Bridget Kane

Spatial Distribution of Farmers Markets in Philadelphia

INTRODUCTION

Ensuring access to healthy, locally grown food has been a challenge for many American cities. The Philadelphia Food Trust has set up numerous benefits to consumers who shop there and to their communities. That said, not all neighborhoods in Philadelphia have access to these markets. Parts of South Philadelphia and North Philadelphia, and nearly all of Northeast Philadelphia, have no markets at all—this deprives those communities of the benefits offered by the city of Philadelphia, including: fresher and healthier foods, a greater variety of foods, reduced overhead (driving, parking, etc.), and a generally relaxing outdoor atmosphere.

This project aims to examine the spatial distribution of farmers markets in Philadelphia by carrying out several types of point pattern analyses to see whether markets are randomly placed, dispersed, or clustered throughout the city. It makes use of a shapefile containing locations of all farmers markets in Philadelphia for the year of 2013, obtained from the PA Spatial Data Access website (www.pasda.psu.edu).

METHODS

In this project we will be working heavily with the placement of points and the concept of randomness. A point process is completely spatially random (CSR) if the probability that a point lands in any of the cells is directly proportional to the area of

the cell. If you have equally sized cells, a point is equally likely to land in any cell, and where one point lands has no effect on where any of the other points land. We will be testing the null and alternative hypotheses for the point pattern analysis of this data:

H_0: Complete spatial randomness

H_a: A lack of complete spatial randomness

Quadrat counting is technique for analyzing spatial point patterns. The window containing the point pattern is divided into grid of rectangular tiles or "quadrats". The number of points of X falling in each quadrat is counted. These numbers are returned as a contingency table.[1] We may choose the number of cells to correspond to our area of study. In the case of farmers markets in Philadelphia, we would use 103,000 feet for the height and 90,000 feet for the width of the study area: this way, the lattice would encompass not just the farmers market points, but the entire Philadelphia area. The size of the quadrant may be arrived at from the following equation:

$$\text{Quadrant (cell) area} = (\sqrt{2} * \textit{(103,000*90,000) / \# of points})^2$$

Two statistical methods can be used to test the null hypothesis, those being the Kolmogorov-Smirnov test and Variance/Mean Ratio (VMR). The Kolmogorov-Smirnov test compares an observed set of frequencies to a theoretical frequency distribution. In the case of quadrat analysis of point pattern, we test the fit of our

[1]http://mapas.mma.gov.br/i3geo/pacotes/rlib/win/spatstat/html/quadratcount.html

observed set of point pattern frequency in quadrats against the random Poisson frequency distribution.

H_0: the pattern is random (the observed frequency is of the random Poisson frequency)

H_a: the pattern is NOT random (the observed point pattern frequency is not the random Poisson frequency)

This does not really distinguish between dispersion and clustering – just random vs. not random. To determine whether it's dispersed or clustered, we must look at the pattern if the test shows that the points are not random.

The test statistic is the largest absolute difference, D, given as D = |Co - Ce| where Co and Ce are the observed and expected cumulative proportions for each value of x. The computed test statistic *D* is compared to the critical value of *D* obtained from a special Kolmogorov-Smirnov table. The value of the critical value depends on the level of significance, α (0.05 for us), and degrees of freedom, ν. Here, for the degrees of freedom, ν is the sum of the observed frequencies (# of quadrats).

The VMA is the variance of the # of points per cell divided by the average of the # of points per cell:

$$VMR = VAR(\#)/MEAN(\#)$$

Here, VMR = 1 indicates we have a random pattern. Mean = Variance = λ, so mean/variance = 1. This is what happens when the observed frequency of points per cell is equal to the expected frequency of points per cell (under Poisson). VMR < 1 indicates uniformity (dispersion), where variance is less than the mean. That is, there is little variability in the number of points per cell. Variance = 0 indicates

perfect uniformity, because it means that there is no variability in the # of points per cell. VMR > 1 indicates clustering, where there is great variability in the number of points per cell. Some may have only a few, and others may have many (i.e., clustering).

Limitations of the Quadrat method include results stemming from differences in cell size and range. If we use the same pattern, same range, and different cell size, it is possible to obtain different results. The same can be said for using the same pattern, same cell size, and different range (causing different results). The quadrat method is a measure of point density, since we're counting the number of points per unit area (i.e. quadrat or cell), though we are not taking into consideration how far apart the points are, and how they're arranged in space. This is a critical limitation and so we should also examine other point pattern methods.

Nearest neighbor analysis examines the distances between each point and the closest point to it, and then compares these to expected values for a random sample of points from a CSR (complete spatial randomness) pattern—this means that it compares observed average distance between each point and its nearest neighbor, as well as the expected average distance between each point and its nearest neighbor if the point pattern were random. The Nearest Neighbor Index (NNI) = Observed Average Distance/Expected Average Distance (when pattern is random) is denoted as follows:

$$= \frac{D_0}{D_E}$$

When the NNI is close to 1, we have a random pattern. NNI = 1 implies that observed average distance is equal to the expected average distance when the pattern is

random. When NNI is close to zero, we have a clustered pattern. NNI = 0 implies that the observed average distance = 0, i.e. that all of the points in our pattern are located in the same spot. When NNI is close to 2 (at most 2.149), we have a dispersed pattern. NNI > 2 implies that the observed average distance is much larger than the expected average distance when the pattern is random. We noted that:

$\boldsymbol{NNI = D_0/D_E}$: This formula describes the observed average distance divided by the expected average distance (when the pattern is random). The formulas for D_0 and D_E are shown below.

$\boldsymbol{D_0 = \sum i = 1^n D_i/n}$: This formula describes the sum of all distances between each feature, *i,* and its nearest neighbor, divided by the number of features, *n*)

$\boldsymbol{D_E = \frac{0.5}{(\sqrt{n})/A}}$: This formula describes *n* as the number of features, and *A* as the area of the minimum enclosing rectangle.

Our hypothesis is as follows:

H_0: The observed point pattern is random (i.e. it isn't significantly different from the expected point pattern)

H_a: The observed point pattern is not random (i.e. there is either significant clustering or dispersion)

Our test statistic has a z (standard normal) distribution and is denoted below:

$$\frac{D_0 - D_E}{SE_{DO}}$$

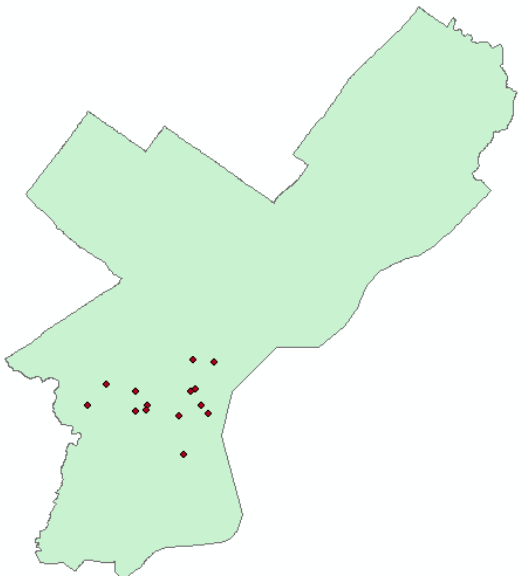
$$z = D_0 - \frac{D_E}{SE} = \sum i = \frac{1^n D_i}{n} - 0.5/(\frac{\frac{\sqrt{n}}{A}}{\frac{0.26136}{\sqrt{n^2}}})/A$$

Here, we are taking D_0, subtracting D_E, the expected value of D_0 when H_0 is true, and dividing the difference by (an estimate of) the standard error of D_0. We are able to

use the standard normal table to get a p-value from calculated z, which corresponds to an α-value and helps to assess significance to reject, or to fail to reject, H_0.

If z > 19.6 or z < -1.96, we can reject H_0 for H_a at α = 0.05. If z > 1.96, we have significant dispersion. This implies that $D_0 > D_E$, i.e. the average observed distance is greater than the average expected distance, meaning that we have significant dispersion. If z < -1.96, we have significant clustering. This implies that $D_0 < D_E$, i.e. that the average observed distance is smaller than the average expected distance, meaning that we have significant clustering.

NNI does not depend on the size of the quadrat because it only uses the distances between points—this accounts for one limitation in VMR we discussed previously. That said, there are still limitations: NNI takes into account average distance to only the nearest neighbor. It depends greatly on the value of A, the area of the study region. Taking the example of hospital locations in Philadelphia, which has an irregular (non-rectangular) shape.



The figure to the left shows hospitals clustered near center city—however, NNI also does not take into consideration the fact that both clustering and dispersion may be present at different scales. Thus, if your study area is irregularly shaped you can only increase the area of the bounding rectangle to account for this shape—quadrat analysis also has this issue.

The idea behind the K-Functions Analysis is to help show how spatial clustering or dispersion changes when the neighborhood size (scale) changes. To further illustrate this idea, the K(d) function is described below, along with a series of steps that k-functions will follow:

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

Assume we have *n* points in our dataset, and that the area of the study region is denoted by *a*. A description of K-functions is as follows:

(1) Place circles, each of radius *d*, around every event (point).
(2) The number of *other* events (points) inside each circle of radius *d* is then counted
(3) From here, it is possible to calculate the average number of *other* events (points) in all circles of radius *d*.
(4) We then divide this average count of other events by the overall study area event density (n/a) *to* get the K-function at distance d, as denoted as K(d).
(5) These steps are repeated for a range of values of *d*.

Under this function:

K(d) > πd^2 implies clustering at scale d
K(d) < πd^2 implies dispersion at scale d

For ease of interpretation, however, many statistical software packages report results in terms of L(d) functions and not K(d) functions. This may be defined as a transformed K(d) function, such as that for all (non-negative) distances *d*:

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

$$L(d) = \sqrt{\frac{a \cdot \#\ of\ points\ within\ circle\ of\ radius\ d\ (excluding\ point\ at\ center\ of\ circle)}{\pi \cdot n \cdot (n-1)}}$$

Here,
$L(d)$ = 0 under CSR
L(d) > 0 when there is clustering at scale d
L(d) < 0 when there is dispersion at scale d

In ArcGIS,
L(d) = $\sqrt{a}$ # of points within the circle of radius d (excluding point at center of circle)/ πn(n-1)

$$Under\ CSR,\ \ L(d) = \sqrt{\frac{\pi d^2}{\pi}} = d$$

Hypothesis testing for K-functions describe:

H0: At distance *d*, the pattern is random (you have Complete Spatial Randomness at distance *d*)

Ha1: At distance *d*, the pattern is clustered

Ha2: At distance *d*, the pattern is uniform

To test significance, we take the randomly generated patterns and for each distance *d* find the lowest value of L(d), denoted by L(*d*) and called lower envelope, and the highest values of L(d), denoted by L(*d*) and called upper envelope. These values represent the lowest and highest values of L(*d*) that you would expect to occur by chance (i.e. under CSR) at each particular distance *d*. Because this confidence envelope is constructed from random permutations, the values defining the confidence envelope will change from one run to the next, even when the parameters are identical. By setting a seed value for the Random Number Generator geoprocessing environment, repeat analyses will produce

consistent results. The number of permutations selected for the Compute Confidence Envelope parameter may be loosely translated to confidence levels: 9 for 90%, 99 for 99%, and 999 for 99.9%. For instance, when you have 999 permutations and L^{obs}(d) < L-(d) at some distance d, then you can be approximately 99.9% confident that you have significant dispersion at that distance d.

For this study, we will be using the following settings in ArcGIS for our analysis to determines whether features, or the values associated with features, exhibit statistically significant clustering or dispersion over a range of distances:

a. Use 10 as the 'Number of Distance Bands'

b. Select 99 Permutations under 'Compute Confidence Envelope'

c. Check 'Display Results Graphically'

d. Leave the 'Weight Field' blank

e. Put in 0 for 'Beginning Distance'

f. Put in 2500 feet for 'Distance Increment'

g. Select Simulate_Outer_Boundary_Values for 'Boundary Correction Method'

h. Select User_Provided_Study_Area_Feature_Class under 'Study Area Method'

i. Select the Philadelphia shapefile under 'Study Area Feature Class')

Points very close to the boundary of a specified region can be accounted for by the Ripley's edge correction, which checks each point's distance from the edge of the study area and its distance to each of its neighbors. All neighbors that are farther away from the point in question than the edge of the study area are given extra weight. Another way of saying this is that neighboring points located

inside the study area are given extra weight to account for the fact that there could never be points in the part of the circle that's outside the study area.

While Ripley's edge correction works only for rectangular study regions, there are other edge corrections in ArcGIS. The Simulate Outer Boundary Values edge correction method 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. The mirrored points are used so that edge points will have more accurate neighbor estimates. In this project we will be utilizing the Ripley's edge correction.

RESULTS

The two-sample Kolmogorov-Smirnov test output is displayed below:

```
Two-sample Kolmogorov-Smirnov test
data:  obscumprop and expcumprop
D = 0.8, p-value = 0.002057
alternative hypothesis: two-sided
```

Because the p-value is small, we may conclude that the two groups were sampled from populations with different distributions. The populations may differ in median, variability or the shape of the distribution.

Results of our Nearest Neighbor analysis are described on the following page. Given the z-score of -0.0699452830094 and p-value of 0.944237, the pattern does not appear to be significantly different than random.

Average Nearest Neighbor Summary

Observed Mean Distance:	3112.9097 US_Feet
Expected Mean Distance:	3127.4314 US_Feet
Nearest Neighbor Ratio:	0.995357
z-score:	-0.069945
p-value:	0.944237

Dataset Information

Input Feature Class:	Philadelphia_Farmers_Markets201302
Distance Method:	EUCLIDEAN
Study Area:	2425645188.585607
Selection Set:	False

Upon re-running this analysis using the area of Philadelphia polygon (as opposed to the minimum enclosing rectangle), however, our results change tremendously. By changing the area of the enclosing rectangle, our z-score becomes -3.344634, and our p-value becomes 0.000824. Given the z-score of -3.34463398096, there is a less than 1% likelihood that this clustered pattern could be the result of random chance. Our output is described in greater detail below:

Average Nearest Neighbor Summary

Observed Mean Distance:	3112.9097 US_Feet
Expected Mean Distance:	4001.3504 US_Feet
Nearest Neighbor Ratio:	0.777965
z-score:	-3.344634
p-value:	0.000824

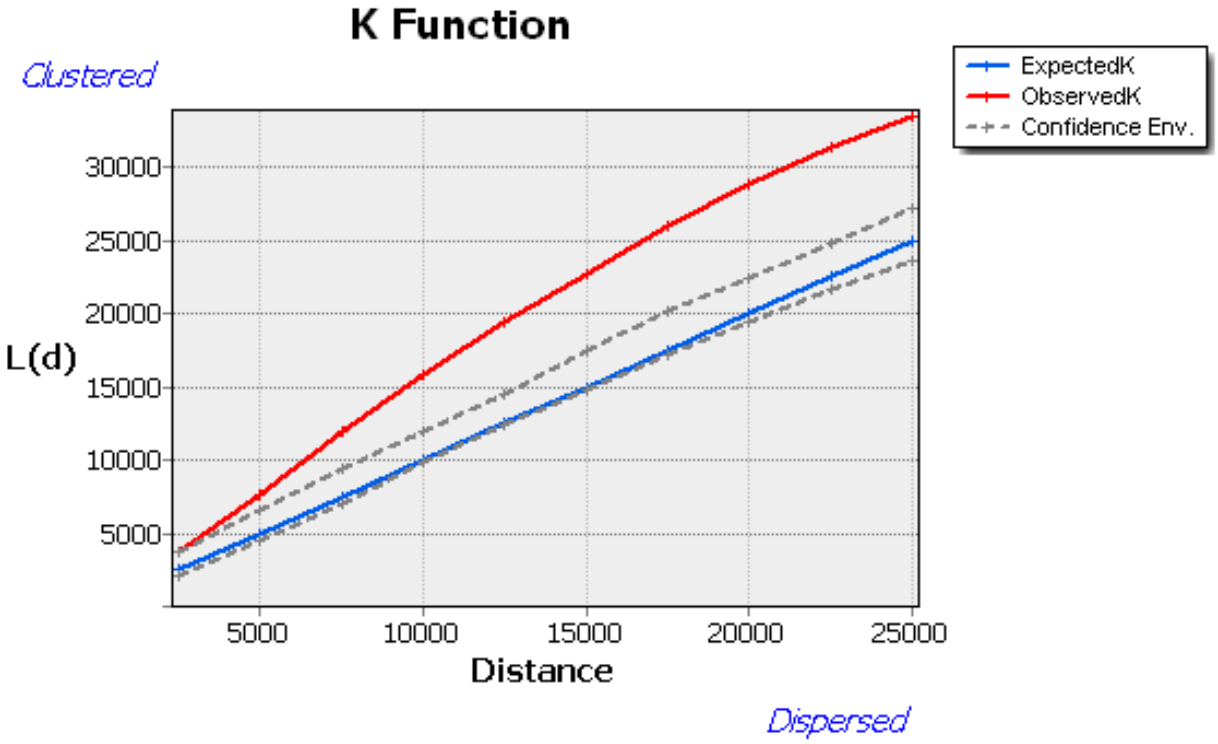
Dataset Information

Input Feature Class:	Philadelphia_Farmers_Markets201302
Distance Method:	EUCLIDEAN
Study Area:	3970679604.941694
Selection Set:	False

When we look at the results of the nearest neighbor analysis in terms of the minimum enclosing rectangle, we are not able to reject the null hypothesis that the observed point pattern is random, in favor of the alternative hypothesis, that the observed point pattern is not random (i.e. there is either significant clustering or dispersion). That said, when we use the area of the Philadelphia polygon, our results change significantly, and we can safely reject the null hypothesis in favor of the alternative hypothesis for significant clustering.

Our K-function analysis in ArcGIS produces this illustration of ExpectedK and ObservedK containing the expected and observed K values, respectively. Here, we can see that our observed K value is larger than the expected K value for a particular distance, meaning that the distribution is more clustered than a random distribution at that distance (scale of analysis). Because our observed K

value is also larger than the HiConfEnv value, especially at larger distances, spatial clustering for that distance is statistically significant.

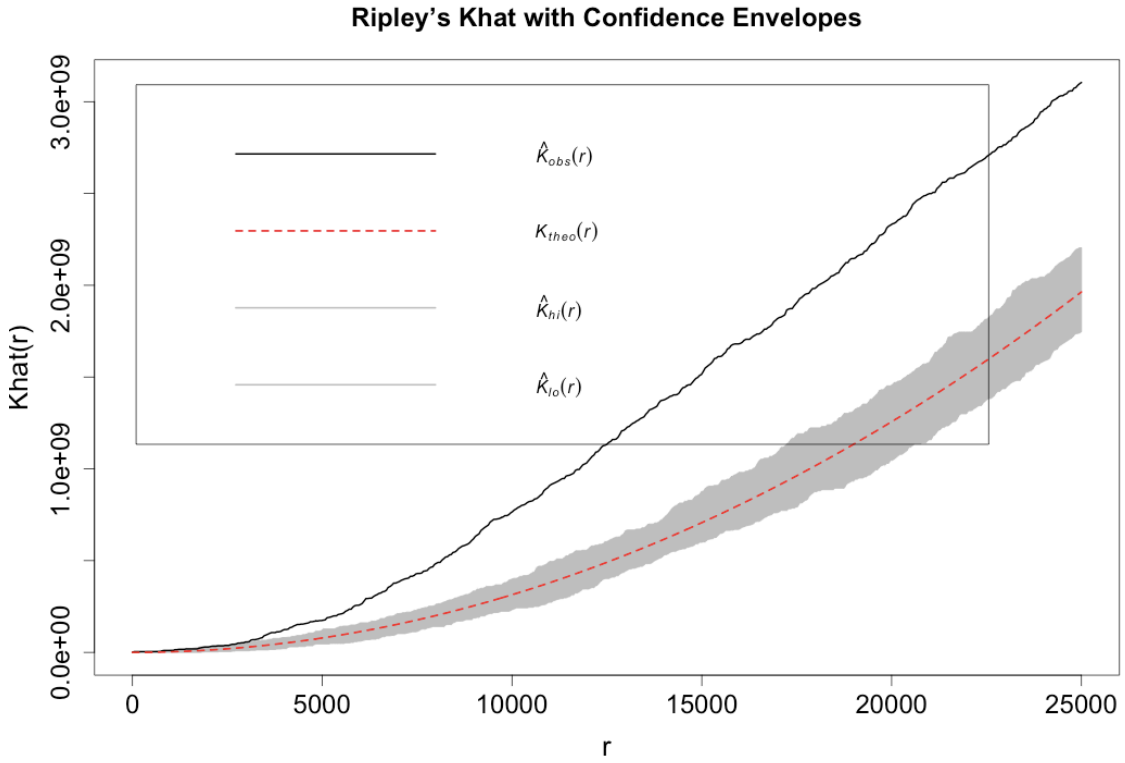


The ArcGIS output also produces the following table, which provides the observed and expected K values, as well as the difference between them (observed k minus expected k), and the high and low confidence interval information for each iteration of the tool, as specified by the number of bands parameter.

OBJECTID	ExpectedK	ObservedK	DiffK	LwConfEnv	HiConfEnv
1	2500	3701.593	1201.593	2163.021	3701.593
2	5000	7691.01	2691.01	4551.898	6641.768
3	7500	11959.65	4459.654	7056.508	9463.769
4	10000	15863.32	5863.325	9844.547	11889.59
5	12500	19372.54	6872.542	12385.2	14544.49
6	15000	22737.46	7737.456	14828.93	17543.92
7	17500	25930.49	8430.489	17197.62	20175.32
8	20000	28846.72	8846.717	19484.35	22456.46
9	22500	31334.49	8834.489	21691.92	24824.31
10	25000	33454.48	8454.481	23581.61	27244.01

We are able to reject the null hypothesis of a random pattern (complete spatial randomness at distance *d*) as well as the second alternative hypothesis (at distance *d*, the pattern is uniform) in favor of the first alternative hypothesis (at distance *d*, the pattern is clustered). Though our graph of the k-function analysis does show that clustering increases at larger distances, it remains that the pattern is, in fact, clustered at all distances and shown to be statistically significant at each distance.

An analysis in R on the following page provides similar results:

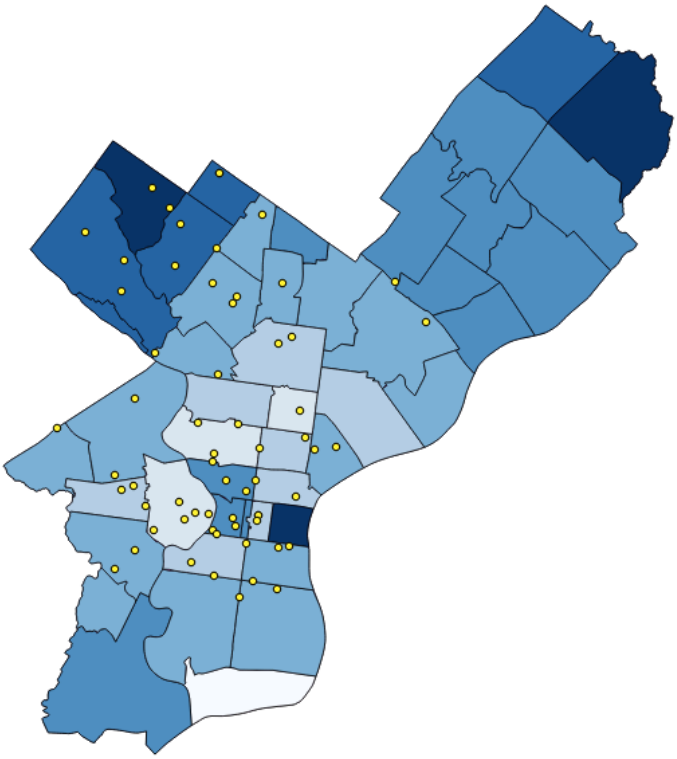


Here, we can see that our observed k-values are well above the high confidence interval for each iteration of the tool. Again, this means that we have significant spatial clustering at distance *d*.

Ideally, we would want to compare our actual pattern to 9, 99, or 999 generated patterns that take into consideration the population density that is being served. It is certainly possible that we would see different results if we were to take into consideration the population density at each zip code—a lack of grocery stores, fresh food access, etc., might sometimes be attributed to lower populations. Despite population density, it is still very important to recognize which communities have little access to these resources.

DISCUSSION

The results obtained with the nearest neighbor analysis (using the area of Philadelphia as the bounding area, as opposed to the minimum bounding rectangle) and k-function analysis are consistent with each other in that we are able to reject the null hypotheses of spatial randomness in favor of the alternative hypothesis of spatial clustering.



Based on the visual examination of the point data, and given the limitations of each method, our results are still consistent with our expectations for significant spatial clustering. A map of median household income at the zip code level is shown to the left. There is some evidence for clustering in these areas, however, clustering seems to be more strongly associated with location (i.e. center city branching off into the northwestern suburbs) as opposed to following an income-based trend.

Results of the k-function analysis for the spatial distribution of farmers markets in Philadelphia conclude that farmers markets are clustered in the city. Implications of these findings may be to take further measures to assure that all

Philadelphians have access to the benefits of these farmers markets. This might be accomplished by delving deeper into the cause of this clustering: perhaps these are areas more heavily frequented by tourists or commuters, or maybe they feature certain characteristics that make people more inclined to shop there. Whatever the reason, it is in the Philadelphia Food Trust's best interests to continue to dig into this issue to help provide farmers market access to all Philadelphians where they need them most.