The Spatial Distribution of Farmers Markets in Philadelphia

Shengao Yi 76852392, Xuening Zhang 65401665, Yu Wang 20166947

1 Introduction

Many American cities have struggled to provide access to healthy, locally grown food. In Philadelphia, the Philadelphia Food Trust has established extensive farmers markets across the city as a method of bringing fresh and healthy food to residents of the city's various neighborhoods. Various benefits were shown to the local people, such as the fresher foods, seasonal foods, and healthier foods are provided to the local shops and communities. However, not all areas of Philadelphia have access to farmers markets; for example, some areas in the South, North, and the whole Northeast Philadelphia do not have any farmers markets, and residents cannot benefit from the markets. To solve the problem, with the help of R and ArcGIS, this project would investigate the spatial distribution of farmers markets in Philadelphia to determine whether they are randomly distributed, dispersed, or clustered across Philadelphia in 2013. The project is based on the shapefile from the PA Spatial Data Access website (www.pasda.psu.edu) with locations of all farmers markets in Philadelphia for 2013, and the data will be analyzed using point pattern analysis - K-Function Analysis and Nearest Neighbor Analysis. Also, the results of regression and analysis from R and ArcGIS will be compared in this project.

2 Methods

2.1 Hypothesis Testing

The point pattern hypothesis test is the basis of the measurements of spatial distribution of farmers markets. The test is related to the Complete Spatial Randomness (CSR),

which explains a point process in which point events occur completely at random in a given study area. There are 2 conditions required to be hold to ensure the point pattern is completely spatially random. The first condition is that the positions of points are independent of one another with no influences to each other. The second condition is in a circumstance that the study area is divided into equal-sized cells or quadrats, the likelihood of a point happening in each cell is the same. The probability that a point is in a specific cell is determined solely by the proportion to the area.

Also, in the point pattern analyses, the hypotheses are as following:

H_0: The observed points are distributed randomly. The observed point pattern isn't different significantly from the expected point pattern.

H_a: The observed points are not distributed randomly. The observed point pattern is distributing in a cluster or dispersion way, which differs significantly from the expected point pattern.

2.2 Quadrat Method

Quadrat method represents an approach to generate the density of points while counting the number of points in each unit area. The quadrat method tends to divide the study area into equal-sized cells or quadrats to count the number of points within every cell or quadrat to assess the distribution of point events in a given area. The variance value, mean value, and VMR value of the points of each quadrat or cell can be used to derive the point pattern. The point pattern is random in a condition of VMR close to 1. The point pattern is uniform in a condition of VMR close to 0. The point pattern tends to be clustered in a condition of VMR is much greater than 1.

However, there are some limitations of the quadrat method:

a) When two sets of points are in the same distribution pattern, the results may vary depending on the size of the cell (keeping the range constant) since the number of points in each cell will change.
b) Also, in the same case, the derived results could be different depend on the size of range (holding cell size constant) for the reason that the total amount of quadrats alters, therefore, the mean value and the variance value change.
c) If the distribution patterns of two sets of point are different, the derived results are depend on the size of both the range and cell size.

Because of the limitations above, the quadrat method is not generally used in practice.

2.3 Nearest Neighbor Analysis Method

To address the problem that the mentioned limitations of quadrat method, we can introduce the Nearest Neighbor Analysis Method. The Nearest Neighbor Index (NNI) is a tool to measure the spatial distribution, to see if it is clustered or dispersal, which can be represented by the ratio of the observed average distance between each point and its nearest neighbor divided by expected average distance. The expected average distance is the average distance between each point and its nearest neighbor in a hypothetical random distribution. The Nearest Neighbor Index can be calculated as:

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

$$\overline{D}_O = \frac{\sum_{i=1}^{n} D_i}{n}$$

Here, $\overline{D}_O$ is the observed mean distance between each feature and its nearest neighbor, D_i is the distance between each feature i and its nearest neighbor, sum them up and divided by the number of features, n. $\overline{D}_E$ can be calculated by the following equation:

$$\overline{D}_E = \frac{0.5}{\sqrt{n/A}}$$

n is the number of features, A is the area of minimum enclosing rectangle, which is the

smallest rectangle that can contain all the points.

NNI can be interpreted by the following guidelines:

- When NNI is close to 1, which means the observed average distance is equal to expected average distance when pattern is random. The points are in random pattern.
- When NNI is close to 0, which means the observed average distance is close to 0, all the points are located close to the same spot. The points are in a clustered pattern.
- When NNI is close to 2, or greater than 2 but less than at most 2.149, which means the observed average distance is much larger than the expected average distance then there is a random pattern. The points are in a dispersed pattern.

There are two hypotheses:

- H_0: The observed point pattern is random.
- H_a: The observed point pattern is NOT random, either significant clustering or dispersion.

We use the average nearest neighbor Z-score to test the two hypotheses, which is calculated by the following equation:

$$z = \frac{\overline{D}_O - \overline{D}_E}{SE_{\overline{D}_O}}$$

Where, $SE_{\overline{D}_O}$ is the standard error of $\overline{D}_O$:

$$SE = \sqrt{\left(\frac{1}{4\,tan^{-1}1} - \frac{1}{4}\right)\frac{A}{n^2}} = \frac{0.26136}{\sqrt{\frac{n^2}{A}}}$$

Thus, the z-score is:

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

We can use the standard normal table to get a p-value from calculated z-score. Since the H_a states that $\overline{D}_O \neq \overline{D}_E$, not like a single dimension comparison < or >. Thus, a two-tailed test is used here. In a two-tailed test, $z = |1.96|$ corresponds to an α-value of 0.05.

If z > 1.96 or z < -1.96, which means p < 0.05, we can reject H_0 for H_a at the level of α = 0.05; and if z > 1.96, we have a significant dispersion pattern; if z < -1.96, the pattern is significant clustering.

The advantage of NNI method is that it only uses the distances between points. However, there are still some limitations in NNI method:

a) It only takes into account the only one nearest neighbor. There are some bias would happen if some special patterns exist. For example, if points are clustered in several pairs, but each pair is different with others.
b) The value of A, the area of the study region, has a great influence on NNI. One way to calculate A is through minimum enclosing rectangle, which would be strongly affected by the outliers; Another way is convex hull, the smallest polygon that encloses all the points.
c) NNI can't identify patterns where both clustering and dispersion are present at different scales.

Compared to the hospital example in lecture slides, this project has fewer points and the farmer's markets just locate in the area of Philadelphia rather than the state of Pennsylvania. The hospitals are clustered in smaller area, but dispersion in a larger scale, as a result of which, the NNI result will change significantly since the scale of study area.

2.4 K-Functions Analysis method

Besides the methods that we have mentioned above, there is another method that exists

for detecting a special point pattern that is clustering at a small scale and dispersion at a large scale. This method is called Ripley's K function, and it is useful when the neighborhood size (scale) changes. Moreover, the steps of performing K-Functions analysis are the following: first select a range of different values of radius (d) of the circles, and the rest of the steps will be repeated for each radius.

1. Place a circle with radius equals to d around every point (farmers market in the study area)
2. Count the number of points inside each circle.
3. Calculate the average number of farmers markets across all the circles with radius d.
4. Divide the average number of other points by farmers markets density in the study region, which is Philadelphia in this case.

The density is calculated by the total number of points divided by the area of the study area. The steps above are called the K(d) function for the specific radius d, and can be calculated by the formula below:

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

The function can be summarized as the mean number of points identified within radius d divided by farmers markets density of Philadelphia. When we have Complete Spatial Randomness (CSR), K(d) equals to the area of a circle with radius d. If there is a clustering or dispersion, the K(d) would change accordingly.

$$\begin{cases} K(d) = \pi d^2, CSR \\ K(d) < \pi d^2, Clustering \\ K(d) > \pi d^2, Dispersion \end{cases}$$

Since many software packages return a result based on the L(d) functions instead of the

K(d) functions, which is a transformed version of K(d) function. Under CSR, the value of L(d) is 0. The common form of the L(d) function is:

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

For this assignment, ArcGIS or R are the two potential tools that we are going to use, and the form of L(d) functions calculated by the Ripley's K function tool in ArcGIS will be slightly different than the common form. Under CSR, the value of L(d) is d. The formula is as follows:

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

When the K-functions are being calculated, there is one more parameter that we need to provide to the tools that we use, which is beginning and incremental distances d. The parameter can be calculated by the formula below:

$$d = \frac{0.5 * \ maximum\ pairwise\ distance}{number\ of\ distance\ bands}$$

The widely used number of distance bands are 10 or 20. The maximum pairwise distance is defined as the distance between the two farthest apart points in the study region.

The null hypothesis is that at distance, the pattern is random. Hypothesis a1 (H_{a1}) is that at distance d, the pattern is clustered. Hypothesis a2 (H_{a2}) is that at distance d, the pattern is uniform. To test the hypothesis, many (e.g., 9, 99, or 999) random point patterns are generated and the number of points is same as the original data. In this project, we will use software to perform L(d) functions 100 times, the original Philadelphia farmers market data and the 99 random patterns created, and when the function is being calculated for each radius d, we will observe the lowest value of L(d) and highest value of L(d) for each radius d.

The lowest value of L(d) for a specific radius d is called the lower Envelope $L^{-}(d)$, and the highest value of L(d) for a specific radius d is called the upper Envelope $L^{+}(d)$. Once we have the lower and upper bounds of the L(d) functions for each radius d, we can form a confidence envelope for that radius. Confidence envelopes are a critical technique that we will be using to test whether we can reject the null hypothesis. For each radius d, we can compare the original $\mathrm{L(d)}^{obs}$ to $L^{+}(d)$ and $L^{-}(d)$.

If $\mathrm{L(d)}^{obs} > L^{+}(d)$, we can reject null hypothesis at distance d for H_{a1} and conclude that there is significant clustering in the distribution of farmers markets at scale d. If $\mathrm{L(d)}^{obs} < L^{-}(d)$, we will reject null hypothesis at distance d for H_{a2} and conclude that there is significant dispersion in the distribution of farmers markets at scale d. If $L^{-}(d) < \mathrm{L(d)}^{obs} < L^{+}(d)$, then we can't reject null hypothesis at distance d. That is, we cannot say that at distance d, the pattern is significantly different from what we'd expect under CSR.

However, if a point located very close to the border of our study area, and some area of the circle will be outside of study area, in which case it will lead to inaccuracy since there may be some points in the outsides aera. Thus, we would require Ripley's Edge Corrections and Simulate Outer Boundary Values Edge Correction. Ripley's Edge Correction adds extra weight to the points that are inside the rectangular region compared to the ones that are located on the border. If the entire circle around a specific market is within the boundary of the rectangular region, Ripley's Edge Correction function in ArcGIS will put a weight of 1 to this market, and less weight on the markets that have its circle partially outside of the boundary. On the other hand, Simulate Outer Boundary Values edge correction method will be the method that we are going to use for this assignment because it is more suitable for irregular study region and Ripley's Edge Correction works only for rectangular study regions.

In some cases where it's not reasonable to assume that points are randomly distributed during permutations, taking into account a reference measure in the K-function analysis

may be helpful, which is called Non-homogeneous K-functions. The problem can be solved by generating point patterns that are influenced by the population of Philadelphia. To create this kind of point patterns, we will use the *Create Spatially Balanced Points* function in ArcGIS, and the function will be more likely to generate points around areas that have high population based on the population layer provided. The steps to generate the desired random patterns are the following:

1. Create a new column called population probabilities in the attribute table and getting its values by using the population of the county divided by the total population across entire Philadelphia region.
2. Use the *Polygon to Raster* function to convert the shapefile containing population probabilities values into a raster layer.
3. In *Create Spatially balanced Points* function window, we will use the raster layer created from the previous step as the input probability raster and enter the number of output points which is same as the number of farmers markets in Philadelphia 62, and then get the point pattern.

Once the steps mentioned above are repeated 99 times to generate enough random patterns, we can use these patterns to calculate L(d) functions and develop confidence envelopes as well. The reason why it is necessary to consider the population of each county when generating point patterns is that markets' locations will be generated at rural areas that have very few populations, and it is unreasonable for farmers markets to be hosted at those kinds of locations.

3 Results

3.1 Nearest Neighbor Results

The nearest neighbor results from ArcGIS and R are shown in Table 1 and Table 2. We have used two different area to conduct nearest neighbor analysis, one is the area of

minimum bounding box, another is the area of Philadelphia.

Table 1. ArcGIS Nearest Neighbor Results

Study area	Expected Mean Distance (feet)	Observed Mean Distance (feet)	z-score	p-value	NNI
Minimum bounding box	3112.9097	3112.9097	-0.069945	0.944237	0.995357
Philadelphia	4001.3504	3112.9097	-3.344634	0.000824	0.777965

Table 2. R Nearest Neighbor Results

Study area	Expected Mean Distance (feet)	Standard Error	z-score	p-value	NNI
Minimum bounding box	3406.1688	226.12	-1.796058	0.036242	0.880768
Philadelphia	4001.3504	265.63	-3.344634	0.000412	0.777965

As for the area of minimum bounding box, the nearest neighbor index is just less than 1, the index in ArcGIS is 0.995 and 0.881 in R. Since the NNI is close to 1, so the farm markets are random in this area. In addition, the z-scores are -0.069945 and -1.796058 from ArcGIS and R respectively, which are between -1.96 and 1.96 (at 0.95 confidence level), indicating that we fail to reject the Null Hypothesis that the locations are random in this area. The different results in the two methods suggest that the calculation method may be different. From the table, we can find that the expected mean distance values are not the same, which is calculated using the area and number of observations, the distance in R is about 300 feet larger than that in ArcGIS.

On the other hand, as for the area of Philadelphia, the NNI values (0.777965) and z-scores (-3.344634) are the same in both ArcGIS and R, which indicate more clustering. The z-score is outside of the range between -1.96 and 1.96. Thus, we can reject the Null Hypothesis in favor of the alternate hypothesis that the locations of farmers markets are not random. The p-value suggests that the clustering is significant.

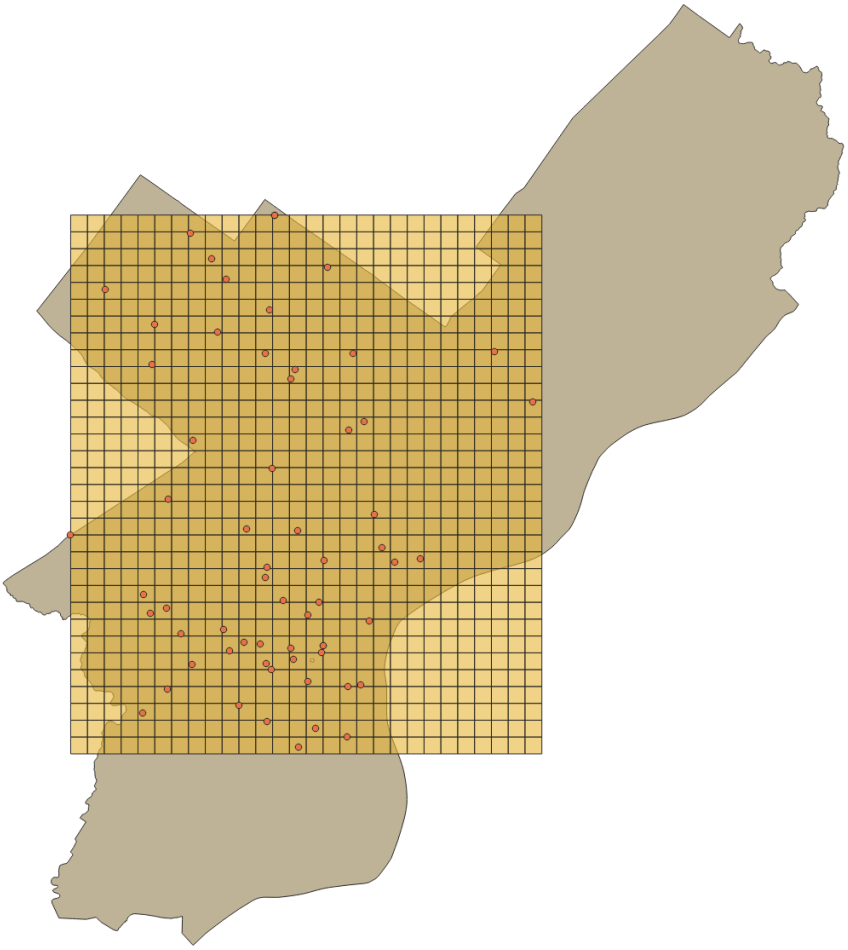

Figure 1. Outlines of Minimum Bounding Box and Philadelphia

Possible reasons for the different results may lie in the fact that the area of Philadelphia is larger than the area of minimum bounding box. Figure 1 shows farmers markets locations and the outlines of Minimum Bounding Box and Philadelphia.

3.2 K-function Analysis Results

To get the K-function results, we chose to use 0 as beginning distance and 2,500 feet as distance increment, and the reason we wanted to use 2,500 feet increment is because that the maximum pairwise distance that we have measured is about 50,000 feet and we chose 10 distance bands, and we got our results by using the equation for beginning and incremental distance:

$$d = \frac{0.5 * 50{,}000}{10}$$

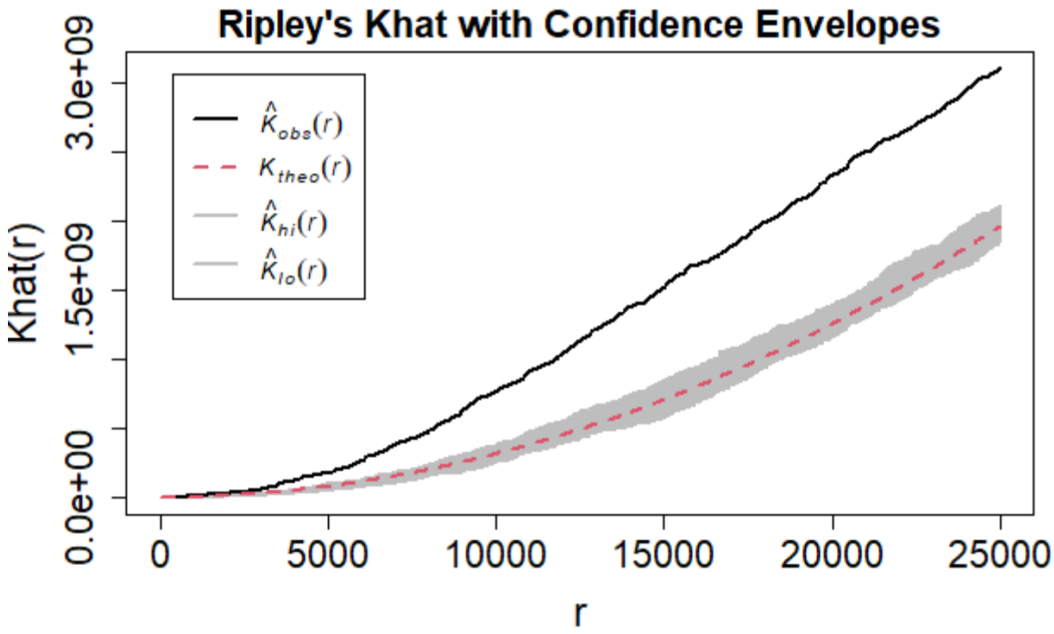

Figure 2. $\widehat{K}$ Plot

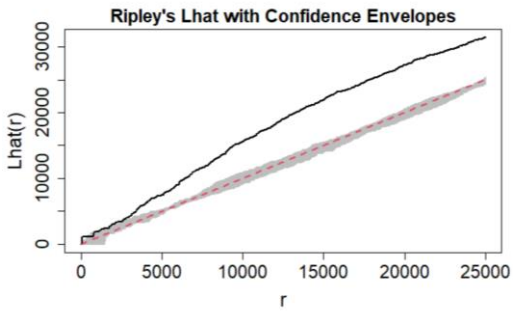

Figure 3. $\hat{L}$ Plot

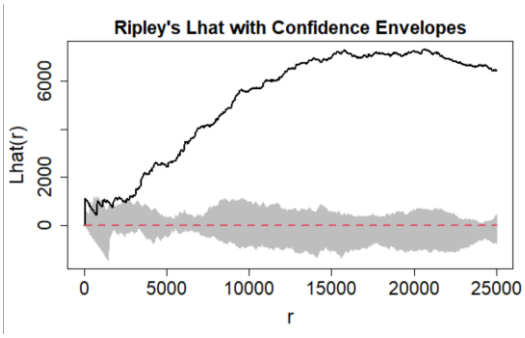

Figure 4. Rotated $\hat{L}$ Plot

We first run the K-function in R, the plots of L(d) functions showed a similar result as the K(d) functions, which the bold observed line is beyond the confidence enveloped shown in grey. By examining the observed K(d) functions, L(d) functions and their confidence envelope, we found out that at about 2,500 feet, the observed K(d) function is clearly beyond the upper limit of the confidence envelope. Thus, we concluded that the farmers markets in Philadelphia are statistically significant clustering at larger distances, and the null hypothesis was rejected by this finding.

Table 3. ArcGIS K-function results

	OBJECTID *	ExpectedK	ObservedK	DiffK	LwConfEnv	HiConfEnv
▶	1	2500	3701.59289	1201.59289	2238.939475	3270.180414
	2	5000	7691.009666	2691.009666	4971.087134	6063.07676
	3	7500	11959.653809	4459.653809	7380.58058	9613.372138
	4	10000	15863.32464	5863.32464	9656.034874	11149.828668
	5	12500	19372.541598	6872.541598	12652.154328	13789.637739
	6	15000	22737.45648	7737.45648	15327.608056	16258.668105
	7	17500	25930.489319	8430.489319	17723.955492	18892.191987
	8	20000	28846.717239	8846.717239	20117.258456	21153.73271
	9	22500	31334.48922	8834.48922	22254.652936	23453.714781
	10	25000	33454.481074	8454.481074	24190.218825	25684.451647

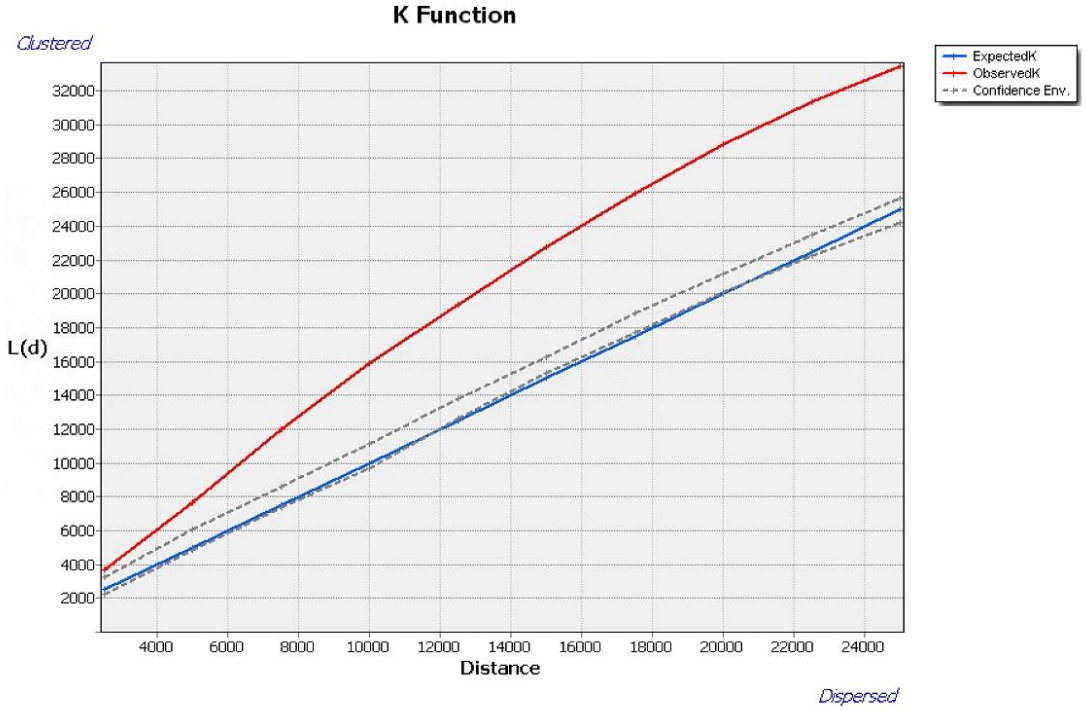

Figure 5. ArcGIS K-Function Plot

The results from ArcGIS are shown in Table 3. and Figure 5. We can know that the Observed K is greater than the Expected K in all increments and the red line is beyond the confidence envelope all the distances. Thus, we can reject null hypothesis and in favor of the first alternative hypothesis that there is clustering for all distances. Compared with the results from R, there is a little different.

Without any further analysis on relationship between the site selections of farmers markets and population of Philadelphia, we believed it is reasonable to assume the absence of farmers markets in Northeast Philadelphia and South Philadelphia is related

to the populations in these areas. Normally, Farmers markets appear in populated areas, where a lot of people would be attracted by the fresh and locally made food products. Therefore, if nonhomogeneous K-Functions were performed using population as a reference measure, we expected to see non-randomly distributed farmers market locations still because farmers markets would be most likely appeared in populated areas in Philadelphia; however, it is hard to make a conclusion on whether the location pattern is clustered or dispersed without further analysis.

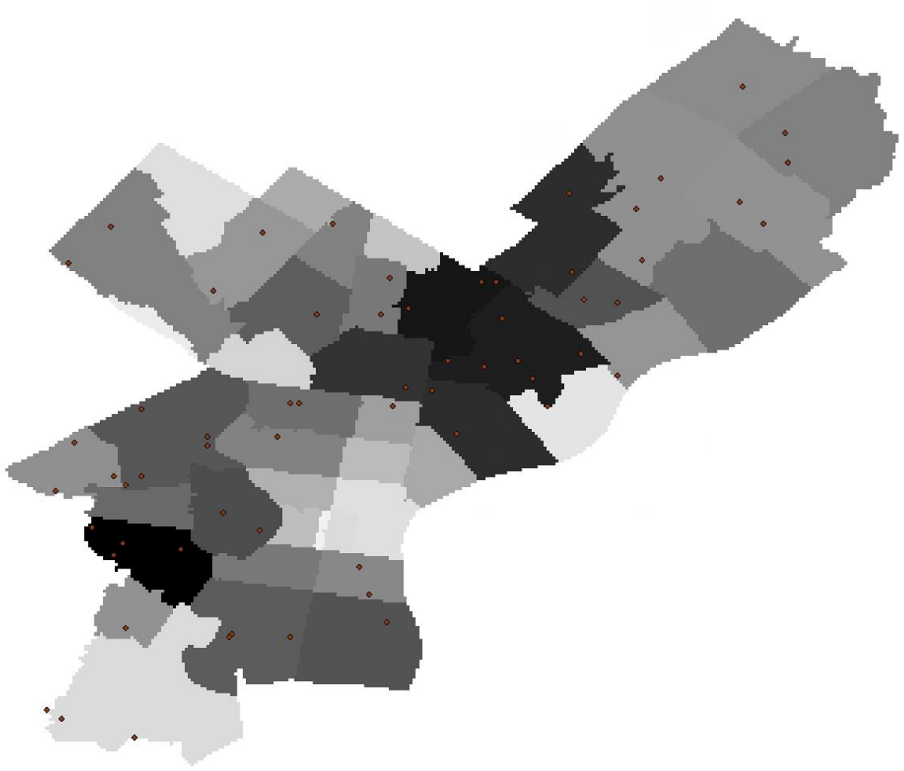

Figure 6. Spatially balanced points result

According to the steps mentioned above, we have created spatially balanced points based on the population. The result is shown in Figure 6. It seems to be dispersed. Then run K-function in ArcGIS and adjust the distance increment to 5,000 feet, the results are shown in Figure 7, which suggest that the newly made farmers markets points based on population density are statistically significant dispersion at all distances.

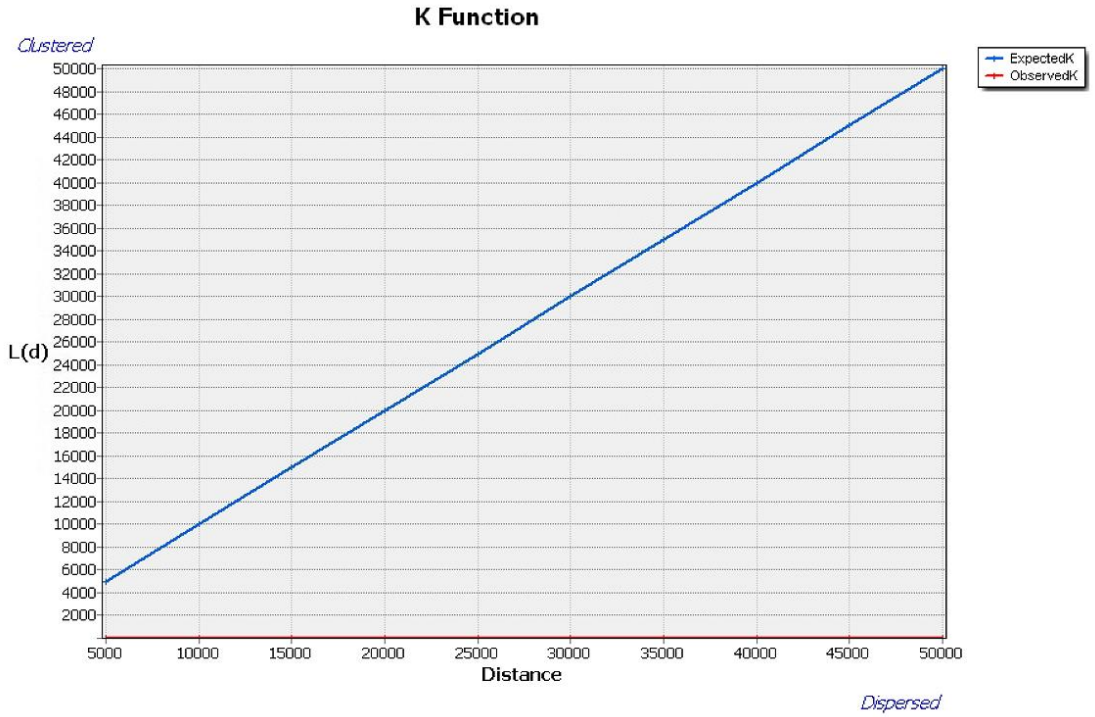

Figure 7. ArcGIS K-Function Plot

4 Discussion

It is represented by the consistent of both K-Function analyses and Nearest Neighbor analyses that the distribution pattern of the farmer markets in Philadelphia are cluster, which is also consistent with the visual examination in Figure 1 as expected. Therefore, the farmers markets of Philadelphia distribute in a relatively cluster way in the city center area, there tends to be some small clusters of markets in 2 or 3 in neighboring lattice cells.

The results of Nearest Neighbor analysis of R and ArcGIS in a minimum enclosing rectangle suggest the positions of farmer markets are distributing in a random way, which is on the contrary to the results of K-Function with smallest intervals that suggest the occurrence of small clusters. The difference between the results indicates the 2 different methods of analyzing the point patterns. In this case, the minimum enclosing rectangle is less relevant than the whole limit of the city because the scale of the problem is related to the entire city. In this circumstance, it could be derived that the farmers markets are distributing in a cluster way in Philadelphia based on both the

results from the nearest neighbor analysis of the entire city, and the results from the K-Function analysis.

The results from ArcGIS and R are relatively consistent, but there are still some small differences between the results. The results of the nearest neighbor of the minimum enclosing rectangle are slightly different, which suggest the same conclusion about the distribution pattern of the farmer markets. It is indicated by the inconsistencies between the results of R and ArcGIS that they used different minimum enclosing rectangle regions. In contrary, the results from the nearest neighbor analysis of Philadelphia and the results from the K-Function are consistent.

Considering the fact that the locations of farmer markets are not chosen in a random way, it is possible to contribute variable to a distribution pattern of clustering, possibly with populations and median household income. The Figure 8 below represents the distribution median household income in Philadelphia with post codes and the distribution of the farmer markets.

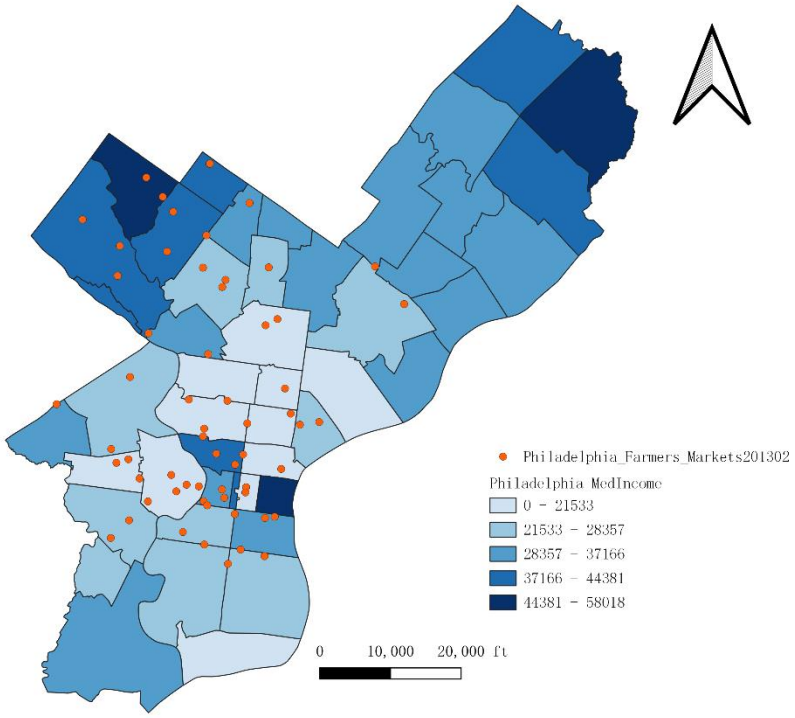

Figure 8. Philadelphia Median Income and Farmers Markets

The plot of farmers markets and the median household income of Philadelphia above represent that the locations of several farmers markets are in the north part of city center area with almost the lowest groups of income, which does not support the opinion that the farmers markets tend to only distribute in more wealthy areas of the city. Also, there are farmers markets in the regions with the highest groups of income in Philadelphia such as the northwest area of Philadelphia. However, there is no farmers markets in the northeast and south Philadelphia.

Based on the derived results, the farmers markets are clustering in Philadelphia. However, there is no strong relationship between the income and the cluster of markets in Philadelphia. Therefore, it is challenging for the city to provide relatively equal chances to residents for accessing the fresh and healthy productions in farmers markets. And to help the organizers to establish farmer markets in the northeast and south parts of Philadelphia, the city could provide tax deduction policy to the market organizers to solve the problem.