SpAM

SpAM (Spatial Analysis and Methods) presents short articles on the use of spatial statistical techniques for housing or urban development research. Through this department of Cityscape, *the Office of Policy Development and Research introduces readers to the use of emerging spatial data analysis methods or techniques for measuring geographic relationships in research data. Researchers increasingly use these new techniques to enhance their understanding of urban patterns but often do not have access to short demonstration articles for applied guidance. If you have an idea for an article of no more than 3,000 words presenting an applied spatial data analysis method or technique, please send a one-paragraph abstract to rwilson@umbc.edu for review.*

Calculating Varying Scales of Clustering Among Locations

Ron Wilson
University of Maryland, Baltimore County

Alexander Din
Maryland Department of Housing and Community Development

The views expressed in this article are those of the authors and do not represent the official positions or policies of the State of Maryland.

Abstract

The Nearest Neighbor Index (NNI) is a spatial statistic that detects geographical patterns of clustered or dispersed event locations. Unless the locations are randomly distributed, the distances of either clustered or dispersed nearest neighbors form a skewed distribution that biases the average nearest neighbor distance used in calculating the NNI. If the clustering or dispersion of locations is moderate to extreme, the NNI can be inaccurate if the skew is substantial. Using Housing Choice Voucher program residential locations, we demonstrate in this article the method to derive an NNI based on a median and two quartiles that more accurately represents the midpoint of a set of nearest neighbor distances. We also demonstrate how to use these alternative point estimates to gauge multiple scales of clustering from different positions across the nearest neighbor distance distribution. Finally, we discuss how to use the average and standard deviation distances from the calculation of each NNI to more comprehensively gauge the scale of the geographic patterns. We also include a Python program that creates a randomized set of locations to calculate statistical significance for the median and quartile NNIs.

Voucher Residence Locations and Clustering

The Housing Choice Voucher (HCV) program enables low-income families to relocate to neighborhoods of their choice (HUD, 2012). A key objective of the HCV program is the deconcentration of families to select better neighborhoods in which to live and improve their lives (Winnick, 1995). A common concern about this relocation freedom is that HCV program participants will reconcentrate in the destination neighborhoods. Research shows that, after receiving assistance, voucher holders often relocate to neighborhoods similar to those in which they previously lived (Freeman and Botien, 2002; Huartung and Henig, 1997; McClure, 2010; McClure, Schwartz, and Taghavi, 2014; Metzger, 2014; Park, 2013; Pendall, 2000; Owens, 2017; Reece et al., 2010; Varady, Walker, and Wang, 2001; Varady et al., 2010; Wang, Larsen, and Ray, 2017; Wang, Varady, and Wang, 2008; Wilson, 2013; Zielenbach, 2015). Relocating to similar neighborhoods subverts the objective of the program, and voucher holders are little better off than they previously were. Therefore, housing authorities need to measure the degree of clustering or dispersion of HCV program participant residences to determine if the objective of deconcentration is being met.

A common measure of location concentration or dispersion is a nearest neighbor analysis using the calculation of the Nearest Neighbor Index (NNI). The NNI is a common global measure of clustering or dispersion, but its accuracy is vulnerable because it is based on an average. Event locations are typically concentrated, which skews the nearest distance distribution positive, because most of the locations are within close proximity to each other. The NNI will be based on a skewed distribution of distances. This problem is especially acute with voucher holders, who often live very close together because the geography of affordable housing stock puts them in the same multifamily building or neighborhood. Another limitation is that these neighborhoods vary in size, meaning that the scale of residential clustering will vary across a geography, from close-quarter environments of multifamily housing, to dense townhomes, to single-family homes with land. This change in scale is not something the standard NNI can take into account, because it is a point estimate for one position on the nearest distance distribution.

As such, a more reliable and multiscale measure must be used to determine the degree of HCV residence concentration. An inaccurate measure can report that voucher holders may be more or less concentrated than they really are, which would have adverse resource ramifications. For example, if the results show that voucher holders are more clustered than they are, then it may appear as though the program is not working and some other solution should be sought. Conversely, if the results show that voucher holders are more dispersed than they are, then it may appear the program is working and that it requires fewer resources.

We demonstrate a method to conduct a more robust nearest neighbor analysis by calculating median and quartile NNIs to overcome the limitations of the common NNI. The medians and quartiles are less susceptible than an average to outliers, and they provide more visibility into concentration patters at multiple spatial scales.

Nearest Neighbor Index

The NNI is an ordinal statistic that reports the existence and degree of clustering or dispersion of locations (geometric points). The NNI is a member of a family of cluster measuring statistics, which includes the more common Moran's *I* and Getis-Ord statistics. The NNI, however, is considered a distance analysis statistic because it strictly measures proximity between locations. In contrast, the Moran's *I* and Getis-Ord *G* statistics can measure proximity between locations not only with distances, but also with buffers around any location or adjacency when data are in areal form (polygon geographies).

Two key assumptions of any analysis involving nearest neighbors are that the sample locations are (1) all included within a finite geography and (2) unimpeded in occurring anywhere in that finite geography (Ebdon, 1985). Empirically, neither assumption ever holds for human or physical events. Related event locations often exist outside of a geography but are unavailable for measurement, leading to missing data that impact the calculation of the NNI statistic. More importantly, it is unrealistic that locations can occur anywhere unimpeded across a geography because other spatial processes either facilitate or prevent locations from occurring anywhere. Nevertheless, these two assumptions are necessary for testing if locations exhibit a clustering or dispersing pattern in order to provide a counterfactual geographic distribution of locations for comparison.[1]

To calculate the NNI, all nearest neighbor distances are summarized into an average. That summarization implies that a distribution of distances exists that can provide more point estimates about those minimum distances. For example, the standard deviation and percentiles can be used to determine patterns at several geographic scales beyond the average. Those estimates can be used as parameter specifications in identifying clusters, such as in kernel density estimation, Knox Test, SatScan, or the local Moran's *I* and Getis-Ord *G*. A key aspect of cluster analysis is determining the distance at which locations are no longer related to each other, that is, the distance at which spatial dependence between the locations ceases. Lacking any theoretical reason or empirical evidence to select that expected distance leaves one to guess what that distance may be. With average, standard deviation, and percentile nearest neighbor distances, the data can be a guide to setting an expected clustering distance.

Calculating the NNI starts with a measurement of the distance between each location to the nearest location. The minimum distance between each location and it nearest neighbor is first summed and divided by the total number of locations in the geography to derive an average minimum distance. The average minimum distance is—

$$\overline{d}(NN) = \frac{\sum_{i=1}^{N} Min(d_{ij})}{N} \quad , \qquad (1)$$

[1] Caution should be exercised in regards to comparing results from any nearest neighbor analysis to which any comparison of techniques should be done in the same study whether (1) between two different location types or (2) the same location type distributions across time. Comparing analyses between any two geographies confronts the problem that is the root of spatial statistics; that is, having to use randomization as the comparison distribution for significance of a clustering or dispersion pattern. The main problem is that each geography is unique in size and shape and will impact the distribution of event locations that directly affect the statistical results for each geography. Another consideration is that two entirely different distributions can have the same result.

where d is the minimum (Min) distance between location i and the closest location j, N is the total number of event locations, and $\bar{d}(NN)$ is the average minimum distance from measuring all nearest neighbors to each location. The average minimum distance is used in comparison—as the numerator—with a random (expected) distance to determine if the locations exhibit an overall pattern of clustering, dispersion, or random distribution.

The random distance represents an expected minimum nearest neighbor distance from which the locations are uninfluenced in that geography by social, economic, physical, or contextual activity—that is, the random locations from a distribution under complete spatial randomness (Cressie, 2015).[2]

The NNI, then, is the observed average minimum distance divided by the expected (random) average minimum distance to produce a relative ratio that is interpreted as an index along a clustering-to-dispersion continuum.

$$NNI = \frac{Observed\ \bar{d}(NN)}{Expected\ \bar{d}(NN)} \tag{2}$$

The NNI is interpreted in relation to a value of 1. Values around 1 indicate the observed distribution is random. Values less than 1 indicate clustering, with values closer to the floor of 0 indicating extreme clustering. An index of 0.0 means all the locations are in exactly the same place. Values greater than 1 indicate dispersion,[3,4] with values closer to the ceiling of 2.149 indicating extreme dispersion.[5] An index of 2.149 means that all the locations are exactly equidistant from each other in a systematic pattern. Exhibit 1 shows the relationship between NNI values and their corresponding patterns along the continuum from 0 to 2.149.

Three limitations hinder the NNI statistic from being more robust. The first is that the NNI is a global statistic; it cannot report where any local pattern of clustering or dispersion is in the geography, only that one of the patterns exists within.

The second is that the index is an average, subject to outliers pulling the observed mean away from its true location in the distribution. Outliers are typically present in location data because most social, economic, physical, or other processes produce clustered locations, with dispersed or random patterns seldom occurring. With a nearest neighbor analysis, a high frequency of short distances is produced with a small number of longer distances that skew the distribution positive.

The third limitation, which is related to the first, is that the average distance is only a single-point estimate for a set of nearest neighbor distances. Single-point estimates give only a partial insight

[2] See appendix A for details on the mechanics of calculating the NNI expected distance.

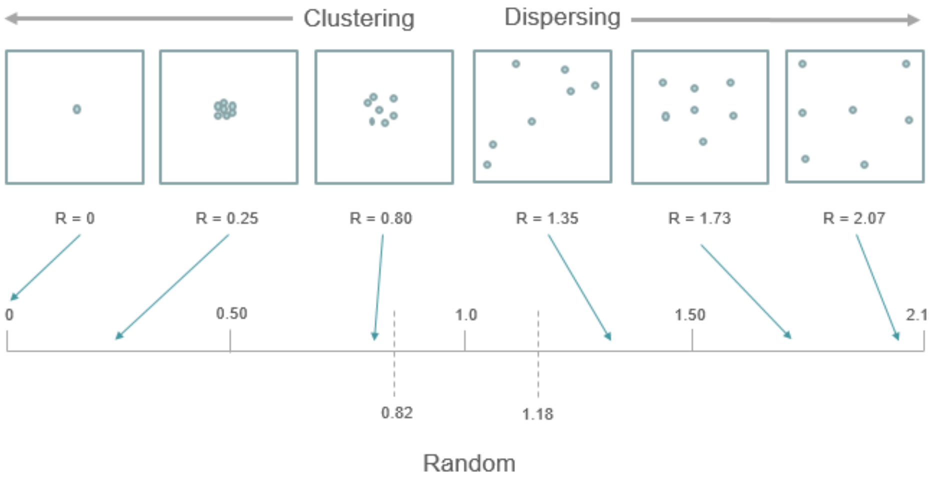
[3] Another way to think about the NNI is that the range of values reflects a progression from absolutely clustered (0, with all points in exactly the same location) to evenly dispersed (2.15, with all points maximally spaced from each other). This approach is useful when comparing indices to determine if one pattern is more or less clustered or dispersed than another pattern.

[4] The NNI can also be interpreted as a percentage more or less than the random distribution, because it is a ratio. For example, an NNI of 0.55 shows the observed nearest neighbor distances are 45 percent closer (1 – NNI = %) than the distances in the random distribution. An NNI of 1.67 shows the observed nearest neighbor distances are 67 percent farther (NNI – 1 = %) than the random distribution distances.

[5] The value 2.149 is the empirical ceiling of the NNI. Theoretically, a value could be higher, but none has been observed in previous research.

Exhibit 1

Range and Patterns of the Nearest Neighbor Index



R = ratio (between observed and expected).

into the spatial relationships between locations when varying scales of spatial relationships are likely in the geographic distribution. That is, different scales of clustering will likely exist across the geography due to variation in the structure of the environment. In this instance, neighborhoods where voucher holders live vary in scale, ranging from multifamily buildings, to dense inner-city blocks, to more spread out suburban, small town, or rural properties.

Nearest Neighbor Index Medians and Quartiles as Indicators of Multiple Spatial Pattern Scales

Most software packages that include the standard NNI technique report only the average nearest neighbor distance, which only allows for the assessment of geographic patterns at only one scale, the average distance between locations. Unless the standard deviation distance is also reported, any pattern variation at distances greater or less than the average is undetectable. However, even if the standard deviation distance is reported, highly clustered locations will have a skewed nearest neighbor distance distribution, and the one-standard deviation distance below the average will likely be less than 0 and useless for identifying any scale changes below the average. Calculating median and quartile distances, however, can reveal differing pattern scales at three locations along the distance distribution. The first scale is that of the densest locations, represented by the first 25 percent of nearest neighbor distances. The second scale is the moderately proximal locations, represented by one-half (50 percent) of the distances at the median. Finally, the third quartile NNI would be the more dispersed distances, represented by 75 percent of all locations, or the top 25 percent furthest distances.

If the scales were all similar, then the NNI would be the nearly the same for each of the point estimates, and the intervals between each index would also be similar. This result will suggest a distance distribution that is normal in shape with a high kurtosis. If the scales are different, then the NNIs will be far apart and the distribution of distances will be spread out, with the intervals between each index varying. This variation in intervals would indicate a distance distribution that is not normal in shape.

Lastly, the median and quartile NNIs can also measure the way in which the differing scales of clusters impact the average. Comparing the observed average with the median and quartiles reveals how far off the average is from the actual mean. Not only can the average distance be compared with the median to gauge how much they differ, but the average can also be compared with the quartiles to determine how far off the average is from different positions in the distance distribution. If the median and quartile NNIs are all below the average NNI, it indicates that the distribution is heavily skewed positive. If the first quartile and median NNIs are lower than the average, but the third quartile NNI is greater, it would indicate that distribution of distances is not too skewed, and the average distance may be acceptable in calculating the NNI, particularly if the average is closer to median NNI.

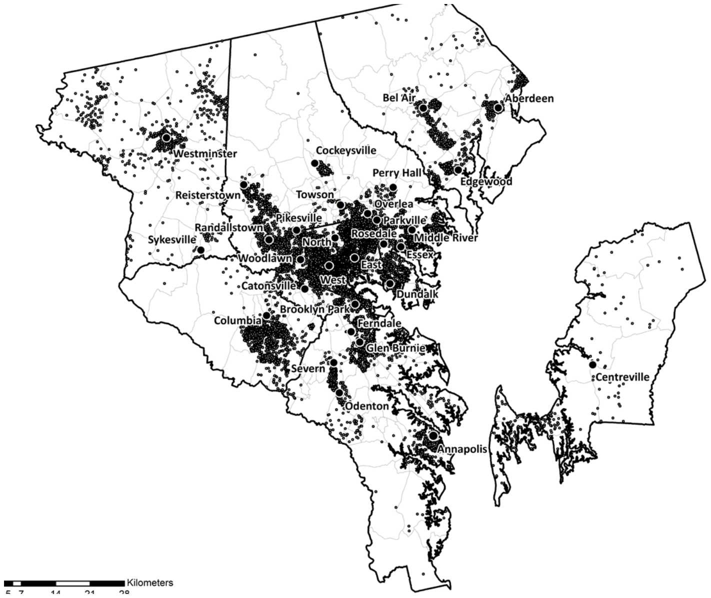
Data

The data used in this example are the counts of 2016 HCV program participants by census tract in the Baltimore-Columbia-Towson, MD Core Based Statistical Area (hereafter, Baltimore CBSA). The data were acquired via the U.S. Department Housing and Urban Development (HUD) enterprise geographic information systems (GIS) storefront. The total number of participant locations is 23,081. Because HUD does not provide program participant locations, we simulated the locations in a two-stage process to create residential locations based on where people actually live, with several being at the same location to emulate residents in the same apartment complex (exhibit 2).

First, the counts of program participants were divided according to the proportion of residents within each block group contained in each census tract. Second, a set of randomly distributed locations was created within each block group to simulate the locations based on known residential patterns. Some of these randomized locations were situated to be in the same coordinates to emulate voucher holders living in apartments or other multiunit residences. This two-stage process allows for a reasonable approximation of where HCV program participants live and reduces the risk of placing them in areas where populations do not reside (for example, forested portions, lakes, parks areas, and industrial sites).

Exhibit 2

Simulated Housing Choice Voucher Program Participant Locations Across the Baltimore CBSA



Baltimore CBSA = Baltimore-Columbia-Towson, MD Core Based Statistical Area.

Calculating the Nearest Neighbor Index Based on the Average Nearest Neighbor Distances

We first conducted a basic nearest neighbor analysis in CrimeStat IV to create the statistics for the NNI on the voucher holder locations in the Baltimore CBSA.[6] Exhibit 3 shows the results.

The results show the HCV program participant locations are moderately clustered, with an NNI of 0.58. This NNI value suggests that voucher holders are not too concentrated. An average distance of about 159 meters and standard deviation distance of about 835 meters suggest that voucher holders are quite spread out at a scale of several neighborhoods—about one-half mile. A standard deviation distance of about five times greater than the mean raises the concern that the NNI may be adversely affected from a skewed distribution of nearest neighbor distances. With geo-processing, we created a variable of nearest neighbor distances and examined the distribution to assess the accuracy of the reported NNI.

[6] We used CrimeStat IV because other GIS programs produce only a minimal listing of statistics, whereas CrimeStat IV provides much more valuable information that allows for more insight into the distance distribution. The nearest neighbor distance is typically all that is reported in most other GIS or spatial statistics programs, leaving the inability to complexly assess the scale of the clustering. With CrimeStat IV reporting the nearest neighbor standard deviation distance, the scale of clustering can be assessed, because it shows how far the locations are spread around the average distance. The standard deviation distances can be used in clustering routines to visualize the concentration of locations, and the distance can be used to gauge the number of blocks or neighborhoods the voucher holders actually cover.

Exhibit 3

Nearest Neighbor Statistics and Diagnostics

Descriptors of Nearest Neighbor Distances	Statistic/Diagnostic
Sample size	23,081
Measurement type	Direct
Mean nearest neighbor distance	158.60 meters
Standard deviation of nearest neighbor distance	834.76 meters
Minimum distance	0.00 meters
Maximum distance	438,731.80 meters
Area of geography (based on user input)	6,819,093,973.49 square meters
Mean random distance	271.77 meters
Mean dispersed distance	584.06 meters
Nearest neighbor index	0.5836
Standard error of random distance	0.94 meters
Test statistic (*Z*)	– 121.0332
p-value (one tail)	0.0001
p-value (two tail)	0.0001

Using the Near tool in ArcGIS, we calculated the distances of each location and its nearest neighbor and added them as a variable to the HCV location layer. The distance distribution showed an extraordinarily positive skew, with a skewness statistic of 14.65—a value far above the 0.50 threshold for skewness. Our analysis showed the NNI based on the average was very inaccurate and that deriving a median and quartile NNIs would prove worthwhile to improve our assessment of voucher holder concentration.

Calculating Median and Quartile NNIs

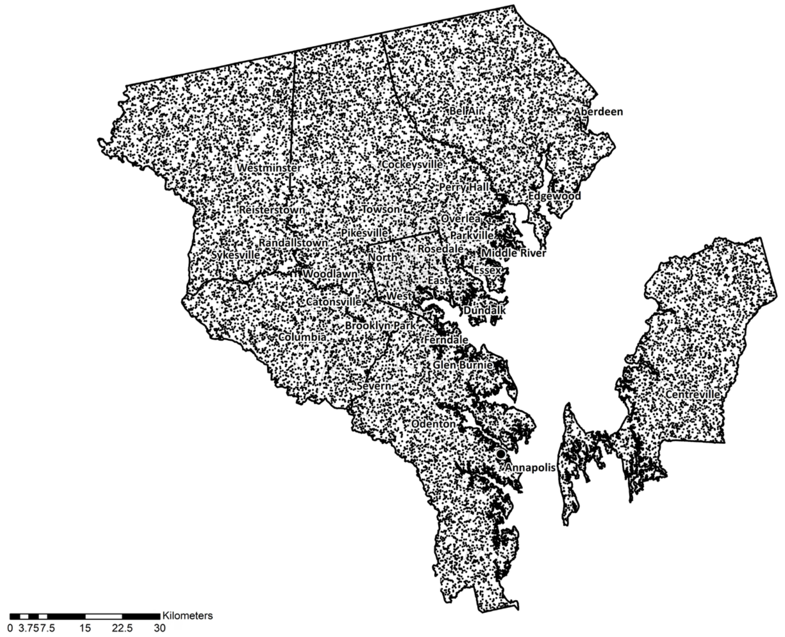
To create median and quartile NNIs, we first randomized 23,081 locations with 999 permutation trials within the Baltimore CBSA to create an expected distribution.[7] Randomizing the locations repeatedly is known as *bootstrapping* and builds an expected distribution against which to compare the observed statistics. The expected distributions represent the distances between locations if no social, physical, economic, or contextual process was influencing their placement. Bootstrapping produces a distribution from which a mean and standard error can be sampled for any point estimate, in this instance quartiles and the median. The mean of any point estimate from the 999 trials becomes the expected value against which to compare the observed statistics, with the standard error used to determine statistical significance of the observed quartile and the median NNIs.

Exhibit 4 shows an example geographic distribution from one permutation trail, in which HCV residences would be under complete spatial randomness with each voucher holder having equal

[7] Appendix B contains the Python code that randomizes the data set of 999 trials, including the output of the descriptive statistics for each trial.

Exhibit 4

Example of Random Permuted Voucher Locations Across the Baltimore CBSA



Baltimore CBSA = Baltimore-Columbia-Towson, MD Core Based Statistical Area.

probability of residing anywhere in the Baltimore CBSA.[8] We then repeated the geo-processing of nearest neighbor distances with the Near tool to create the random distribution of the 999 permutation trials and visualize the difference with the observed voucher holder nearest neighbor distances. The random distribution has an approximately normal shape, with a skewness value of 0.71.[9]

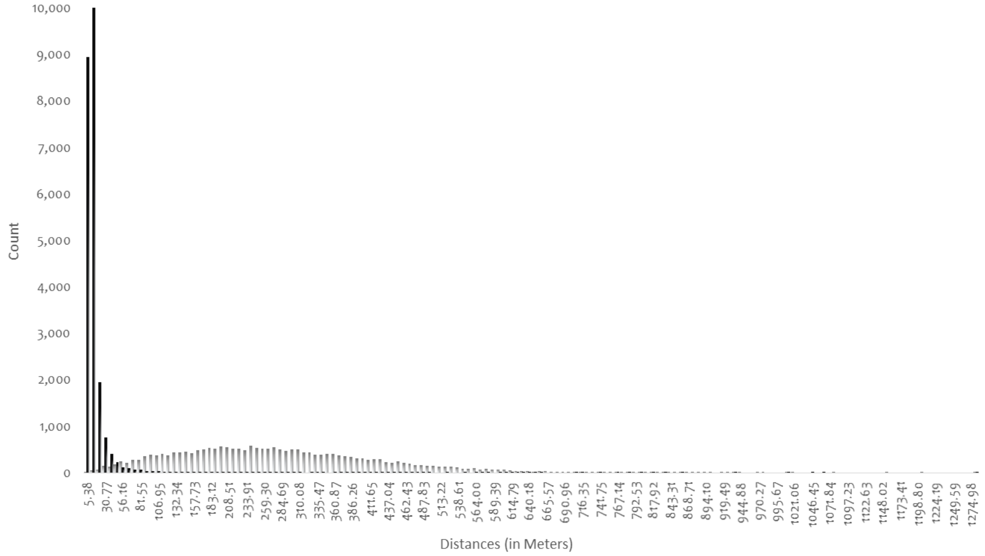
Exhibit 5 shows both the observed and randomized frequency nearest neighbor distance distributions for comparison.

[8] To make a more reasonable random distribution, the randomization process could be restricted to only geographies where voucher holders could live. The use of census tracts or block groups that show residential populations would comprise the area within which the randomization process would be distributed. The use of these geographies would give a more accurate expected average distance and standard error with which to compare the observed average and more precise NNI. Having the ability to randomize in a GIS allows for creating more realistic randomization processes. One of the fallacies of the randomization process under complete spatial randomness is that a location has equal probability of being anywhere in a geography, because nothing should prevent it from being anywhere. That may be likely for a physical process, but not for human settlement patterns. Only so many places exist in which voucher holders have an equal likelihood of residing, which would be places where affordable housing options are available, even if they are not likely to be housing options voucher holders can afford. However, that is where complete spatial randomness matters. If nothing prevents the voucher holders from residing in any housing unit, then randomizing across those units is more reasonable, because the voucher holders are not going to live where no housing exists at all. Therefore, the comparison is between the observed and the random places where a person could actually live. To be even more reasonable, only a certain percentage of housing units in a tract keeps the randomization process from locating a random location in just the areas with few housing units.

[9] Even a random distribution will produce some locations that are far apart, skewing the otherwise normal distribution positive.

Exhibit 5

Observed Versus Random Distances Between Nearest Neighbor Locations



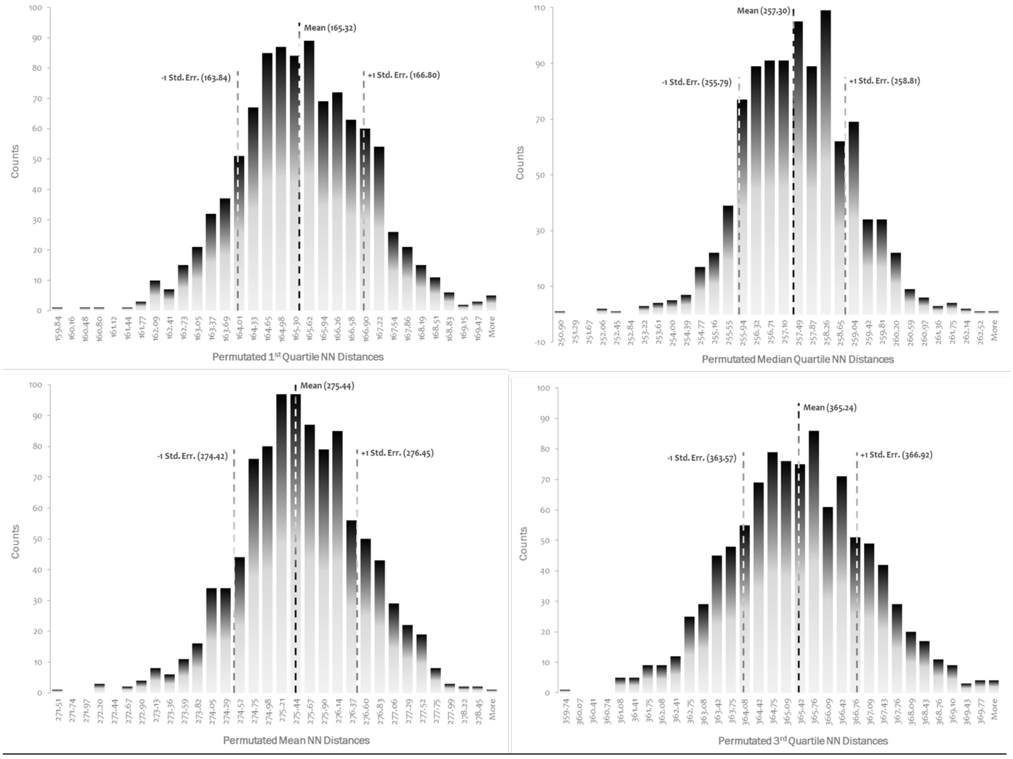
To ensure we calculated our 999 permutation trials correctly, we compared the random means and standard errors with results from the CrimeStat and ArcGIS formulas, noting that the differences were slight.[10] The comparison showed that our randomized trial results could be used to calculate the statistics for the median and quartile NNIs.

We then calculated the median and the quartile distances from the observed nearest neighbor distance distribution (exhibit 6), which were 0.00 meters for the first quartile (25th percentile), 16.20 meters for the median (50th percentile), and 118.82 meters for the third quartile (75th percentile). The average nearest neighbor distance showed to be greater than the third quartile of distances and nearly equivalent to the 82nd percentile of distances. Thus, the average distance used in the standard NNI calculation is representing a larger proximity scale—more spread out—between

[10] We compared the results with the NNI and z-score from CrimeStat and ArcGIS to compare the accuracy of our randomization process. The results between our randomization process and that of the software are similar enough, but not exact, because the two programs use a formula to estimate a standard error of a random distribution. CrimeStat and ArcGIS produced an average expected distance of 271.77 meters, and our randomization process produced 275.44 meters, a difference of 3.67 meters. CrimeStat produced a standard error of 0.97 meters, and ours was 1.01 meters, a difference of 0.04 meters. ArcGIS does not produce the standard error to allow for a comparison. We did a difference of means test comparing the CrimeStat formula results and the randomization results, which show the two are statistically different. This comparison shows that the formula is only an estimate that produces a result that is close enough but not as precise. Even though the formulas in CrimeStat and ArcGIS are reasonable approximations, they still are not as truly representative of a random distribution as permutation. Our results appear to back this conjecture, because the formulas produce a result that is close enough for practical purposes. Whether the average and standard errors from the formula or random process are used, the resulting NNIs will not be different enough to affect interpretation and will be identical if rounding to two decimal points to the right. Using the expected average nearest neighbor distance from CrimeStat and ArcGIS, the NNI is 0.584. The expected average distance from the randomization process is 0.576, a difference of 0.008. Nothing changes in the interpretation of the NNI, in that they both indicate a moderate level of clustering. Rounded to two decimal places to the right, as the NNI is often reported, both become 0.58, that is, identical.

Exhibit 6

Quartile, Median, and Mean Distance Distributions of Randomized Trials



NN = nearest neighbor. Std. Err. = standard error.

locations than ground truth. Although the NNI based on the average shows that voucher holders are clustered, for program evaluation purposes it shows them to be not nearly as concentrated as they really are.

With the statistics from the randomized trials, we calculated the average median and quartiles to create corresponding expected statistics for three NNIs, representing three different scales of location patterns (exhibit 7). The NNI for the first quartile is 0.00, median is 0.06, and the third quartile is 0.33. Each of these NNIs is less than the NNI based on the average, further showing that the NNI from the average distance is unreliable when a large number of geographic locations are in very close proximity.

To determine if our median and quartile NNIs were statistically significant, we used the standard errors of the median and quartiles from the 999 trials to calculate corresponding z-scores. The first quartile NNI of 0.00 has score of -111.63, the median NNI of 0.06 has a score of -159.34, and the third quartile NNI of 0.33 has a score of -146.91. All three NNIs are highly statistically significant. The intervals between the quartiles and the median are imbalanced, showing a change in clustering scales across the observed distance distribution. The difference between the first quartile and the median is 0.06, but the difference between the median and the third quartile is 0.27, showing that

Exhibit 7

Observed and Random HCVP Location Nearest Neighbor Index Statistics

	HCVP Location Distances (in Meters)			
	Observed	Random	NNI	*z*-score
Average	158.60	275.44	0.576	– 115.22
Standard deviation	740.03	1.01		
1st quartile (25th percentile)	0.00	165.32	0.000	– 111.63
Standard error	—	1.48		
Median (50th percentile)	16.20	257.30	0.063	– 159.34
Standard error	—	1.51		
3rd quartile (75th percentile)	118.82	365.24	0.325	– 146.91
Standard error	—	1.68		

HCVP = Housing Choice Voucher participant. NNI = Nearest Neighbor Index.

the 50 percent of voucher holders above the median are about 4.5 times more dispersed than the 50 percent of voucher holders below the median. This finding indicates the voucher holders are, indeed, clustered at different scales. The use of median and quartile NNIs, therefore, can provide more information about geographic patterns in the data.

With the nearest neighbor distances created from geo-processing, the scale of the clusters can be assessed with the percentiles around the quartile and median to evaluate the varying scales of clustering among the locations. Exhibit 8 shows the NNI values at +10 percentiles around the quartiles and the median are used to reveal the scales of clustering by depicting the spread around each point estimate, including the interquartile range.[11]

The interquartile range shows the clustering of locations in the upper outer quartile to be about 6.3 times more dispersed than those in the inner quartile. This outcome indicates that locations with distances below the median are clustered in very close proximity. This finding reveals that 50 percent of voucher holders likely live on the same block, and the other 50 percent likely live in neighboring city blocks.

The percentile ranges provide more detail about the scale of clustering at each of the quartile and median distances. For the first quartile, the range of 0.00 meters between the 15th and 35th percentiles of nearest neighbor distances reports that the scale of clustering for 35 percent of voucher

Exhibit 8

Cluster Scales

	Percentile and Quartile Distance Ranges (in Meters)										
Ranges	*Min*	*15th Pctl*	*1st Qrtl*	*35th Pctl*	*40th Pctl*	*Median*	*60th Pctl*	*65th Pctl*	*1st Qrtl*	*85th Pctl*	*Max*
	0.00	0.00	0.00	0.00	4.05	16.20	39.26	57.55	118.82	187.34	23,444.30
Percentile			0.00			35.22			129.80		
Inter-quartile				*Lower* 16.20				*Upper* 102.62			

Max = maximum. Min = minimum. Pctl = percentile. Qrtl = quartile.

[11] Any percentile range can be used to examine the ranges around the quartile and median NNI.

holders is that of the same building or complex; the minimum is also 0.00 meters and indicates that two or more locations are in the exact same place. At the median, the range of nearest neighbor distances between the 40th and 60th percentile is 35.22 meters, indicating that 20 percent of voucher holders live on blocks with dense housing. Finally, between the 65th and 85th percentiles of nearest neighbor distances, 20 percent of voucher holders live within 129.80 meters of each other and are likely on contiguous blocks in a neighborhood. However, after the 85th percentile of nearest neighbor distances (187.3 meters), the voucher holders spread out substantially and are isolated from others given they are up to 23,444.30 meters away from the nearest voucher holder.

Summary and Extensions of the Interquartile NNIs

With highly clustered locations, we showed that the average nearest neighbor distance proves to be an inaccurate base to calculate the NNI. When using the statistic to help in assessing a program's performance, such as the HCV program, an alternative measurement must be used. The median and quartile NNIs not only provide more accurate results of the geographic pattern of clustering, but the statistics also provide more information about changes in clustering at differing geographic scales. In this example, the NNI based on the average distance inaccurately showed that voucher holders are only moderately concentrated, when they are actually far more clustered. Using the median and quartile NNIs, however, revealed that at least 25 percent of voucher holders were highly concentrated at the same location, with 26 to 75 percent likely living together in small neighborhoods. Given the nature of housing availability for HCV program participants, future analyses of the program will likely need to be analyzed with the median and quartile NNIs.

We offer a final thought about the geographies used to randomize when calculating any NNI. Typically, randomization is either implemented in the permutation process or is estimated with a formula that uses the area of the geography in which the locations occur. The use of the entire geography is based on the assumption that the locations can be equally likely to occur anywhere within that boundary. This assumption is unrealistic, because locations do not have equal probability of being anywhere, which is due to physical and human influences on a geography that restrict the occurrence of locations. We suggest running a second analysis limiting the geography to only areas in which the locations can actually occur. With voucher holders, these geographies would only be the areas in which rental housing is available. The boundary of the limited geography can be used in the permutation process with GIS as an alternative, by identifying all the areas that the analyzed locations can actually have the opportunity to occur.

Appendix A: Formula for Calculating an Expected Nearest Neighbor Distance

A random (expected) distance is generated by one of two randomization methods. The first method—which is rarely implemented in software—is to randomly distribute the same number of observed locations within an area that is either the size of their minimum bounding rectangle or within the study geography.[12,13] This randomization process is known as permutation, by which the observed data are used to generate a counterfactual (expected) distribution for what would occur in that unique geography. The second method uses a formula to approximate the Monte Carlo process, which is—

$$\bar{e}(NN) = 0.5\sqrt{A/_{N}} \tag{3}$$

where $\bar{e}(NN)$ is the expected average distance, A is the total from the study geography in which the locations occur, and N is the total number of locations. The ratio produces a density, of which the square root is taken to produce a linearized value.[14] The constant 0.5 is multiplied to the linearized density ratio to rescale it and prevent the expected average distance from being larger than the study geography.

Appendix B: Python Code for Creating the Randomized Locations

```
# -*- coding: utf-8 -*-
"""
#This file creates random points in a given geography then takes statistics of distances of
the  ## nearest points.

## Author: Alex Din
"""
import arcpy
import csv
import numpy  as np
import os
from   random import randint
import time
##
start_time = time.time()
## this needs to be 1000 in order to get 999 iterations because of the range loop starts
at 1, not 0
iterations = 1000
## the number of random points to be created each iteration
pointsNum  = 23081
## the prjArea is the project area area geography which is the Baltimore CBSA in this
example
prjArea    = r"C:\Path\to\the\feature\class\for\the\project\geography"
## the csvName is the name of your output file, it MUST have '.csv' appended to the string
csvName    = "Baltimore_CBSA.csv"
## workspace is the geodatabase where functions will be performed
workspace  = r"C:\Path\to\the\working\geodatabase.gdb"
## dirspace is where your csv will be written, the directory must already exist prior to
running the script
dirspace   = r"C:\Path\to\the\working\directory"
## csvPath is the combination of the csvName and dirspace for outputing the final CSV
csvpath    = os.path.join(dirspace,csvName)
```

[12] Randomizing observed data is known as permutation.

[13] A minimum bounding rectangle is the outermost boundary of the furthest locations in each Cartesian plane orthogonal direction.

[14] The square root is taken to transform the two-dimensional density ratio into a one-dimensional distance so that it is geometrically comparable with the observed distance.

```
## the name of the random points
ptsName    = "samplepoints"
## the random number is used to export a random set of points for visualization purposes
randomNumb = randint(1, iterations)
## the work environments are set
arcpy.env.overwriteOutput = True
arcpy.env.workspace = workspace
os.chdir(dirspace)
## the print statement informs the user which
print "Iteration %s will be exported as a random copy for map purposes" %(randomNumb)
##
for number in range(1,iterations):
    print "Processing number %s" %(number)
    small_time = time.time()
    # create a set of random points within the project area
    try:
        arcpy.CreateRandomPoints_management(workspace, ptsName, prjArea, "", pointsNum,
"", "POINT", "")
    except Exception as e:
        print e
    # compute nearest neighbor distance
    try:
        nearValueList = []
        arcpy.Near_analysis(ptsName, ptsName, "", "NO_LOCATION", "")
        with arcpy.da.SearchCursor(ptsName,["NEAR_DIST"]) as cursor:
            for row in cursor:
                nearValueList.append(row[0])
        nearValueList.sort()
        # print the values in the IPython console to inspect while processing
        p25  = round(np.percentile(nearValueList, 25), 2)
        p50  = round(np.percentile(nearValueList, 50), 2)
        p75  = round(np.percentile(nearValueList, 75), 2)
        mean = round(np.mean(nearValueList), 2)
        std  = round(np.std(nearValueList), 2)
        var  = round(np.var(nearValueList,ddof=1), 2)
        maxx = round(np.max(nearValueList), 2)
        minn = round(np.min(nearValueList), 2)
        del cursor, row
    except Exception as e:
        print e
    # get the time it took to run just this one iteration
    small_time_end = time.time()
    small_elapse   = round((small_time_end - small_time),2)
    print "This iteration took %s seconds" %(small_elapse)
    # log the information to a CSV file
    # if the CSV does not yet exist, the CSV will be created with headers and append the
first iteration of data
    # else, if the CSV does exist, the information will be appended to a new row
    try:
        headRows = ["Number", "Seconds","25P", "Median", "75P", "Mean","STD", "Variance",
"Maximum","Minimum"]
        dataRows = [number,small_elapse,p25,p50,p75,mean,std,var,maxx,minn]
        if not os.path.exists(csvpath):
            with open(csvName, 'wb') as f:
                wtr = csv.writer(f, delimiter= ',')
                wtr.writerow(headRows)
                wtr.writerow(dataRows)
        else:
            with open(csvName, 'ab') as f:
                wtr = csv.writer(f, delimiter= ',')
                wtr.writerow(dataRows)
    except Exception as e:
        print e
    del f, wtr
    # if the iteration matches the random number, export the sample data set for visual-
ization purposes
    try:
        if number == randomNumb:
            out_name = "%s_random_%s" %(ptsName,number)
            arcpy.FeatureClassToFeatureClass_conversion(ptsName,workspace,out_name)
    except Exception as e:
        print e
    print("-------------------------------------------------------------------")
##
end_time = time.time()
print("Total time elapsed was %g seconds" % round((end_time - start_time),2))
```

Acknowledgments

The authors thank Danielle Wilson from the University of Maryland, Coro Chasco of the Universidad Autónoma de Madrid, Julia Koschinski of the University of Chicago, and Jay Lee of Kent State University for reviewing this method and for providing valuable comments toward improving the manuscript.

Authors

Ron Wilson is an adjunct faculty member of the Geographic Information Systems Program at the University of Maryland, Baltimore County.

Alex Din is a housing research and GIS analyst with the Maryland Department of Housing and Community Development.

References

Cressie, Noel. 2015. *Statistics for Spatial Data*, revised ed. New York: Wiley.

Ebdon, David. 1985. *Statistics in Geography Second Edition: A Practical Approach*. Malden, MA: Blackwell Publishing.

Freeman, Lance, and Hilary Botein. 2002. "Subsidized Housing and Neighborhood Impacts: A Theoretical Discussion and Review of the Evidence," *Journal of Planning Literature* 16 (3): 359–378.

Hartung, John M., and Jeffrey R. Henig. 1997. "Housing Vouchers and Certificates as a Vehicle for Deconcentrating the Poor: Evidence From the Washington, D.C. Metropolitan Area," *Urban Affairs Review* 32 (3): 403–419.

McClure, Kirk. 2010. "The Prospects for Guiding Housing Choice Voucher Households to High Opportunity Neighborhoods," *Cityscape* 12 (3): 101–122.

McClure, Kirk, Alex F. Schwartz, and Lydia B. Taghavi. 2014. "Housing Choice Voucher Location Patterns a Decade Later," *Housing Policy Debate* 25 (2): 215–233.

Metzger, Molly W. 2014. "The Reconcentration of Poverty: Patterns of Housing Voucher Use, 2000 to 2008," *Housing Policy Debate* 24 (3): 544–567.

Owens, Anne. 2017. "How Do People-Based Housing Policies Affect People (and Place)?" *Housing Policy Debate* 27 (2): 266–281.

Park, Miseon. 2013. "Housing Vouchers as a Means of Poverty Deconcentration and Race Desegregation: Patterns and Factors of Voucher Recipients' Spatial Concentration in Cleveland," *Journal of Housing and the Built Environment* 28 (3): 451–468.

Pendall, Rolf. 2000. "Why Voucher and Certificate Users Live in Distressed Neighborhoods," *Housing Policy Debate* 11 (4): 881–910.

Reece, Jason, Samir Gambhir, Craig Ratchford, Matthew Martin, Jillian Olinger, John A. Powell, and Andrew Grant-Thomas. 2010. *The Geography of Opportunity: Mapping To Promote Equitable Community Development and Fair Housing in King County, WA*. Columbus: The Ohio State University, Kirwan Institute for the Study of Race and Ethnicity. http://kirwaninstitute.osu.edu/docs/KingCounty.pdf.

U.S. Department of Housing and Urban Development (HUD). 2012. "Public and Indian Housing Tenant-Based Rental Assistance: 2012 Summary Statement and Initiatives." https://www.hud.gov/sites/documents/TENANT_BR_ASSIS_2012.PDF.

Varady, David P., Carole C. Walker, and Xinhao Wang. 2001. "Voucher Recipient Achievement of Improved Housing Conditions in the US: Do Moving Distance and Relocation Services Matter?" *Urban Studies* 38 (8): 1273–1304.

Varady, David P., Xinhao Wang, Yimei Wang, and Patrick Duhaney. 2010. "The Geographic Concentration of Housing Vouchers, Blacks, and Poverty Over Time: A Study of Cincinnati, Ohio, USA," *Urban Research & Practice* 3 (1): 39–62.

Wang, Ruoniu, Kristin Larsen, and Anne Ray. 2017. "Rethinking Locational Outcomes for Housing Choice Vouchers: A Case Study in Duval County, Florida," *Housing Policy Debate* 25 (4): 715–738.

Wang, Xinhao, David Varady, and Yimei Wang. 2008. "Measuring the Deconcentration of Housing Choice Voucher Program Recipients in Eight U.S. Metropolitan Areas Using Hot Spot Analysis," *Cityscape* 10 (1): 65–90.

Wilson, Ron. 2013. "Using Near-Repeat Analysis To Measure the Concentration of Housing Choice Voucher Program Participants," *Cityscape* 15 (3): 307–318.

Winnick, Louis. 1995. "The Triumph of Housing Allowance Programs: How a Fundamental Policy Conflict Was Resolved," *Cityscape* 1 (3): 95–121.

Zielenbach, Sean. 2015. "Moving Beyond the Rhetoric: Section 8 Housing Choice Voucher Program and the Lower-Income Urban Neighborhoods," *Journal of Affordable Housing & Community Development Law* 16 (1): 9–39.