Zhenzhao Xu, Hanpu Yao, and Ben Aiken
Statistical and Data Mining Methods - MUSA 500
Professor Eugene Brusilovskiy
Assignment 4

The Spatial Distribution of Farmers Markets in Philadelphia

Introduction

Many people prioritize healthy, locally grown food these days. Reasons include reducing driving and parking demand, fresher and healthier foods, and providing a place to meet neighbors.[1] One way to get locally grown food is from farmers markets throughout Philadelphia set up by the Philadelphia Food Trust. However, according to the Philadelphia Food Trust[2], improving healthy food access in low-income communities and communities of color continues to be an urgent need. Parts of South Philadelphia and North Philadelphia, and essentially all of Northeast Philadelphia have no farmers markets at all.

This report examines whether the farmers markets are randomly placed, dispersed, or clustered throughout Philadelphia in 2013. Data, provided by the PA Spatial Data Access website (www.pasda.psu.edu), will be analyzed with several types of point pattern analyses including Nearest Neighbor Analysis and K-Function Analysis to determine the spatial relationship among farmers markets.

The regressions and analysis will be completed in both ArcGIS and R. A comparison between the results is also included in the report.

Methods

a) Hypothesis Testing

The basis of the spatial distribution measurements of Farmers' Markets is the point pattern hypothesis test. These tests relate to Complete Spatial Randomness or CSR. CSR means all point events in a given area occur completely randomly. Specifically, there are two needed conditions to determine if a point process is completely spatially random:

1. When the study area is subdivided into cells or quadrats of equal size, the probability of a point occurring is the same in all cells. The probability that a point is in a certain cell only depends on the proportion to the area.
2. The locations of points are independent and have no effect on others.

In point pattern analyses, the hypotheses are:

[1] Farmers' market, https://www.canr.msu.edu/news/public_markets_differ_from_farmers_markets
[2] The Food Trust, http://thefoodtrust.org/food-access

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

H_a: The observed points are not randomly distributed in space. The observed point pattern is cluster or dispersion, significantly different from the expected point pattern.

b) Quadrat Method

The quadrat method is a way to measure point density when counting the number of points per unit area. To determine how point events are distributed in a particular space the quadrat method subdivides the study area into cells or quadrats of equal size and counts the number of points in each quadrat. The mean, variance and VMR (Variance/mean) of the number of points in each quadrat is calculated to determine what point pattern it is. If VMR is close to 1, it is random; if VMR is closed to 0, it is uniform; if VMR is much greater than 1, it is more clustered.

However, the quadrat method has several limitations:

a. When two groups of points are in the exact same pattern, the results may vary based on the cell size (holding range constant). This is because the number of points in each cell changes.
b. When two groups of points are in the exact same pattern, the results may vary based on the range size (holding cell size constant). This is because the total number of cells changes, so the variance and mean also change.
c. When two groups of points are in different patterns, the VMR and results can be identical depending on the range and cell size.

These are three possible situations that can present inconsistent points pattern and quadrat analysis results. As a result of these limitations, the Quadrat method is rarely used in practice.

c) Nearest Neighbor Analysis

The nearest neighbor index (NNI) is a measure of clustering or dispersal with more consistent results than the Quadrat method. NNI is expressed as the ratio of the observed average distance between each point and its nearest neighbor divided by expected average distance. The expected distance is the average distance between each point and its nearest neighbor in a hypothetical random distribution. The nearest neighbor index is 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, *D₀* is the observed mean distance between each feature and its nearest neighbor, D_i is the sum of distances between point *i* and its nearest neighbor, *n* is the number of points. $\overline{D}_E$ is calculated using the following formula.

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

n is the number of features. *A* is the area of the minimum enclosing rectangle, which is the smallest rectangle around the points.

NNI is interpreted using the following guidelines:

a. When NNI is close to 1, the points are in a random pattern, because the observed average distance = expected average distance in random pattern.
b. When NNI is close to 0, the points are in a clustered pattern. This means the observed average distance is close to 0, so all the points are close to the same spot.
c. When NNI is close to 2 or greater than 2 and less than maximum 2.149, the points pattern is dispersed since the observed average distance is much larger than the expected average distance of random pattern.

The hypothesis testing method of NNI uses the Z-score, which is calculated. The hypotheses are:

H_0: the observed point pattern is random
H_a: the observed point pattern is not random, either significant clustering or dispersion.

The average nearest neighbor z-score for the statistic is calculated using the equation below:

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

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

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

Therefore, the z-score is:

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

A p-value is calculated by comparing the z score in a standard normal table. Since the H_a compares the NNI to some given value (0,1,2), it does not rely on a single dimension comparison like < or >. Instead, a two-tailed test is used here.

If $z > 1.96$ or $z < -1.96$, meaning $p < 0.05$, the hypothesis test at the 0.05 α level will reject H_0 for H_a. And if $z > 1.96$, the pattern is significant dispersion. If $z < -1.96$, the pattern is significant clustering.

However, there are also some limitations in NNI method.

a. It calculates the distance to only one nearest neighbor. This may bring some bias when there are some special patterns in points. For example, points are very close in pairs, but each pair is far away from each other.
b. The value of A, the area of study, has great impact in NNI. The way it is defined makes a significant difference to NNI results. One way is smallest rectangle around all the points which will be affected greatly by outliers. Another way is convex hull, the smallest polygon that encloses all the points.
c. NNI cannot identify patterns where both clustering and dispersion are present at the same time at different scales.

Compared to the hospital case in lecture slides, the Farmer's Market investigation case has fewer points, and they are only distributed in Philadelphia, rather than the entire state of Pennsylvania, used for the hospitals. The hospitals appear to be clustered in smaller scale, likely in areas of high population density, but dispersion or perhaps randomness on a larger scale. As a result, the NNI will change greatly with the scale of the study area. This may also be the case for the Farmers' Markets if they are indeed clustered on a small scale but dispersed in the city as a whole.

d) K-Function Analysis

It is common that both clustering and dispersion are present at different scales. However, the methods above cannot take this pattern into account. Ripley's K-Functions Analysis is introduced to detect patterns at different scales. K(d) function can be calculated by the formula below.

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

To calculate K(d), the number of other points in a given distance d is counted and averaged. K(d) is the quotient of the mean number and the point density. Meaning that if there are more points in the circle with radius d, K(d) will be bigger. Same steps are repeated to get the results at different distances.

Under CSR, K(d) equals to the area of a circle with radius d. If there is clustering or dispersion, the K(d) will change accordingly.

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

L(d) is introduced for ease of interpretation. It can be calculated by the formula below. Under CSR, the value of L(d) is 0, and if there is clustering or dispersion, L(d) will be greater or less than 0, so that the results is more interpretable.

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

However, ArcGIS uses a slightly different definition of L(d). Under CSR, L(d) in ArcGIS is d instead of 0.

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

Generally, the beginning and incremental distances d are set as follows. The number of bands are often 10 or 20. Maximum pairwise distance is the distance between 2 farthest points.

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

The null hypothesis is that at distance d, the pattern is random. Hypothesis a is that at distance d, the pattern is clustered. Hypothesis b is that at distance d, the pattern is uniform. To test the hypothesis, 9 or 99 or 999 patterns are generated, and the number of points is the same as the original dataset. In this project, 99 permutations are generated. Then, L(d) is calculated for each pattern, and the highest and lowest L(d) is recorded, denoted as $L^{+}(d)$ (*Lower Envelope*) and $L^{-}(d)$ (*Upper Envelope*).

For each distance d, we can then compare the original $L^{obs}(d)$ to $L^{+}(d)$ and $L^{-}(d)$. 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 expect under CSR. If $L^{obs}(d) < L^{-}(d)$, we can reject H_0 for H_1, so there is significant clustering at scale d; if $L^{obs}(d) > L^{+}(d)$, we can reject H_0 for H_2, so there is significant dispersion at scale d.

However, if a point is close to the border, some area of the circle with radius d will be outside of research area, which will lead to inaccuracy because there may be some points potentially in this outside area. Thus, Ripley's Edge Correction and Simulate Outer Boundary Values Edge Correction are introduced to counteract the border effect.

Ripley's Edge Correction takes neighboring points which are inside the study area and gives extra weight. For example, if a circle is fully in the study area, the weight of this point will be 1; if 50% of a circle is in the study area, its weight will be 0.5. The Simulate Outer Boundary Values Edge Correction mirrors points across the study area boundary to correct for underestimates near edges.

In this project, Simulate Outer Boundary Values Edge Correction is chosen because Ripley's Edge Correction works only for rectangular study regions.

In some situations, it is not reasonable to assume that points are randomly distributed in space during the permutations, here a reference measure may be helpful to the analysis. For example, if one wants to test whether the nest of squirrels is clustered in the park, they may want to know the distribution of trees in the park. The solution is called nonhomogeneous K-Functions. To do a nonhomogeneous K-Function, the reference measure (usually by polygons) is first transformed into probability by dividing its maximum value. Then, instead of generating points randomly, the points are generated according to the probability. And for each distance d, $L^{+}(d)$, $L^{+}(d)$, and $L^{obs}(d)$ are calculated, and the hypothesis is tested, which is the same as regular K-Functions.

Results

a) Nearest Neighbor Analysis

The nearest neighbor results from ArcGIS and R are displayed in Tables 1 and 2.

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	3127.4314	-0.069945	0.944237	0.995357
Area of Philadelphia	4001.3504	3112.9097	-3.344634	0.000924	0.777965

In both programs the nearest neighbor was first calculated using the area of a minimum bounding box. However, for the second area, the programs measure the nearest neighbor for the area of Philadelphia.

Table 2 – R Nearest Neighbor Results

Study Area	Expected Mean Distance (feet)	Standard Error	z-score	p-value	NNI

Minimum bounding box	3353.28	222.61	-1.080	0.140	0.928
Philadelphia	4001.35	265.63	-3.345	0.000	0.778

When the minimum bounding box is used, the nearest neighbor index is just below 1. The index is 0.995 in ArcGIS and 0.928 in R. Since these values are closer to 1 than 0 or 2, the farmers markets are random within this area. Furthermore, the z-scores, -0.069945 and -1.080, from ArcGIS and R respectively, are between -1.96 and 1.96 (at 0.95 confidence level), failing to reject the null hypothesis meaning the locations are random in this area. The differences in the results of the two programs suggest that the nearest neighbor index may be calculated differently. At the very least, it seems that a different minimum bounding box is used because the expected mean distance values are not the same, which is calculated using the area and number of observations (which would not change by program).

On the other hand, when the area of Philadelphia is used to calculate these same statistics, the results indicate more clustering. The z-score of -3.45 (consistent in both programs) is outside of the -1.96 to 1.96 range that rejects the null hypothesis in favor of the alternate hypothesis proving that the results are not random. Since the nearest neighbor index is closer to 0 than 2 the locations are clustered. The p-value indicates that this clustering is, in fact, significant.

The difference in results from minimum bounding box and Philadelphia area are likely due to the fact that the area of Philadelphia is considerably larger than the minimum bounding box.

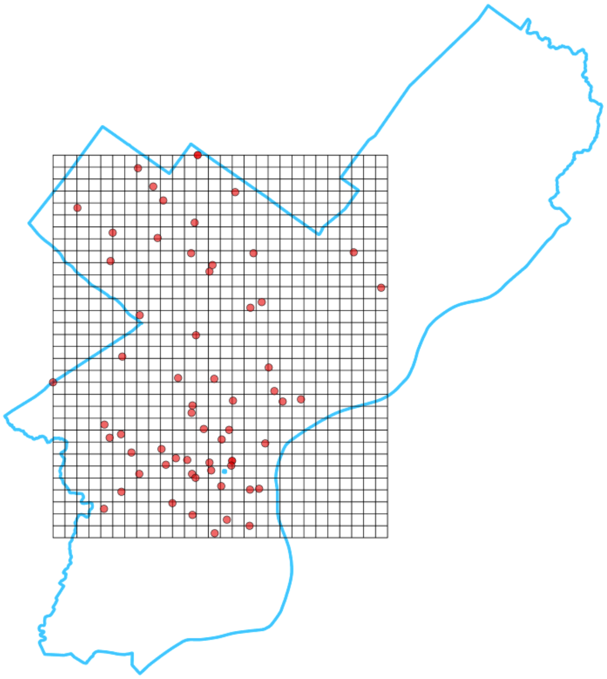

Figure 1 – Minimum Bounding Box and Philadelphia Outline

Figure 1 shows the farmers markets and minimum bounding box lattice with the outline of Philadelphia. The areas of Philadelphia that the bounding box does not cover are relatively large.

b) K-function Analysis

The k-function results from ArcGIS are displayed in Table 3. The beginning and incremental distances were calculated using the following formula:

$$distance = \frac{\frac{1}{2} * maximum\ distance}{\#\ of\ distance\ bands}$$

In this case, the maximum distance was 50,000 feet and 10 bands were used, so each band was set to increase by 2,500 feet.

Table 3 – ArcGIS K-Function Results

OBJECTID	L(d)	ExpectedK	ObservedK	DiffK	LwConfEnv	HiConfEnv
1	2500	2500	3701.593	1201.593	1917.313	3834.626
2	5000	5000	7691.01	2691.01	4551.898	6666.879
3	7500	7500	11959.65	4459.654	7403.186	9481.409
4	10000	10000	15863.32	5863.325	9707.81	12153.68
5	12500	12500	19372.54	6872.542	12084.74	14601.82
6	15000	15000	22737.46	7737.456	14578.92	17352.38
7	17500	17500	25930.49	8430.489	17031.64	20034.03
8	20000	20000	28846.72	8846.717	19312.07	22560.39
9	22500	22500	31334.49	8834.489	21397.2	24958.57
10	25000	25000	33454.48	8454.481	23073.03	27237.87

Table 3 shows that the Observed K is greater than the Expected K in all increments and, with the exception of the first distance, all are outside of the confidence envelope; therefore, the null hypothesis is rejected in favor of the first alternative hypothesis for all distances: there is clustering. This is reflected in Figure 2, which shows the Observed K line above Expected K. In fact, the space between the lines increases as the distance increases, suggesting that the extent of clustering increases as the distance increases.

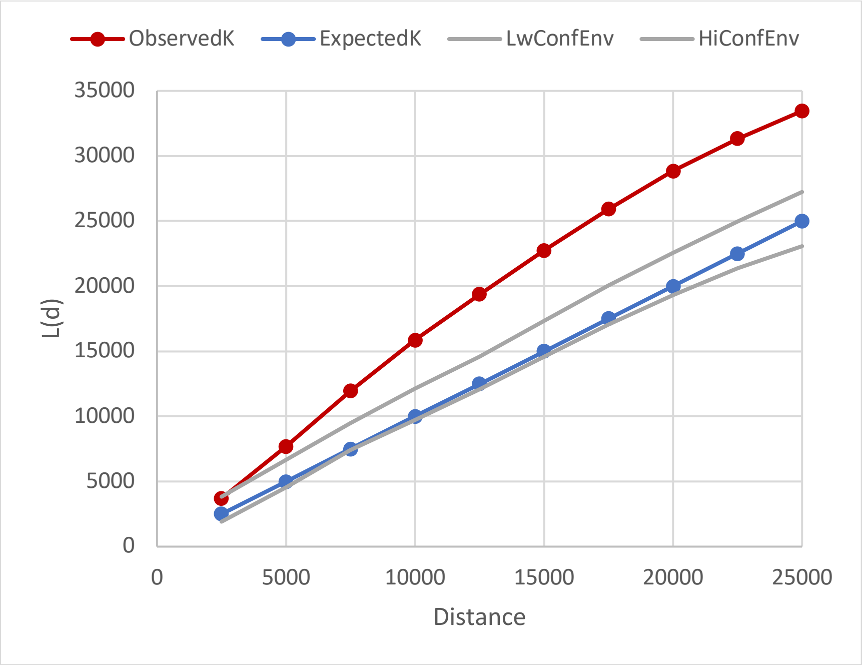

Figure 2 – ArcGIS K-Function Plot

These results from ArcGIS are supported by the results from R shown in Figures 3, 4, and 5.

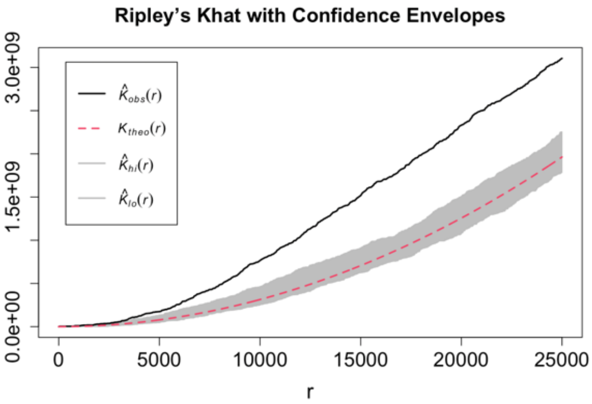

Figure 3 -$\widehat{K}$ Plot

Figure 3 shows the K-function results, which also show that the markets are clustered; however, the plot is considerably different than the ArcGIS results. The K-function is translated into an L-function and plotted in Figure 4. This plot from R is nearly identical to the ArcGIS results.

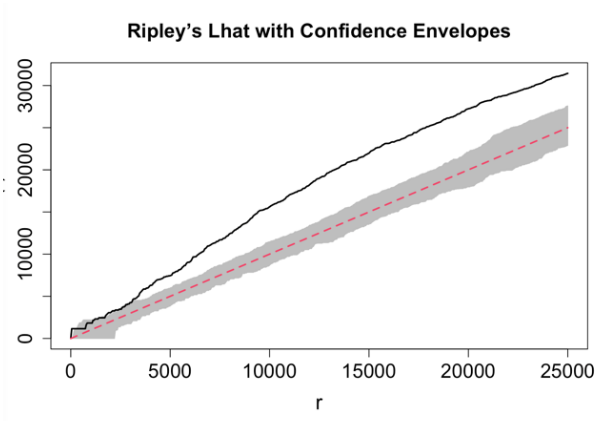

Figure 4 -$\hat{L}$ Plot

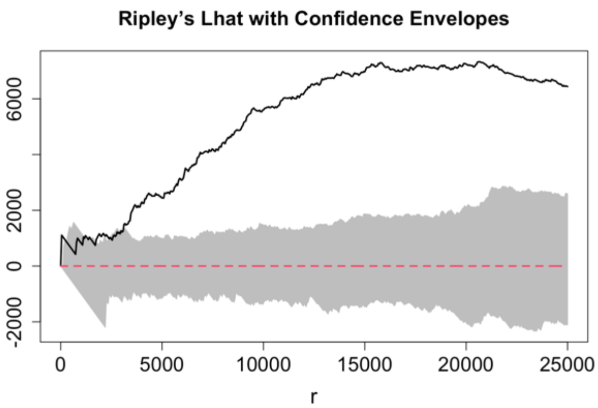

Figure 5 - Rotated $\hat{L}$ Plot

Figures 4 and 5 both plot the $\hat{L}$; however, to better see the amount of clustering, Figure 5 shows the data rotated 45 degrees. Again, there is clear clustering, even in smaller areas. The amount of clustering steadily increases until roughly 2,000 feet where it appears to start to decline slightly.

The clustering of farmers markets in central and Northwest Philadelphia may be due to external variables such as population. It would make logical sense that the market organizers would choose locations that are convenient for the most people possible. Therefore, an investigation that takes population by zip code or census tract into account might uncover relevant relationships.

Discussion

The Nearest Neighbor Analyses and K-Function Analyses are generally consistent in indicating that there is clustering of the Philadelphia Farmer's Markets. This agrees with a visual examination of the plot of Figure 1. Overall, the markets appear clustered in the center of the city, and within the central area some of the markets appear to form small clusters with groups of two or three in neighboring cells of the lattice.

The analyses differ, however, with their measurements of clustering within an area smaller than the city limits. The Nearest Neighbor results from both ArcGIS and R when using a minimum bounding box suggest that the locations of markets are random. This is opposed to the K-Function results in the smallest intervals, which still indicate some clustering, although not as significant as the clustering once the increments increase, and a larger area is considered.

These differences are a clear demonstration of the different approaches to point pattern analysis of these two functions. The minimum bounding box in this case is not nearly as relevant as the entire city limits since the question regards the entire city, not just a specific part of the city. With that in mind, it is fair to focus interpretation on the nearest neighbor results for the whole city paired with the k-function analysis to conclude that the farmer's markets are indeed clustered.

The results produced by R and ArcGIS are generally consistent, with some minor differences. The Nearest Neighbor values for the minimum bounding box are somewhat inconsistent; however, they suggest the same overall result of randomness. The inconsistencies suggest that the programs use different minimum bounding box areas. Conversely, the Nearest Neighbor values for the Philadelphia area and all K-function results are essentially identical.

As discussed in the end of the results section, the locations of farmers markets are not selected at random, they are planned. With that in mind, there are likely contributing variables to the clustering pattern, which could include population or, perhaps, median household income. To explore this theory, Figure 6 shows the median household income of Philadelphia zip codes and the locations of all farmers markets.

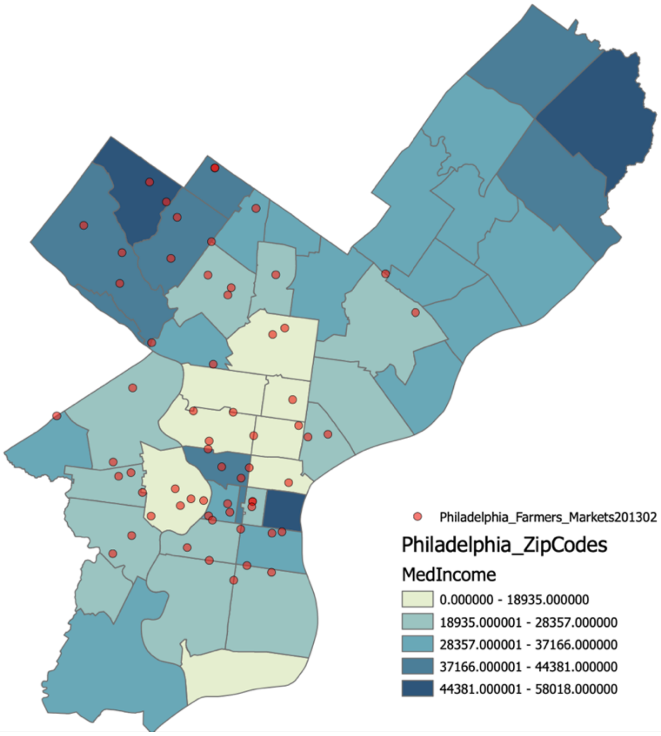

Figure 6 – Philadelphia Median Income and Farmer's Markets

This plot largely dispels the theory that the markets are only in wealthier parts of the city since several markets are in zip codes just North of Center City, which fall into the lowest income group. At the same time, some of the zip codes with the highest income, such as 19118 in the Northwest corner of the map, also have farmers markets. Finally, South Philadelphia and Northeast Philadelphia do not have any markets but show all five income groups. This suggests that there must be another factor leading the markets to cluster in this pattern. Population would be useful to investigate next.

Given the clustering, yet no clear sign that income level is a contributing factor (at least on the zip code level), the city is left with a challenging situation if it seeks to provide equal access to fresh produce for all residents. What is the current rationale for the locations of these markets? If there are financial constraints that prevent organizers from setting up markets in South and Northeast Philadelphia, perhaps the city could provide tax breaks or incentives to alleviate these challenges. If there are more logistical roadblocks, such as a suitable space or necessary parking for vendors, again, the city may be in a position to support. Undoubtably, the first step to answering these questions is to understand how the locations are chosen, which is beyond the scope of this report.