A Point Simulation Approach for Evaluating the Cluster Detection Capabilities of Local Network Statistics

TBD

July 16, 2010

Abstract

Despite advancements in spatial statistics for network-based phenomenon in recent years, several questions about these methods persist. One of these questions concerns the impact of the distribution of events on the cluster detection capability of these statistics. This study will evaluate the cluster detection power of local network statistics for a variety of simulated point patterns. Specifically, the network kernel density, the local k-function, and network versions of the local Moran and local G statistic will be evaluated under a variety of event distributions. Results demonstrate.....

1 Introduction

One of the simplifying assumptions made in studies of spatial phenomenon is that events are distributed on an isotropic plane in Euclidean space. This simplification ignores the fact that some phenomenon like traffic accidents, retail outlet locations, certain plant species, and certain kinds of crime, are more likely to occur on or near road networks, and thus behave differently than if they were distributed in Euclidean space. Critiques of the application of planar methods to network-constrained phenomenon cite several reasons why this application is inappropriate. One reason offered is that network events actually occur in a subset of continuous two-dimensional space (Yamada and Thill, 2004; Borruso, 2008) and therefore it is incorrect to treat them as continuous. Another related critique of the uniform use of planar methods on network-based phenomenon is the fact it is physically impossible for some processes to occur off of the street network because of physical or legal constraints (Yamada and Thill, 2004). Finally, the use of planar methods in place of network methods confuses the definitions of proximity and accessibility. Two crash locations on a network may be spatially proximal, but the route on a street network a rescue vehicle would travel between the two crash locations may be dramatically different than their distance in Euclidean space (Yamada and Thill, 2004).

In response to these critiques, statistical analyses of network-constrained phenomenon have evolved from the application of planar methods to network based phenomenon (Black, 1991; Levine et al., 1995; Miller, 1994; Okabe and Kitamura, 1996; Kwan, 1998; Black and Thomas, 1998; Okabe and Okunuki, 2001) to the development of network specific statistics for network-based phenomenon. The development of network-specific methods recognizes the necessity for different treatments of network events compared to their Euclidean-space counterparts. Innovations in network statistics include network Huff models (Miller, 1994; Okabe and Kitamura, 1996; Okabe and Okunuki, 2001), network based kernel density methods (Porta et al., 2009; Borruso, 2008; Xie and Yan, 2008; Okabe et al., 2009), local indicators of spatial association (Yamada and Thill, 2010), network k-functions (Okabe and Yamada, 2001; Lu and Chen, 2007), local network k-functions (Yamada and Thill, 2007), and network quadrat analysis (Shiode, 2008). These statistical innovations are particularly important

given the inherent differences in network vs. Euclidean phenomenon. In fact, studies evaluating network-based methods demonstrate these methods produce visually different results than their planar counterparts (Yamada and Thill, 2004; Cheng and Washington, 2005; Downs and Horner, 2007).

Despite this visual demonstration of differences in the results produced by network and planar methods, the circumstances under which these methods are likely to diverge remains unexplored in the literature, and has been recommended as an area for future research (Lu and Chen, 2007; Borruso, 2008). The existing literature suggests differences in these methods are likely to occur when the characteristics of the underlying event distribution and the characteristics of the network vary (Yamada and Thill, 2004). Specifically, it is anticipated the degree of divergence between network and planar methods is likely to be smaller for highly connected street networks and dense points distributions (Lu and Chen, 2007)because this kind of distribution "*fills in*" or approximates Euclidean space. The anticipated difference between the two methods is expected to be greater for less well connected street networks and more sparse point distributions (Lu and Chen, 2007). More importantly, the impact of the characteristics of street networks and point distributions on network-based statistics is not well understood.

This study is a first step in exploring the circumstances under which the results produced by planar and network methods are likely to diverge. It proposes to evaluate the impact of varied event distributions on the statistical power of network-based statistics. In the analysis that follows, a simulation will be designed that evaluates the power of three network-based statistics for a variety of simulated point patterns, holding the features of the street network constant. Specifically, the cluster detection capabilities of the network kernel density function (Okabe et al., 2009), the network local k-function (Yamada and Thill, 2007), and the network local indicator of spatial association (Yamada and Thill, 2010) will be evaluated.

The remainder of this paper is organized as follows. Section two provides an overview of network analyses from a variety of disciplines including geography, mathematics, ecology, criminology, and marketing. Section three provides a brief review of power evaluations from the cluster literature, which are the basis for the simulation design in this study. Section four discusses the street network data and the simulated event data used in this study. Section

five provides an overview of the simulation design, including the distribution of events under the null and alternative hypothesis. Section six contains a description of the three methods evaluated in this study while section seven discusses the results of the simulation and the cluster detection capabilities of the network statistics of interest. The study concludes in section eight with a discussion of extensions to this simulation and the practical implications of the results produced in this study.

2 Network Constrained Phenomenon

Although techniques, like the shortest-path algorithm, (Dijkstra, 1959) used to estimate network based statistics are rooted in mathematics and graph theory, the analysis of network-constrained events is relevant to a variety of disciplines including geography, ecology, epidemiology, communications studies, and engineering. For example, network-based statistics may be used to evaluate the spread of a contagious disease along public transportation routes. These statistics are also useful for plants found growing along transportation networks, like the Acacia plant (Spooner et al., 2004). Network-based market area demand models have also been utilized in the retail location literature (Okabe and Okunuki, 2001). Perhaps one of the most common applications of these statistics is to examine clusters of traffic accidents on road networks (Yamada and Thill, 2004) (Okabe et al., 2009).

While the utility of network-based methods is certainly widespread, the accuracy of these approaches over their planar counterparts to date is unclear, as are the kinds of data for which these statistics are most useful. These methods are certainly useful for data that are network constrained, but their utility for events that occur near networks or conform mostly, but not entirely to networks, is rather nebulous. Within criminology for example, the usefulness of the application of network statistics to events less constrained to street networks like burglary and robbery, as opposed to motor vehicle thefts, remains unclear. Yamada and Thill (2007) provide some clues about the utility of these statistics via a typology based on the movement and location constraints of various phenomena. However, this is certainly an open research question in the spatial statistics literature.

In addition to uncertainty regarding the variety of data for which these statistics are

useful, the impact of variations in the street network and the relative clustering of events of interest on the cluster detection power of these statistics remain unexplored in a simulation context. Variations in the attributes of street networks, such as street network length, street length density, number of street segments, and street segment density are mentioned in the literature (Yamada and Thill, 2004; Lu and Chen, 2007), however, the impact of other street network characteristics (connectivity, redundancy, and nodal accessibility) on the power of network-based statistics remain open areas for future research. The simulation analysis in this paper will evaluate the impact of varied point distributions on the cluster detection capabilities of three network statistics: the network kernel density function, the network local k-function, and the network LISA. The simulation analysis will vary the number and strength of clusters on the street network, and the ability of these statistics to detect clusters analyzed.

3 Cluster Statistic Simulations

The basis for the simulation design in the present analysis is derived from the literature evaluating the statistical power of various cluster statistics. These studies evaluate a subset of over 100 test statistics that have been proposed to evaluate the deviation of point patterns from a random distribution (Kulldorff, 2006). These studies may be subdivided in a number of ways based on the kinds of cluster statistics evaluated (global or local), the type of inputs into the simulation (empirical or simulated data), and the number of clusters used in the power evaluation. The goal of this study is to evaluate the ability of the three network statistics mentioned (network kernel density, network local k-function, network LISA) to accurately detect the number of clusters simulated. Therefore, studies evaluating the power of local or focused cluster statistics via simulated datasets are of particular interest for the purposes of this study. Table 1 contains the sample of studies considered in the simulation design.

Two aspects of the simulation designs contained in these studies were of particular interest for the design of the simulation in the present study. One, the number of clusters evaluated in the study and the manner in which these clusters were created. Two, the manner in

which the strength of these clusters, or the relative risk, was determined. For example, Tango (1999) and Swartz (1998) evaluated the power of test statistics based on the location of a single cluster, while Aldstadt and Getis (2006) used multiple clusters to evaluate the capabilities of their AMOEBA statistic. A common feature of these studies is that the locations of these clusters as well as their relative strengths are typically specified a priori. The Kulldorff and Nagarwalla (1995) study is a good example of this. They placed a cluster in the center of a square of 100 cells and then distributed the disease cases amongst the cells so that the cell representing the "true cluster" had a relative risk that was rr higher than all of the other cells. In a similar fashion, Rogerson (1999) elevated the risk of disease at a single point where the relative risk was defined to be three times higher than all of the other locations. The relative risk then declined exponentially from this designated point and locations with a resultant relative risk less than 1.5 assigned a risk value of one.

Additional analysis of the studies in Table 1 in this context reveals a great deal of variety in simulation designs. As mentioned previously, a common feature of these simulations is an a priori designation of the cluster center/s and a designation of relative risk throughout the study area and at the cluster locations. The simulation design in this study is loosely based on these power evaluation s because of its a priori designation of cluster locations and strengths. However, this study is different from these simulation studies and other network studies in two key ways. First, the simulation is designed for clusters located on a network and not clusters located in Euclidean space. Second, the simulation assigns events directly to the network and does not require observations be snapped to the network as do other studies of network-constrained phenomenon (Borruso, 2008; Xie and Yan, 2008; Okabe et al., 2009)

4 Data

Street network data utilized in this study pertain to the city of Mesa, which is located to the east of Phoenix, Arizona. Mesa is a medium-sized city in the Phoenix metropolitan statistical area (Figure 1) These data were obtained from the GIS Data Repository of Arizona State University (2010). The data points assigned to the street network are a function of the simulation design, which will be discussed in the next section.

5 Simulation Design

The simulation in this paper opts for the use of simulated rather than empirical data. This selection is made because the use of simulated data allows one to know the true data generation process, and the type of clustering present in the data (Kulldorff et al., 2006). This is not the case with real data and therefore it is more difficult to evaluate the clustering process present in these data (Kulldorff et al., 2006) and the subsequent statistical power of the statistics of interest. Before discussing the details of the simulated data in the present study however, it is necessary to discuss some of the necessary pre-processing of the street network in preparation for the simulation.

5.1 Preparation of Street Network Information

Although the street network used in the study is not simulated, it was necessary to perform two operations on the dataset to prepare it for use with the simulated dataset. First, it was necessary to ensure the length of the street is calculated for each component street of the larger network. Second, it was necessary to assign each street a unique identifier so that the streets may be ordered, and the simulated points assigned accordingly. Once these initial calculations are made, it is possible to assign each street a probability based on its length, in proportion to the total length of all the streets in the network. The steps to do this are as follows:

Step 1: Obtain the total length of all the streets in the network.

Step 2: Calculate the *proportional length* of each street. This is obtained by dividing the length of the street by the total length of all the streets in the network.

Step 3: Assign a probability to each street segment based on its cumulative proportional length. This is done by ordering the streets according to their unique identifiers, and then calculating the cumulative proportional length for each street. For example, the cumulative proportional length for the street with a unique identifier of three is the sum of the proportional lengths of streets one, two, and three. To obtain the probability for street 100, the proportional lengths of streets 1 through 99 are added. This step results in streets laid end to end, according to their unique identifier, on an imaginary line starting at 0 and ending at

1, with a corresponding probability for each equal to their proportional length.

At the end of the preparation of the street network data, each street segment will have the following key attributes: a length, a unique id, a proportional length, and a probability. The first three attributes in this list were key inputs to obtain the probability of each street. The probability assigned to each street will be the means by which points are matched directly to the network.

5.2 Null Hypothesis

The distribution of observations under the null hypothesis is random. 1,500 random numbers between 0 and 1 are generated from a uniform distribution. The random numbers are then matched to the streets in the network according to their correspondence with the calculated probability of the street. After matching events to the streets, the events are then located on that street in proportion to their distance on the imaginary 0 to 1 line generated by laying all the streets in the network end to end. For example, if an event with a probability value of 0.70 is matched to a street that begins at a point with a probability value of 0.60 and terminates at a point with a probability value of 0.75, that point consumes 10% of the length of that particular street. Accordingly, that point with a probability value of 0.70 should be placed on its corresponding street so that is consumes approximately 10% of the total length of that street. A count of the number of events on the street concludes the distribution of points under the null hypothesis.

Although the additional step of placing points in specific locations on the component streets of the network is not important in the context of the LISA results produced in this paper, it is important in the estimation of the network kernel density function and the network local k-function. While the network LISA statistics merely rely upon counts of events on the component streets of the network, both the network kernel density and the local k-function operate on individual points. Therefore, it is paramount to locate events at specific points on the streets instead of stacking them in the center of the streets; such an approach would not provide any information about the density of events on the individual streets of the network.

5.3 Alternative Hypothesis

The process for generating the distribution of points under the alternative hypothesis, which allows for multiple clusters of points, is somewhat similar to the process for distributing the points under the null hypothesis, with a few important exceptions. First, it is necessary to identify the locations and the relative strengths of the clusters, per the precedent set by power evaluations of planar cluster statistics discussed in the previous section. To do this, the street network of interest is divided into four quadrants and a street segment selected in each of these quadrants to serve as the cluster center. Second, it is necessary to distribute some percentage of the total points in the simulation (1500) amongst these four clusters. To begin, 10% percent of the 1500 points to be distributed are allocated evenly amongst the four clusters. The remaining 1350 points are distributed randomly across the street network.

The random distribution of the remaining 1350 events follows the same process as for the null hypothesis described above. However, it is necessary to remove the four streets serving as cluster centers from the set of streets that will randomly receive additional events from the set of 1350 to be assigned. After these four streets are removed from consideration, the remaining street segments are reassigned unique ids and steps one through three for the preparation process of the street network described above repeated. This is necessary to rescale the probabilities of the remaining street segments so the sum of their probabilities equals one. After the remaining street segments are assigned new probabilities, the remaining 1350 points are randomly matched to and located on the street segments according to the process outlined for the null hypothesis above.

Modifications to the locations of the four clusters and the strengths of the four clusters are considered for the alternative hypothesis. The locations of the four clusters are varied so that four scenarios are considered:

Scenario 1: One cluster per quadrant. *Scenario 2*: One cluster per quadrant and two clusters in quadrant I. *Scenario 3*: Two clusters in quadrants I and III. *Scenario 4*: Three clusters in quadrant II and one cluster in quadrant III.

In addition to considering each of the four scenarios detailed above, five different cluster strengths will be considered for each of the scenarios. The cluster strength in the alternative

hypothesis corresponds to the percent of all observations assigned to clusters, where the larger the proportion of observations assigned to the clusters corresponds to stronger clusters. Proportions of 10%, 20%, 30%, 40%, and 50% will be considered for each of the four scenarios for a total of 20 different combinations of cluster locations and cluster strengths.

6 Methods

The cluster detection powers of three kinds of network-based statistics are evaluated in this paper: the network kernel density function, the network local k-function, and the network versions of the local Moran (Anselin, 1995) and the Getis and Ord local G statistic (Getis and Ord, 1992). The principle difference between these methods and their planar counterparts is the use of network distance rather than Euclidean distance in the computation of results. The incorporation of network distance however does present present challenges, especially as the complexity of the network increases.

6.1 Network Kernel Density Estimation Method

The specification of the network kernel density function is based on the formulation developed by (Okabe et al., 2009). This kernel density estimation method is discontinuous at each node on the network and assigns the value of the kernel function to all links equally at each node along the shortest path (p) from location x to location y (Okabe et al., 2009). The equal-split non-continuous kernel function specified in this paper is used for two reasons. One, other network kernel function options, like the similar-shape kernel function, are biased (Okabe et al., 2009). Two, of the unbiased kernel functions mentioned in the study, the computational and practical considerations recommend the non-continuous version of the equal-split kernel function over the continuous version. Not only does implementation of the continuous version of the equal-split function require the consideration of many special cases, but it also requires more time to compute the results. Okabe et al. (2009) found the computation time for the continuous version of the equal-split function increased rapidly for bandwidths greater than 200m.

6.2 Network Local k-function

The network local k-function developed by (Yamada and Thill, 2007) is based on the planar local k-function developed by Getis and Franklin (1987) and the representative location approach utilized by (Openshaw et al., 1987) in their Geographical Analysis Machine (GAM) analyses. The statistic computes the number of observations within a pre-specified network distance h of a reference point (Yamada and Thill, 2007) and is considered a departure from its planar counterpart because it examines clustering around reference points and not even locations (Yamada and Thill, 2007). It is important to note that unlike the network local Moran and local G statistic discussed in this piece, the network local k-function is not a local indicator of spatial association (LISA) because of its use of reference points and not actual event locations to detect local clusters; this prevents the sum of local indicators from summing to the global k-function, proportional to some constant. (Yamada and Thill, 2007)

6.3 Network Local Indicators of Spatial Association (LISAs)

The network versions of the local Moran (ILINCS) and the local G statistic (GLINCS) [1] used in this paper, are extensions of the LINCS (local indicators of network-constrained clusters) framework developed in Yamada and Thill (2007). The computation of these statistics is much the same as for their planar counterparts. However, an important difference is the incorporation of a network-based weights matrix in the calculation of the two statistics. This weights matrix, in its simplest form, may reflect the connectivity between two links or streets on the network, where a 1 represents that two links (i and j) share a node (Yamada and Thill, 2010). A more complex version of a network weights matrix might compute the network distance between nodes on a network (Yamada and Thill, 2010), as designated by the shortest path distance between the two nodes. As per their planar counterparts, these statistics are best used in conjunction with one another and decompose the global Moran (Moran, 1948) and the global G statistic (Getis and Ord, 1992) to uncover local "pockets" of spatial dependence that may be overlooked when using global statistics (Getis and Ord, 2010).

[1] Please see Anselin (1995) and Getis and Ord (1992) for the specification and a more thorough discussion of these statistics.

7 Results

-we will need to mention bandwidth for the network kernel density function here.

8 Discussion and Conclusion

References

Aldstadt, J. and Getis, A. (2006). Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38(4):327–343.

Anselin, L. (1995). Local indicators of spatial association-lisa. *Geographical analysis*, 27(2):93–115.

Black, W. (1991). Highway accidents: a spatial and temporal analysis. *Transportation Research Record*, 1318:75–82.

Black, W. and Thomas, I. (1998). Accidents on belgium's motorways: a network autocorrelation analysis. *Journal of Transport Geography*, 6(1):23–31.

Borruso, G. (2008). Network density estimation: A gis approach for analysing point patterns in a network space. *Transactions in GIS*, 12(3):377–402.

Cheng, W. and Washington, S. (2005). Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention*, 37(5):870–881.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1:269–271.

Downs, J. and Horner, M. (2007). Characterising linear point patterns. In *Proceedings of the 2007 GIS Research UK Annual Conference*, pages 11–13.

Getis, A. and Franklin, J. (1987). Second-order neighborhood analysis of mapped point patterns. *Ecology*, 68:473–477.

Getis, A. and Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206.

Getis, A. and Ord, J. (2010). The analysis of spatial association by use of distance statistics. *Perspectives on Spatial Data Analysis*, pages 127–145.

Kulldorff, M. (2006). Tests of Spatial Randomness Adjusted for an Inhomogeneity. *American Statistical Association*, 101(475):1289–1305.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8):799–810.

Kulldorff, M., Song, C., Gregorio, D., Samociuk, H., and DeChello, L. (2006). Cancer map patterns are they random or not? *American journal of preventive medicine*, 30(2S):37–49.

Kwan, M. (1998). Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30(3):191–216.

Levine, N., Kim, K., and Nitz, L. (1995). Spatial analysis of honolulu motor vehicle crashes: I. spatial patterns* 1. *Accident Analysis & Prevention*, 27(5):663–674.

Lu, Y. and Chen, X. (2007). On the false alarm of planar k-function when analyzing urban crime distributed along streets. *Social Science Research*, 36(2):611–632.

Miller, H. (1994). Market area delimitation within networks using geographic information systems. *Geographical Systems*, 1(2):157–173.

Moran, P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B*, 10(2):243–251.

Okabe, A. and Kitamura, M. (1996). A computational method for market area analysis on a network. *Geographical Analysis*, 28:330–349.

Okabe, A. and Okunuki, K. (2001). A computational method for estimating the demand of retail stores on a street network and its implementation in gis. *Transactions in GIS*, 5(3):209–220.

Okabe, A., Satoh, T., and Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a gis-based tool. *International Journal of Geographical Information Science*, 23(1):7–32.

Okabe, A. and Yamada, I. (2001). The k-function method on a network and its computational implementation. *Geographical Analysis*, 33(3):271–290.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Science*, 1(4):335–358.

Porta, S., Latora, V., Wang, F., Strano, E., Cardillo, A., Scellato, S., Iacoviello, V., and Messora, R. (2009). Street centrality and densities of retail and services in bologna, italy. *Environment and Planning B: Planning and Design*, 36(3):450–465.

Rogerson, P. (1999). The detection of clusters using a spatial version of the chi-square goodness-of-fit statistic. *Geographical Analysis*, 31(1):130–147.

Shiode, S. (2008). Analysis of a distribution of point events using the network-based quadrat method. *Geographical Analysis*, 40(4):380–400.

Spooner, P., Lunt, I., Okabe, A., and Shiode, S. (2004). Spatial analysis of roadside acacia populations on a road network using the network k-function. *Landscape ecology*, 19(5):491–499.

Swartz, P. (1998). An entropy-based algorithm for detecting clusters of cases and controls and its comparison with a method using nearest neighbours. *Health and Place*, 4(1):67–77.

Tango, T. (1999). Comparison of general tests for spatial clustering. *Disease Mapping and Risk Assessment for Public Health*, pages 111–116.

Xie, Z. and Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5):396–406.

Yamada, I. and Thill, J. (2004). Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2):149–158.

Yamada, I. and Thill, J. (2007). Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis*, 39(3):268–292.

Yamada, I. and Thill, J. (2010). Local indicators of network-constrained clusters in spatial patterns represented by a link attribute. *Annals of the Association of American Geographers*, 99999(1):1–1.