

Gabriel Hernandez, Timothy Oliver, Ziyi Tang

Professor Eugene Brusilovskiy

MUSA 5000/CPLN 6710

10 Dec 2023

Assignment 4: The Spatial Distribution of Farmers Markets in Philadelphia

1. Introduction

Farmers markets play an integral role in urban life, offering myriad benefits to city dwellers. These markets are more than just sources of fresh, seasonal, and healthy food options; they are vibrant community hubs where residents can engage with neighbors and enjoy the outdoors while shopping for groceries. The availability of farmers markets contributes significantly to the timely access of nutritious food, promoting a healthy and enjoyable lifestyle. However, the distribution of these markets is not uniform across Philadelphia. Particularly in southern, northern, and northeast parts of the city, the absence of farmers markets deprives residents of their advantages. This study delves into the geographic spread of farmers markets in Philadelphia, employing nearest neighbor and K-function analyses in ArcGIS to investigate the extent of their clustering. The goal is to provide insights and recommendations to local authorities, addressing the disparities in access to these vital community resources.

2. Methods

2.1 Quadrat Method

In our study, we aim to discern the spatial arrangement of farmers markets in Philadelphia, specifically examining if their distribution is random or exhibits clustering. The core hypotheses tested are:

Null Hypothesis (H_0): The distribution of farmers markets follows a pattern of Complete Spatial Randomness (CSR).

Alternative Hypothesis (H_1): The distribution of farmers markets is not random but clustered.

To understand CSR, we must consider two key conditions:

Uniformity: The likelihood of a market being located in any given area is proportional to the size of that area. In a scenario where areas (or quadrats) are of equal size, each has an equal chance of containing a point.

Independence: The placement of one market does not influence the location of another. Essentially, the existence of a market in one area does not affect the probability of another appearing nearby.

Although not employed in this study, the Quadrat method is relevant in understanding spatial patterns. It involves dividing the study area into equally sized cells (quadrats) and then counting the number of points (in this case, farmers markets) in each cell. The Variance-to-Mean Ratio (VMR) of these counts is then computed to assess the point pattern.

However, the Quadrat method has limitations, primarily its sensitivity to the size and shape of the quadrats. The chosen dimensions of these cells can significantly influence the results, leading to potential misinterpretation of spatial patterns. Small quadrats might suggest clustering where none exists, while large quadrats could obscure existing clusters. Due to these drawbacks, the Quadrat method is often bypassed in favor of more reliable techniques, such as the nearest neighbor and K-function analyses, which we will utilize in our study to more accurately discern the spatial distribution of farmers markets.

2.2 Nearest Neighbor Analysis

2.2.1 Description

The Nearest Neighbor Analysis (NNA) scrutinizes the spatial relationship between each point in a dataset and its closest neighbor. This method gauges whether the distribution pattern of these points aligns with what is expected from a Completely Spatially Random (CSR) arrangement. The measure used for this assessment is the Nearest Neighbor Index (NNI), which is calculated using the following equation:

$$NNI = \frac{\text{Observed Average Distance}}{\text{Expected Average Distance (when pattern is random)}} = \frac{\bar{D}_o}{\bar{D}_e} \quad (1)$$

Here, \bar{D}_o represents the mean of the observed distances from each point to its nearest neighbor, while \bar{D}_e is the expected mean distance for a random pattern, calculated as

$$\bar{D}_e = \frac{0.5}{\sqrt{\frac{n}{A}}}, \text{ where } n \text{ is the count of points and } A \text{ is the area encompassing the points.}$$

The NNI value is interpreted to determine the nature of the pattern. An NNI approaching 1 indicates a random distribution, aligning with CSR expectations. An NNI significantly less than 1 suggests a clustered pattern, implying that points are more closely spaced than what randomness would predict. Conversely, an NNI considerably greater than 1 signals a dispersed pattern, where points are more uniformly distributed across the area than expected by chance.

2.2.2 Statistical Test

In our analysis, we will perform a hypothesis test using the NNA to discern if there is a pattern of significant clustering or dispersion among our data points. The null hypothesis posits that the distribution of points is random, not notably different from what would be expected by chance. Conversely, the alternative hypothesis suggests that the distribution is non-random, characterized by meaningful clustering or dispersion.

The statistical test utilizes a z-score, which follows a standard normal distribution, to evaluate the deviation from randomness. The test statistic is calculated as follows:

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

where:

- \bar{D}_O is the mean observed distance between each point and its nearest neighbor,
- \bar{D}_E is the expected mean distance for a random pattern,
- $SE_{\bar{D}_O}$ is the standard error of \bar{D}_O

Utilizing the standard normal distribution table, we can determine the p-value from the computed z-score. In the context of a two-tailed hypothesis test, the condition for the alternative hypothesis is that the expected average distance (\bar{D}_E) is not equal to the observed average distance (\bar{D}_O). The critical z-score values of ± 1.96 correspond to a significance level (α) of 0.05. As such, a z-score exceeding 1.96 or falling below -1.96 leads us to dismiss the null hypothesis in favor of the alternative hypothesis (H_0) at $\alpha=0.05$. Concretely, a z-score greater than 1.96 indicates a significant level of dispersion, suggesting that the observed average distance is substantially larger than what would be expected by chance. Conversely, a z-score less than -1.96 signals significant clustering, meaning the observed average distance is notably smaller than the expected average distance under a random distribution.

2.2.3 Limitations of the Nearest Neighbor Analysis

However, the NNA has its constraints. It solely evaluates the average distance to the nearest neighbor, which may not reflect complexities such as the geographical shape of the study area

or the scale at which clustering occurs. For example, when considering the distribution of hospitals within an urban area like Philadelphia, the NNA might be limited by its sensitivity to the city's irregular boundaries and the scale of observation. Hospitals may cluster in the city center, but this method, with its reliance on a minimum enclosing rectangle, might not accurately reflect such a pattern, particularly if the hospitals are not evenly distributed throughout the city. This illustrates the method's susceptibility to outliers and its potential to misrepresent the true nature of spatial patterns in certain contexts.

2.3 K-function Analysis Method

2.3.1 $K(d)$ and $L(d)$ functions

The K-function is a statistical measure that assesses the variance in clustering or dispersion across varying scales of neighborhood size. This function operates by positioning circles of a specific radius d around each point on a plane and tallying the count of additional points residing within each of these circles. It then computes the mean count of these points across all circles with radius d . This average is then normalized by the total density of points across the plane to yield the K-function at distance d , represented as $K(d)$. The K-function is articulated mathematically as:

$$K(d) = \frac{\text{Mean count of points within circles of radius } d}{\text{Mean point density in entire study region } A} = \frac{(\sum_{i=1}^n \# [S \in \text{Circle}(s_i, d)]) / n}{n/A} \quad (3)$$

Where:

- n is the total number of points in the dataset,
- A symbolizes the area of the study region,
- S_i is the center point of each circle,
- d is the radius of the circles,

- $K(d)$ reflects the average observed number of points within a circle of radius d , adjusted for the overall point density in the study area.

In a scenario where points are distributed in a Completely Spatially Random (CSR) fashion, the K-function is expected to equal πd^2 . If $K(d)$ exceeds πd^2 , this indicates a clustering at the scale of d ; conversely, if $K(d)$ falls below πd^2 , this points to a dispersion at the scale of d .

Some statistical programs use an alternative function called $L(d)$ for analyzing spatial point patterns, which is derived from the K-function. The $L(d)$ function is defined by the equation:

$$L(d) = \sqrt{\frac{K(d)}{n}} - d \quad (4)$$

- 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:

$$L(d) = \sqrt{\frac{K(d)}{n}} \quad (5)$$

- Under CSR, $L(d) = 0$;
- When $L(d) > d$, clustering;
- When $L(d) < d$, dispersion.

2.3.2 Beginning and incremental distances

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 is 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:

$$d = \frac{0.5 \text{ maximum pairwise distance}}{\# \text{ of distance bands}} \quad (6)$$

2.3.3 Testing procedure and the concept of confidence envelopes

The evaluation of the K-function relies on point patterns that are shuffled randomly. Our hypothesis for testing at a specific distance d is framed as follows:

Null Hypothesis (H_0): The spatial pattern is random at distance d .

Alternative Hypothesis 1 (H_1): There is clustering at distance d .

Alternative Hypothesis 2 (H_2): The pattern is uniform at distance d .


To test these hypotheses, we simulate a series of random point patterns, each with n points, and calculate the L-function, $L(d)$, for each pattern. We identify the minimum L-function value, called the Lower Envelope ($L - (d)$), and the maximum L-function value, called the Higher Envelope ($L + (d)$). The observed L-function value, denoted $L_{obs}(d)$, is then compared against these envelopes at each distance d :

- If $(L - (d)) < L_{obs}(d) < (L + (d))$, the Null Hypothesis H_0 cannot be rejected at distance d , indicating that the observed pattern does not significantly deviate from randomness.
- If $L_{obs}(d) > (L + (d))$, we reject H_0 in favor of H_1 at distance d , indicating significant clustering at this scale.


- If $L_{obs}(d) < (L - (d))$, \mathbf{H}_0 is rejected in favor of \mathbf{H}_2 at distance d , indicating significant dispersion at this scale.

The construction of a confidence envelope is based on the method of randomly distributing points across the study area. Each iteration of this random distribution is referred to as a "permutation." Since the values that form the confidence envelope are subject to change with each iteration, it's crucial to establish a consistent starting point, known as a "seed," for the number of permutations to ensure reproducibility in results. The amount of permutations chosen corresponds with specific confidence intervals: 9 permutations for a 90% confidence level, 99 for 99%, and 999 for a confidence level of 99.9%. For example, with 999 permutations, if the observed L-function value, $L_{obs}(d)$, is less than the lower envelope, $(L - (d))$, at any given distance d , it implies with approximately 99.9% certainty that there is significant dispersion at that distance.

2.3.4 Ripley's Edge Correction and the Simulate Outer Boundary Values Edge Correction

Border effects are a common challenge in spatial analysis, where points on the perimeter may be inaccurately assessed due to partial circle overlaps. To mitigate this, ArcGIS uses two correction methods: Ripley's Edge Correction Formula, which adjusts the weights of the circles based on how much of their area falls within the boundary, and the Simulate Outer Boundary Values, which doubles the border points across the boundary to provide a better estimate of neighbors. This study opts for the latter method, Simulate Outer Boundary Values, since it is not constrained to rectangular areas like Ripley's Edge Correction. 

2.3.5 Nonhomogeneous

When analyzing the K-function, it is essential to consider underlying factors that may influence clustering, such as the distribution of social resources correlating with population density. In ArcGIS, we utilize the "Spatially Balanced Points" tool which distributes sample points proportionately to the inclusion probability depicted in a density raster. This entails converting population figures into probabilities and transitioning from shapefile to raster format, reflecting population density. Subsequently, spatially balanced points are generated, and random point patterns are simulated multiple times (9, 99, or 999). In computing $L(d)$ for each pattern, we choose the appropriate number of permutations to form the confidence envelope and apply the "Simulate Outer Boundary Values" for correction. The final step is to correlate the output table with the permuted point patterns by distance to the Expected K.  This process ultimately determines at which distances the patterns exhibit clustering, randomness, or dispersion, as shown in the output tables.

3. Results

3.1 Initial Nearest Neighbor Analysis Result

Utilizing the nearest neighbor analysis with the bounds of Philadelphia defined by the smallest enclosing rectangle, the findings are depicted in figure 1. The nearest neighbor ratio yielded a value of 0.9954, which approximates 1. The z-score obtained is -0.069945, which does not exceed the critical value of 1.96, and the accompanying p-value stands at 0.9442, surpassing the threshold of 0.05. Interpreting the z-score of -0.069945, it can be inferred with 94.42% confidence that the observed spatial pattern of farm markets could have arisen from random distribution. Consequently, we do not dismiss the null hypothesis, suggesting that the distribution of farm markets across Philadelphia occurs randomly.

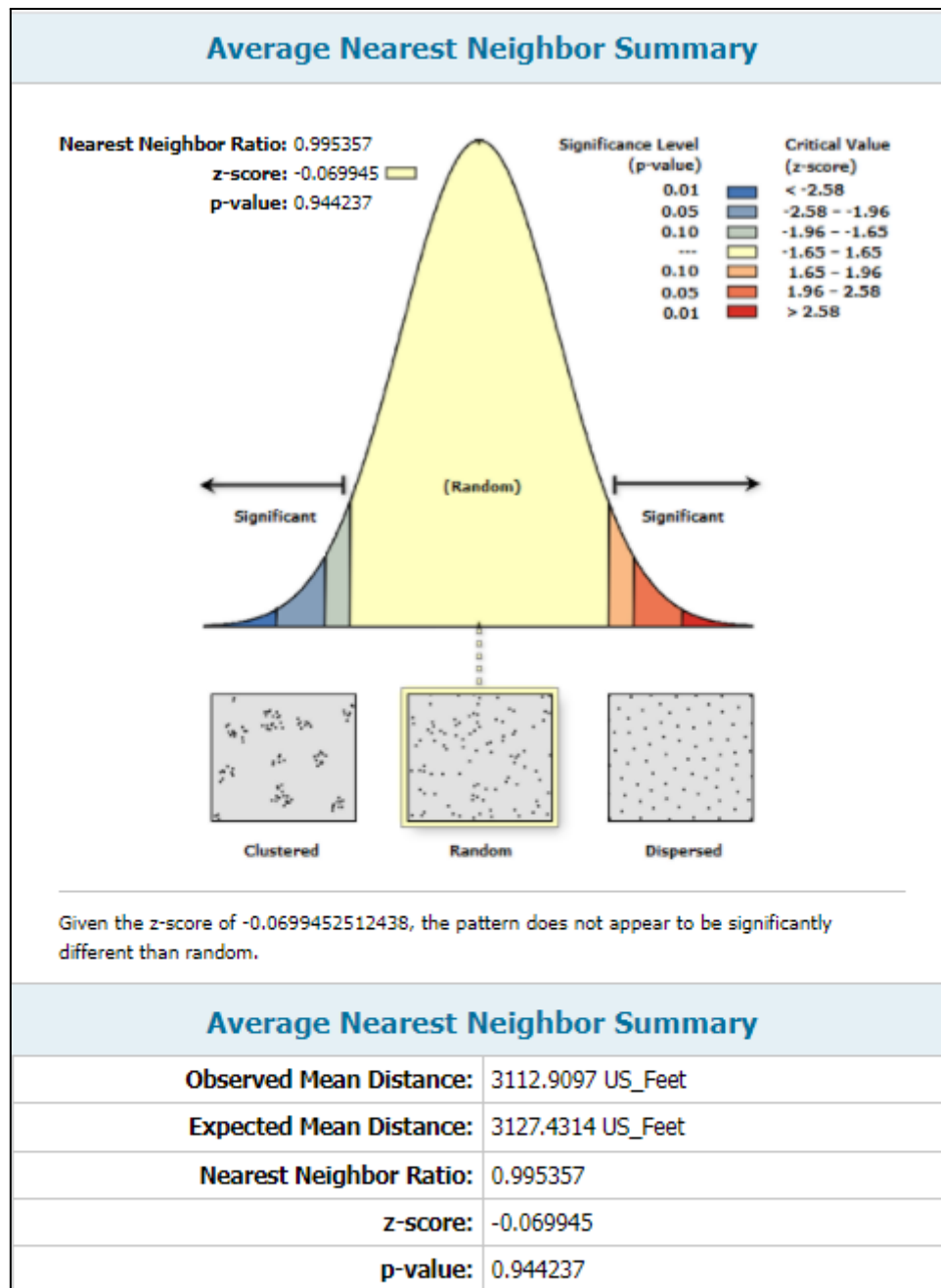


Figure 1. Results of the Nearest Neighbor Analysis

3.2 Second Nearest Neighbor Analysis Result

To address the issues that often arise with nearest neighbor analysis, the analysis was conducted anew, this time considering the actual area of Philadelphia instead of the minimum enclosing rectangle as illustrated in figure 2. The calculated z-score, at -3.344634, falls substantially below the critical value of -1.96, and the p-value is significantly low at

0.000824, well under the 0.01 mark. With such a z-score, the probability that the observed clustering of farm markets is due to random distribution is less than 1%. Therefore, the null hypothesis is rejected in favor of the alternative hypothesis, indicating that the farm markets are clustered within Philadelphia.

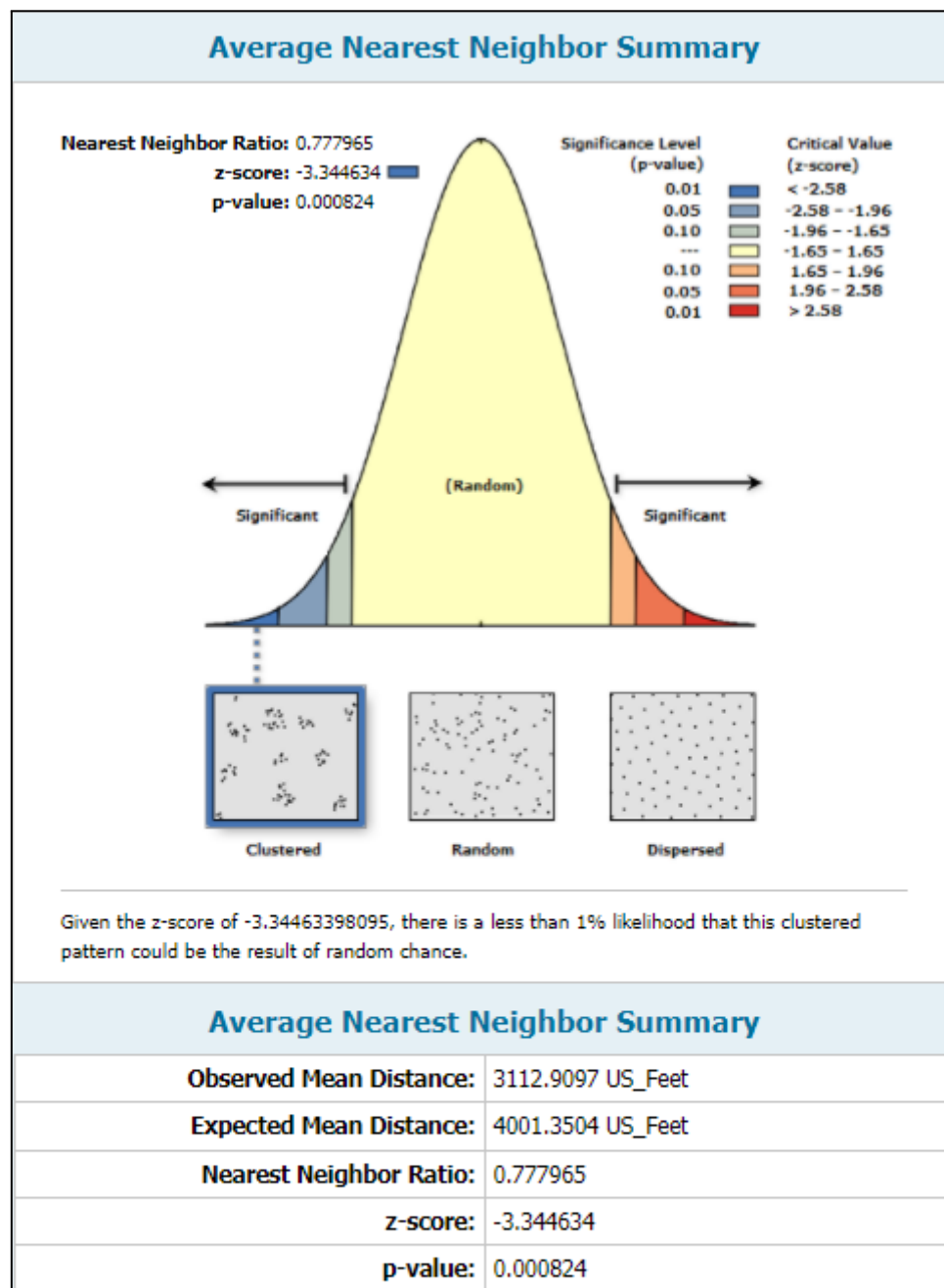


Figure 2. Results of the Nearest Neighbor Analysis, using the area of Philadelphia

3.3 K-function Analysis Result

During the K-function analysis, we selected a starting and incremental distance of 2500 feet, which aligns with the typical dimensions of urban block groups, providing a relevant scale for our measurements. As indicated by the data in the subsequent Table 1 and Figure 3, the Observed K-function consistently exceeds the Expected K-function across all measured distances. Furthermore, the observed K values surpass the upper limit of the confidence interval as presented in Figure 3. Additionally, the Expected K-function falls outside the confidence interval at both shorter and longer distances. These findings allow us to dismiss the null hypothesis and confirm that the spatial clustering at these distances is not random but indeed significant.

	OBJECTID *	ExpectedK	ObservedK	DiffK	LwConfEnv	HiConfEnv
▶	1	2500	3701.59289	1201.59289	2163.02103	3834.626438
	2	5000	7691.009666	2691.009666	4696.439064	6514.762449
	3	7500	11959.653809	4459.653809	7357.905932	9231.383945
	4	10000	15863.32464	5863.32464	9725.007429	12112.365962
	5	12500	19372.541598	6872.541598	12385.201216	14601.822449
	6	15000	22737.45648	7737.45648	14896.380728	17629.427056
	7	17500	25930.489319	8430.489319	17207.333855	20150.455275
	8	20000	28846.717239	8846.717239	19612.564542	22567.799122
	9	22500	31334.48922	8834.48922	21691.922862	25032.104318
	10	25000	33454.481074	8454.481074	23517.74749	27305.272127

Table 1. K Function Results Output Table

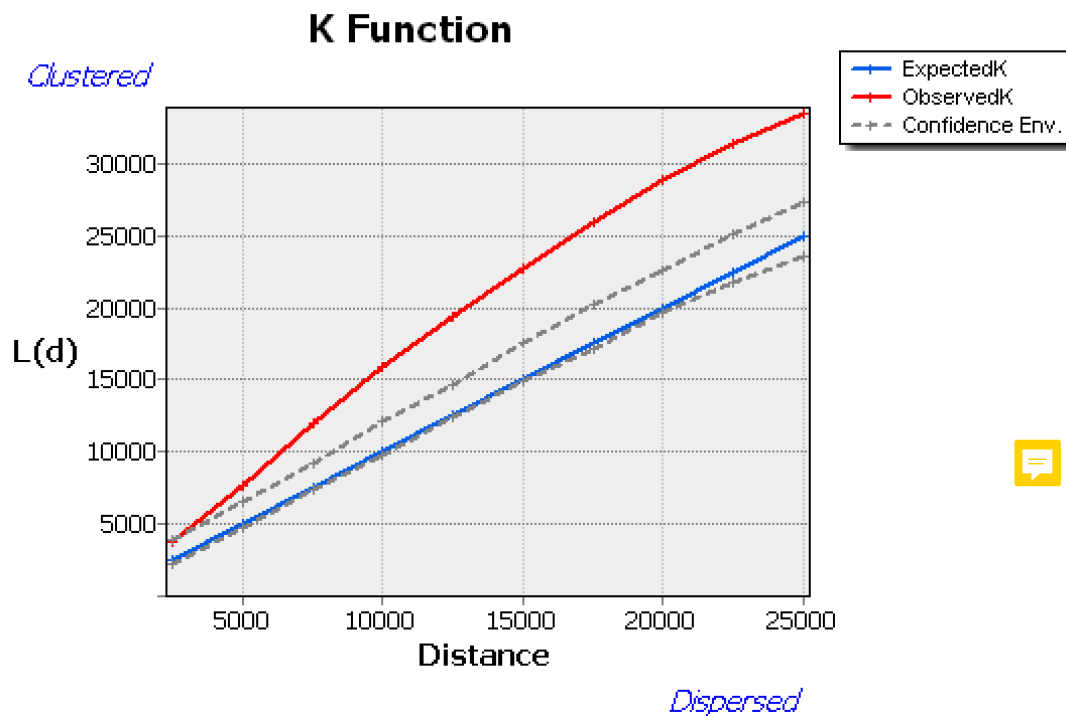


Figure 3. K Function Results Plot

In exploring the reasons behind the clustering, it is evident that farmers markets are scarce in the Northeast, North, and South regions of Philadelphia. This scarcity, however, cannot be attributed merely to population density, as North Philadelphia is densely populated and South Philadelphia has a moderate population density, as shown in figure 4. It is posited that the primary causes are the high poverty rates in these areas, rather than a simple deficit in population numbers. Incorporating population data into the analysis might yield different insights into clustering patterns. Nonetheless, considering population density as a factor in this particular analysis would not yield relevant results.

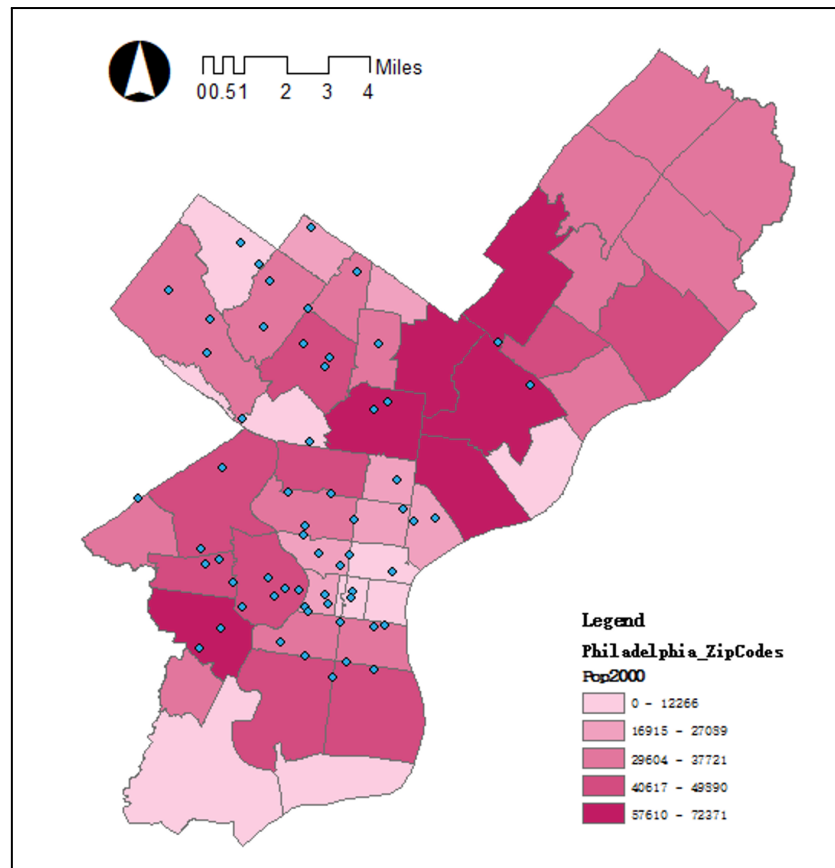


Figure 4. Population Density and Farmers Markets Distribution Map

4. Discussion

The findings from the initial Nearest Neighbor Analysis and the subsequent K-function Analysis revealed incongruities, primarily due to differing boundary delineations used in each method. Initially, the Nearest Neighbor Analysis, utilizing the minimum enclosing rectangle, suggested a random distribution of farmers markets in Philadelphia, failing to reject the null hypothesis. However, after adjusting the analysis area to accurately reflect Philadelphia's actual boundaries, the revised Nearest Neighbor Analysis indicated significant clustering, aligning with the K-function Analysis's indications of notable spatial clustering.

The results matched what I expected after looking at where the points were placed on the map. The main problem with the Nearest Neighbor Analysis is that it only looks at the closest

point and uses the smallest box to cover all points, which can mess up the results. When we used the actual shape of Philadelphia instead of this small box, we got better results that made more sense and agreed with what the K-function analysis showed us. However, the K-function analysis isn't perfect either; it doesn't think about important things like how many people live in an area. Figure 5 shows us that in Philadelphia, places with lower median incomes, like North and South Philly, don't have many farmers markets. But even areas with medium to high median income, like Northeast Philly, also don't have many farmers markets.

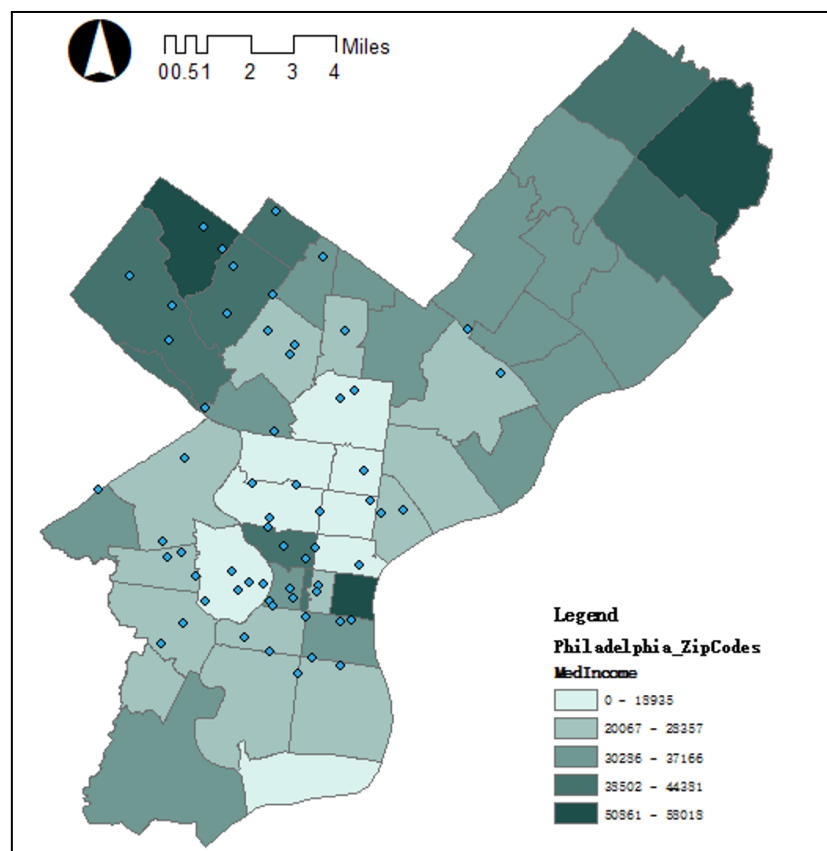


Figure 5. Median Household Income and Farmers Markets Distribution Map

In summary, our analyses using Nearest Neighbor and K-function methods indicate that farmers markets in Philadelphia tend to be concentrated in certain areas. This clustering is observed both in less affluent neighborhoods like West Philadelphia and in wealthier areas such as the city center. Therefore, it's not accurate to assert that these markets are only found in low-income areas. Utilizing the smallest possible bounding box to define the study area

was inadequate, as it overlooked the absence of markets in North, Northeast, and South Philadelphia, leading to incorrect assessments of their distribution patterns. Beyond household income and population density, other societal elements should be taken into account when evaluating the placement of farmers markets. These factors include racial preferences, education levels, and local zoning laws. Different cultural groups may prioritize various aspects of food quality, and individuals with less education might not focus as much on maintaining a healthy diet.

We also must acknowledge that farmers markets aren't the only places offering fresh and nutritious food. Large supermarkets like Walmart and ShopRite also supply these products, which helps compensate for the lack of farmers markets in certain regions. In light of our findings that farmers markets in Philadelphia are concentrated in specific areas, we recommend that the government examine a range of factors before making policy decisions. If the goal is to introduce more farmers markets in underserved areas, it's essential to understand the local needs and the reasons for the current scarcity, while also considering alternative sources of healthy food available to the community.

Works Cited

Ceadserv1.nku.edu. 2023. Nearest Neighbor Analysis. [online] Available at:

<<http://ceadserv1.nku.edu/longa/geomed/ppa/doc/NNA/NNA.htm>> [Accessed 10

December 2023].

Desktop.arcgis.com. 2023. 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 10 December 2023].