundary areas of the city. For instance, Glenside Farmers Market is closed to

Northeast Philadelphia, and Swarthmore Farmers Market is closed to lower west end of Philadelphia, but both farmers markets are located outside the city boundary. In addition, farmers market is not the only resource to provide fresh and healthy food. Most of large grocery stores, like Walmart and ShopRite provide healthy food and fill in the gaps of lacking farmers markets in the areas we discussed above.

Finally, we suggest government should consider comprehensive factors when make policy decisions. Based on the statistics analysis, we think farmers markets are clustered in Philly. Whether the government should introduce more farmers markets in the areas with few farmers markets, they need to analyze the local demand and supply, factors that result in lacking farmers markets, as well as other resources to provide heathy food other than farmers markets.

References

Ceadserv1.nku.edu. 2021. Nearest Neighbor Analysis. [online] Available at: <http://ceadserv1.nku.edu/longa/geomed/ppa/doc/NNA/NNA.htm> [Accessed 22 December 2021].

Desktop.arcgis.com. 2021. How Multi-Distance Spatial Cluster Analysis (Ripley's K-function) works—Help | ArcGIS Desktop. [online] Available at: <https://desktop.arcgis.com/en/arcmap/10.6/tools/spatial-statistics-toolbox/h-how-multi-distance-spatial-cluster-analysis-ripl.htm> [Accessed 22 December 2021].