

Temporal Extensions of K Function

Sungsoon Hwang

Department of Geography

SUNY at Buffalo

Shwang5@buffalo.edu

Abstract

This study is aimed at extending K function to temporal dimensions. Given the formulation of (spatial) K function, two kinds of temporal K function have been formulated: (1) time K function (2) space-time K function. Time K function can be used to detect whether or not there is any unusual temporal pattern (i.e. regular, clustered). Space-time K function is broken down to three distinct types: The first type examines spatial pattern of observations disaggregated into temporal category. The second type explores temporal pattern of observations disaggregated into spatial category. The third type detects any interaction between space and time. As a case study, we tested a set of K function methods on multi-year fatal crash data. Slicing the locations of crashes by the appropriate resolutions of space and time allows us to detect hot spots that would have been unnoticed otherwise. This study demonstrates the explicit treatment of temporal dimensions can enhance the quality of knowledge discovery when time is crucial information for detecting “interesting” patterns. A collection of newly formulated K functions can provide a consistent method for exploring spatial, temporal, and spatio-temporal patterns of point events across varying scales.

1. Introduction

Spatial statistics has been widely used for hot spot detection of point events such as traffic accidents, epidemics, and crime. In particular, K function has been a popular choice for hot spot detection in that it can measure both intensity and spatial dependency over a wide range of scales without too much restriction on spatial measurements. Mapping the location of multi-period spatial data sets, we are easily forgetful of the fact that the mapped data are aggregated over time. K function does not specifically deal with temporal dimension. Rather, all spatial point events are treated as the phenomenon free of time dimension in K function. Taking time into consideration can provide us with a chance to examine the role temporal elements play in the phenomenon concerned. In some applications, temporal order is critical in understanding the behavior of the geographic phenomenon. The intensity of some geographic phenomenon may vary by time. Thus, it is worthwhile considering

temporal dimension in that task. There are several ways to count in time; (1) exploring temporal pattern (2) exploring spatial pattern across a wide range of temporal resolution in addition to spatial resolution (3) exploring spatio-temporal pattern, that is to say, the interaction between space and time. Moreover, spatio-temporal data have been available in an increasing rate while tools for exploring the data set in both dimensions do not keep pace with it. Therefore, it is timely procedures developing hot spot detection methods in spatial, temporal, and spatio-temporal dimension.

Different temporal resolutions have to be applied depending on the phenomenon in hand; For example, temperature (continuously) changes say every minute (i.e. process) while traffic crashes (discretely) occurs say every day (i.e. event) given the same unit of spatial area. Not only resolution is a kind of derived behavior associated with the phenomenon, but also resolution can be seen as defining properties; For instance, temporal resolution itself defines the phenomenon somehow like weather versus climate as does spatial resolution such as geography versus astronomy. While it is relatively easy to determine the appropriate time scale for the process (because it changes in quite a predictable manner), it is hard to say there is a reasonable time scale for the event because of its abruptness in nature. Our focus is placed on time scale for (discrete) events in which we find it hard to determine the optimal scale.

In the light of those concerns mentioned above (i.e. importance and timeliness of including temporal dimension in hot spot detection tasks, and difficulty in determining the appropriate scale in time for spatial point event), we proceed to modify K function into spatio-temporal clustering algorithms fully integrated in a geographic information system. To this end, we reformulate K function (known for analyzing spatial point pattern) into (1) space K function (which is the same as K function), (2) time K function, and (3) space-time K function. More specifically, time K function is formulated to detect clusters in time, and space-time K function is broken down to (a) the extension of space K function where time is treated as the attribute of spatial event, (b) the extension of time K function where space is treated as the attribute of temporal event, and (c) spatio-temporal K function where space and time are jointly treated. Determining the optimal scale for each of K functions will be presented also in the corresponding sections.

The remainder of this paper will be organized as follows: Section 2 will review the related literatures. Section 3 will present the modified version of K function for the analysis of spatio-temporal data, in particular for clustering pattern. In Section 4, we implement multi-scale spatio-temporal clustering algorithms based on K for the analysis of fatal crashes. Finally, Section 5 will conclude this study.

2. Related literatures

In this section, we consider methods for the analysis of a set of event locations, often referred to as

a point pattern analysis. As mentioned above, spatial statistical approach can be used to determine if events exhibit some form of clustering. The statistical formula given in this section attempts to “estimate how the intensity of events varies over the study area” (Bailey and Gatrell 1995).

The simplest way of measuring the intensity is to report the count of events in the equally-sized sub-regions, which is often referred to as *quadrat method*. There is a trade-off in choosing the size of sub-regions; large sub-regions may mask too much detail while small sub-regions may yield high variability. Not only it is difficult to decide what size to use, but also it does not take into account the relative location of events within the sub-regions.

To get around the fluctuation in variability, *kernel estimation* attempts to obtain a smooth estimate of probability density from an observed sample of observations. Similar to quadrat method, it superimposes spatial frames over observations, but this time instead of crude count within the frame, the count of events and distance from events to a priori given general location are translated into bivariate probability density function, known as kernel. Except that smoothing techniques take care of high variability, kernel estimation preserves two limitations discussed in quadrat method.

Two methods mentioned above are global in the sense that they use the general location as a reference to measure the distance, instead of inter-events distance. In this sense, the statistics is sensitive to the size of frame or bandwidth depending on methods. Those methods are used to summarize an overall pattern at a global scale over the study area, or large scale variation. Now the attention turns to local statistics – how intensity varies at a local scale of study area. Local statistics is used to detect spatial dependence or inter-event interaction.

One way of investigating the degree of spatial dependence is to examine the observed distribution of *nearest neighbor distances* (Bailey and Gatrell 1995). The analysis begins with setting a specific distance, say w , and calculating the probability that nearest neighbor distance will be less than w . The tendency for inter-event attraction or repulsion can be visualized when the empirical cumulative probability distribution is plotted against w . Graph that climbs very steeply in the early part of its range suggest clustering at a local scale. The limitation of nearest neighbor analysis is that it uses distances only to the nearest events and therefore only considers the smallest scales of pattern (Bailey and Gatrell 1995).

Capability of exploring spatial pattern across multiple resolutions (not limited to nearest neighbor) is provided in *K function*. In comparison with nearest neighbor analysis, K function allows us to measure the degree of spatial dependence over varying range of scales. K function is estimated as follows:

$$\hat{K}(h) = \frac{R}{n^2} \sum_{i \neq j} \frac{I_h(d_{ij})}{w_{ij}} \quad (\text{Equation 1})$$

Where R is area of study area, n is the total number of observed events, h is the distance considered

for local scale variation (or band size), d_{ij} is the distance between event i and event j , I_b is 1 if $d_{ij} < b$, or is 0 otherwise, and W_{ij} is the adjustment factor of edge effect. Since $K(b)$ would be greater than πh^2 under clustering, comparing $K(b)$ estimated from the observed data with πh^2 will provide a way to detect clustering. $L(b)$ provides the measure of spatial clustering as given in Equation 2. The high positive value of $L(b)$ will indicate the tendency for clustered spatial pattern.

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{\pi}} - h \quad (\text{Equation 2})$$

There are several questions to be addressed regarding Equation 2. First, “how can we determine the optimal scale where clusters are most effectively identified?” One way of doing this is to graph $L(b)$ against b where peaks in graph indicate clustering, and thus the corresponding value of b can be set to the optimal scale. Second, “how can we determine whether the magnitude of these peaks is significant?” It can be done by comparing $L(b)$ with the upper simulation envelope. If $L(b)$ is higher than the upper bound of simulated $L(b)$, then the significance of clustered pattern can be confirmed.

Spatial statistical approaches are largely credited for statistical elaboration of spatial phenomenon. Various methods can be devised under the assumption that spatial phenomenon can behave differently (e.g. continuous versus discrete, global versus local scale trend, non-stationary versus stationary). Above all, K function can be considered to be the most robust method for detecting hot spots over varying scales. Given the self-evaluation, K function is chosen for the further elaboration to treat temporal dimensions over other methods on the several scores. (1) It detects clusters over multiple resolutions; Interesting patterns may be associated with specific scales which could be missed if the whole range of possible scales was not explored; (2) It provides us with the robust method for choosing the optimal scale; Choosing the right scale is critical to knowledge discovery; (3) It is not a computationally expensive method (e.g. polynomial order); In such a way, we will be guaranteed to lower computational complexity in the face of large data sets.

3. Temporal extensions to K function

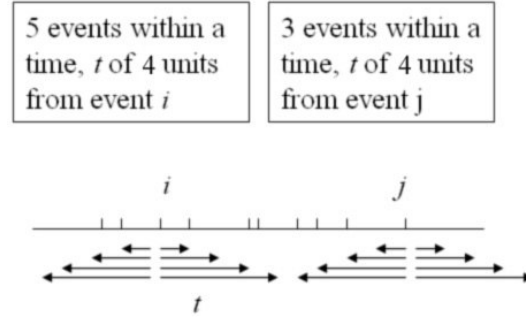
3.1. Time K function

Time K function is different from space K function in that it considers one-dimensional space (i.e. length) instead of two-dimensional space (i.e. area). Unlike space K function, time interval t (Equation 3) is considered instead of circular area with the radius b (Equation 1). Consequently, the size of study area R is replaced by total duration L . Similarly, adjustment factor of edge effect is derived from the duration addition of the start and end period in comparison to space K function

where areal addition of portions in edges. In general, time K function provides the measure of temporal dependence over varying time scales.

$$\hat{K}(t) = \frac{L}{n^2} \sum_{i \neq j} \sum \frac{I_t(d_{ij})}{w_{ij}} \quad (\text{Equation 3})$$

- L : total duration
- n : total number of observed events
- t : time interval
- d_{ij} : interval between i and j
- I_t : 1 if $d_{ij} < t$, 0 otherwise
- w_{ij} : adjustment factor of edge effect



Under regularity, $K(t)$ is expected to be $2t$, thus subtracting $2t$ from $K(t)$ will provide some way to detect temporal clustering. When $L(t)$ (Equation 4) is graphed against t , peaks will indicate clustering at the corresponding scales of time interval t . The significance of $L(t)$ can be tested by comparing $L(t)$ with the upper simulation envelope.

$$\hat{L}(t) = \hat{K}(t) - 2t \quad (\text{Equation 4})$$

3.2. Space-time K function

Space-time K function can be broken down to three cases. The first one can be seen as the extension of space K function where time is treated as an attribute of spatial event. The second one can be seen as the extension of time K function where space is treated as an attribute of temporal event. The third one measures the space-time interaction where space and time are jointly treated unlike the previous two cases.

3.2.1. The extension of space K function

This version of K function (Equation 5) provides the measure of spatial dependence at the spatial scale b wherein observations are disaggregated into the categorical type of temporal attributes (e.g. month, season, weekday versus weekend). Since the equation does not take account of numerical

attributes of time in an absolute scale, it does not measure the temporal pattern (such as clustered, random, and regular), thereby giving the name “the extension of *space* K function”.

$$\hat{K}_{ij}(h) = \frac{R}{n_i n_j} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \frac{I_h(d_{ij})}{w_{ij}} \quad (\text{Equation 5})$$

Where n_i is the number of observations for temporal attributes associated with event i , and n_j is the number of observations for temporal attributes associated with event j . Equation 5 can be useful in examining spatial patterns that vary by temporal category. For example, it can be used to test the hypothesis that crime is likely to be spatially concentrated in a specific portion of the total duration considered. Since it considers events to be spatial, Equation 2 can be used to test the evidence of spatial clustering.

3.2.2. The extension of time K function

This version of K function (Equation 6) provides the measure of temporal dependence at the temporal scale t wherein observations are disaggregated into the categorical type of spatial attributes (e.g. county, city, urban versus rural area). Since the equation does not take account of numerical attributes of space in an absolute scale, it does not measure the spatial pattern (such as clustered, random, and dispersed), thereby giving the name “the extension of *time* K function”.

$$\hat{K}_{ij}(t) = \frac{L}{n_i n_j} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \frac{I_t(d_{ij})}{w_{ij}} \quad (\text{Equation 6})$$

Where n_i is the number of observations for spatial attributes associated with event i , and n_j is the number of observations for spatial attributes associated with event j . Equation 6 can be useful in examining temporal patterns that vary by spatial category. For example, it can be used to test the hypothesis that epidemics is likely to be temporally concentrated in a specific sub-area. Since it considers events to be temporal, Equation 4 can be used to detect temporal clustering.

3.2.3. Spatio-temporal K function

This version of K function (Equation 7) provides the measure of the interaction between space and time at both spatial scale h and temporal scale t . The equation takes account of numerical attributes of both space and time in an absolute scale. Therefore, it can measure spatio-temporal pattern

(often referred to as “space-time clustering” or “space-time interaction” test in spatial statistics literatures). Equation 5 and 6 are concerned with a labeling of events in terms of different types, whether temporal (for space K function) or spatial (for time K function) (e.g. spatially concentrated crimes in the summer, temporally concentrated SARS cases in the southern China). In contrast, Equation 7 concerns whether space and time interact – that is, if a pair of events is close in space *and* close in time. This is particularly useful in studying a dynamic process such as in an epidemiological context.

$$\hat{K}(h,t) = \frac{LR}{n_i n_j} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \frac{I_{h,t}(d_{ij})}{w_{ij}} \quad (\text{Equation 7})$$

Where n_i and n_j are respectively the number of observations within distance h and time interval t for a pair of events considered, that is i and j . $I_{h,t}$ will be 1 if event i and j are both within distance h and time interval t , or be 0 otherwise. The more a pair of events are closed in space and time, the higher the estimated value of $K(h,t)$ will be. Therefore, it is a good measure of space-time interaction. No interaction between space and time can be formulated as $K(h)K(t)$, thus by comparing $K(h,t)$ with $K(h)K(t)$, we can test the tendency for space-time interaction. Consequently, the test of space-time interaction will look like Equation 8.

$$\hat{D}(h,t) = \hat{K}(h,t) - \hat{K}(h)\hat{K}(t) \quad (\text{Equation 8})$$

High positive value of Equation 8 will suggest the evidence of space-time interaction. On the contrary, low negative value of Equation 8 will indicate that there is no tendency for space-time interaction. In the context of epidemiology, high positive value of Equation 8 will indicate the high degree of infection.

4. GIS implementation for the analysis of fatal crashes

4.1. Study areas

In this section, we present the implementation of extended K functions integrated in GIS for analyzing spatio-temporal pattern of fatal crashes. Considering the relevance to applications, we implement (1) space K function (to detect spatial clustering pattern), (2) time K function (to detect temporal clustering pattern), and (3) space K function extended to time (to detect spatial clustering pattern that may vary by categorical temporal attributes). We chose two study areas in terms of

spatial scales to test the consistency of proposed methods. Study area 1 covers New York State (Figure 1) while study area 2 covers the part of New York City (King and Queens County) (Figure 2). We consider the time periods from year 1996 to year 2001 for both study areas. To correct for edge effect in space, areas within a reasonable range of edges are excluded (Geocoded data sets are not available beyond New York State at this point). On the contrary, we add additional time periods beyond study periods instead of excluding data in edge periods due to the availability of data beyond time periods.

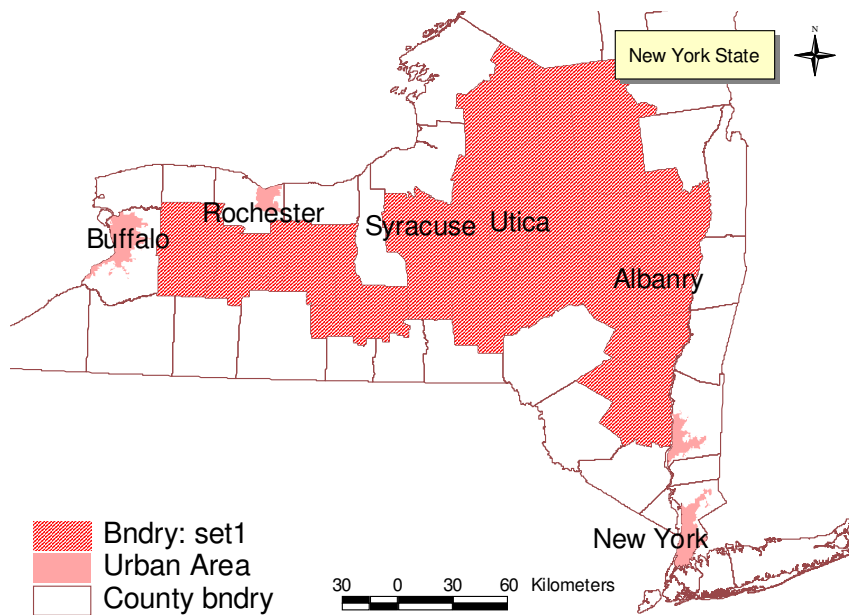


Figure 1. Study Area 1: marked as red oblique line

The map of observed events for study area 1 is depicted in Figure 3. The number of observed events considered for study area 1 is 1508 where the size of study area is 48306.5110 square kilometers and the total duration is 2192 days. It turns out 0.0312 fatal crashes per square kilometers and 0.688 fatal crashes per day are reported in the study area.

The observed events for study area 1 are shown in Figure 4. The number of observed events considered for study area 2 is 689 where the size of study area is 210.5319 square kilometers during 2192 days. It turns out 3.27 fatal crashes per square kilometers and 0.3143 fatal crashes per day are reported in the study area. The spatial density of fatal crashes in New York City is approximately hundred times higher than that of New York State based on observations of two study areas.

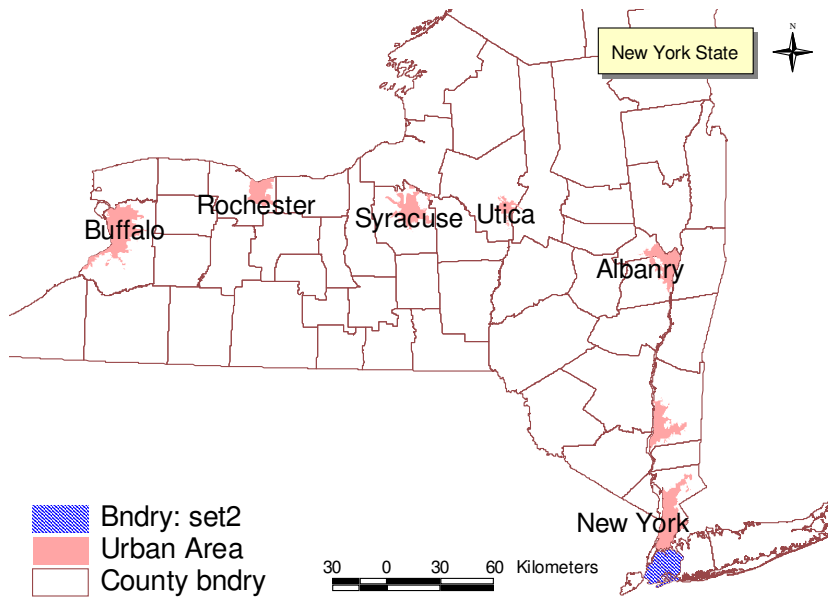


Figure 2. Study Area 2: marked as blue oblique line

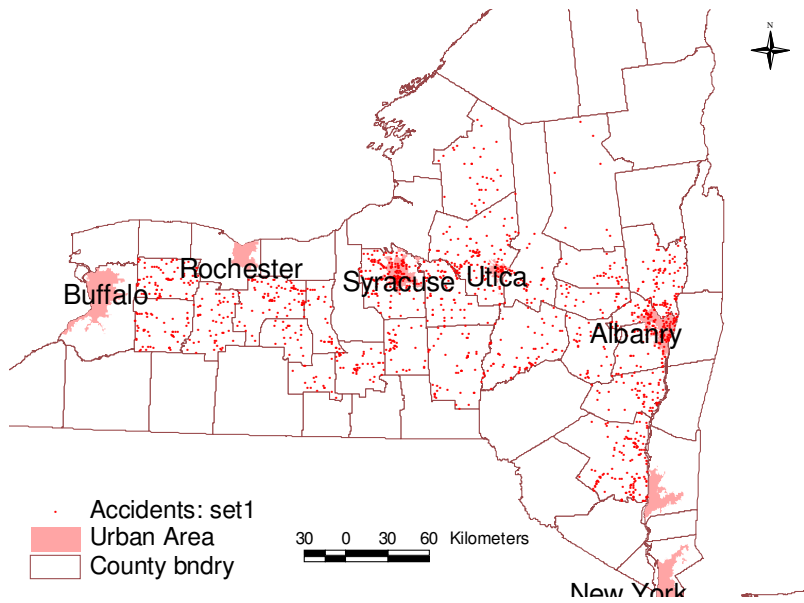


Figure 3. Observed fatal crashes in study area 1 (Source: FARS 1996-2001)

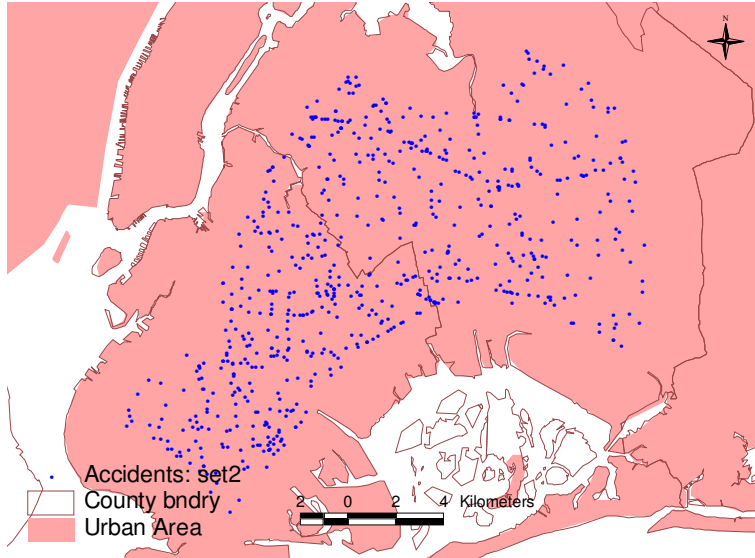


Figure 4. Observed fatal crashes in study area 2 (Source: FARS 1996-2001)

4.2. Finding the optimal scale

Based on observed fatal crashes in study area 1, estimated value of $L(b)$ from Equation 2 is graphed against b (Figure 5). The graph shows clear evidence of spatial clustering as can be seen from the positive value of estimated $L(b)$ across the whole range of b . To test the significance of a clustered pattern, we obtain lower and upper bound based on Monte Carlo Simulation in the iterations of 100. The significance of spatial clustering is confirmed by the fact estimated $L(b)$ is way over the upper bound of the simulated $L(b)$. Furthermore, one peak in the value of estimated $L(b)$ is clearly pronounced, thus we can safely conclude the optimal scale for spatial clusters is around 16 kilometers. To show the visual evidence of clustered spatial pattern in a 2-dimensional Cartesian space, we will present the kernel density map of spatial clusters (where 16 kilometers are set to bandwidth), which can be deferred to the section 4.3.

To examine if there is any tendency for temporal clustering in study area 1, estimated value of $L(t)$ from Equation 4 is graphed against t (Figure 6). It is evident that observed fatal crashes are not temporally clustered as seen from the negative value across the whole range of t (simulation is omitted as a consequence). Therefore, it is concluded there is no tendency for temporal clustering in this data set. Neither can we choose the optimal scale for the possible clusters in time as a result. We do not present the map of temporal clusters due to the insignificance of clusters evident in this graph.

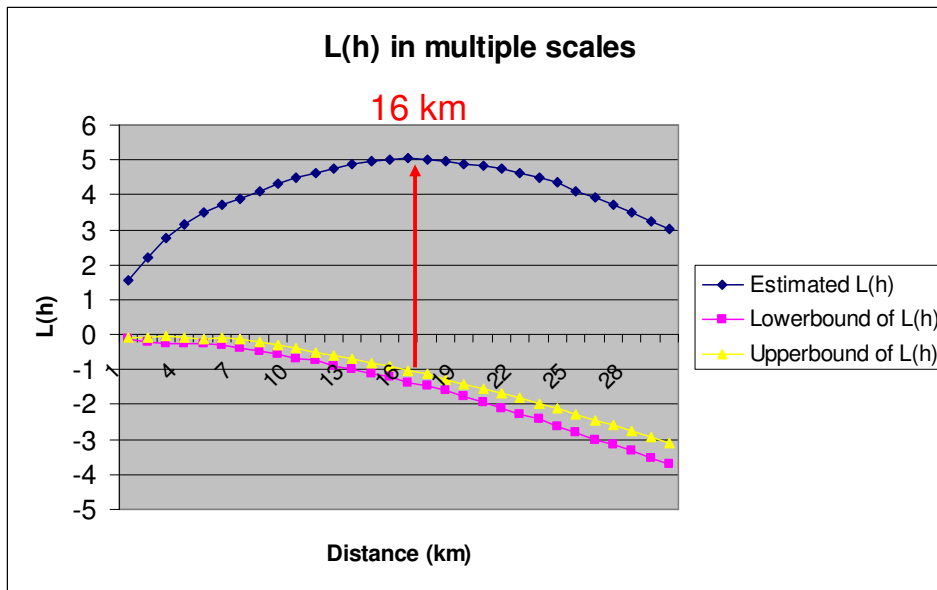


Figure 5. The graph of estimated $L(h)$ across h for study area 1

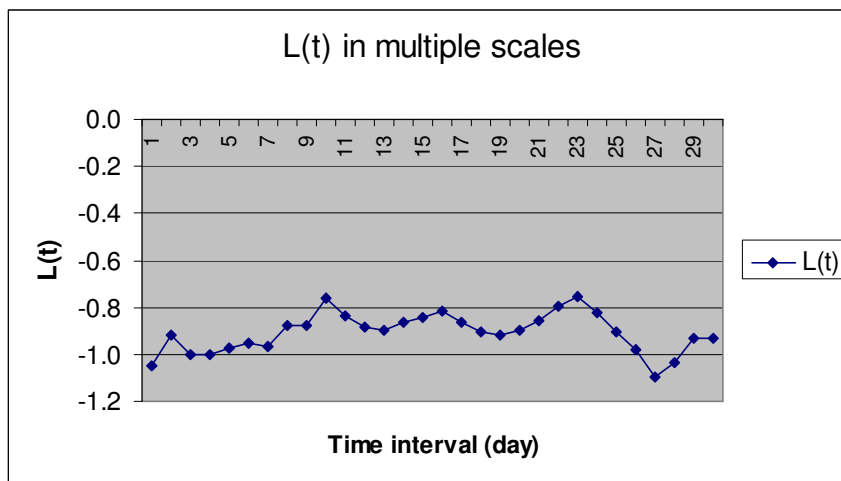


Figure 6. The graph of estimated $L(t)$ across t for study area 1

We implement Equation 5 (extension of space K function) where observations are disaggregated into 12 month periods using data sets of study area 1. The K function measures the tendency for spatial clustering when observations are categorized into 12 different types (here month). That way, the function allows us to examine if data shows any clear spatial pattern for each month, thus to compare spatial variations across months. Evident in Figure 7, fatal crashes in study area 1 tend to be more spatially clustered in the month of May than any other months as can be seen in the pronounced peak from the Month axis. Since observations are disaggregated into each month, the number of observations examined in each month is significantly reduced. As a result,

optimal scale (peak from the Distance axis) is shown to be the maximum given the range, 30 kilometers in comparison to 16 kilometers in Figure 5. The visual evidence for spatial clustering in May will be provided in the section 4.3.

Now we turn to study area 2, a highly urban setting. Due to the significantly smaller size of study area, the scale of distance h is proportionally reduced. As a consequence, $L(h)$ of study area 2 is a fraction of that of study area 1 (see Figure 5 for comparison). Anyhow, study area 2 also shows the clear evidence of clustered spatial pattern as can be seen from the value of $L(h)$ higher than the upper envelope of simulated $L(h)$. In comparison to Figure 5, $L(h)$ shows large fluctuations in value due to relatively reduced number of observations tested in the detailed spatial scale. Two candidate scales are chosen from peaks in the estimated value of $L(h)$: 0.09 kilometer and 0.18 kilometer, as marked in Figure 8.

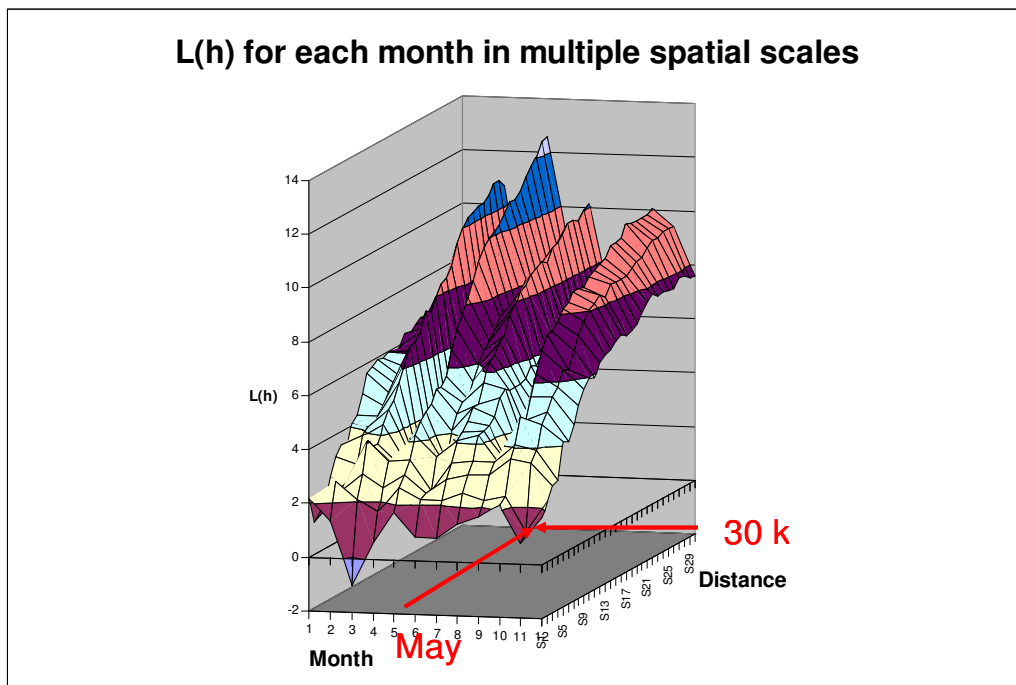


Figure 7. The graph of estimated $L(h)$ for space K function extended to temporal attributes across space (h) and time (months) (obtained from study area 1)

To examine the tendency for temporal clustering in study area 2, $L(t)$ is graphed against t . In Figure 9, a pronounced peak in the time scale of 30 days may suggest a cluster of fatal crashes seems to be concentrated in a way that varies by month, which is quite contrastive to Figure 6 with no interesting peak whatsoever. Nevertheless, it is hard to find the significant evidence for temporally clustered pattern in traffic crashes in this study area judging from rarely positive value of $L(t)$.

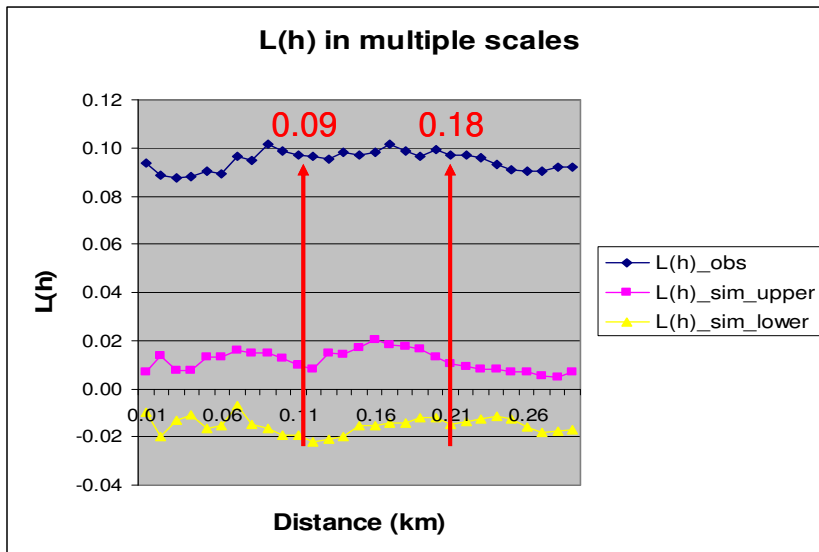


Figure 8. The graph of estimated $L(h)$ across h for study area 2

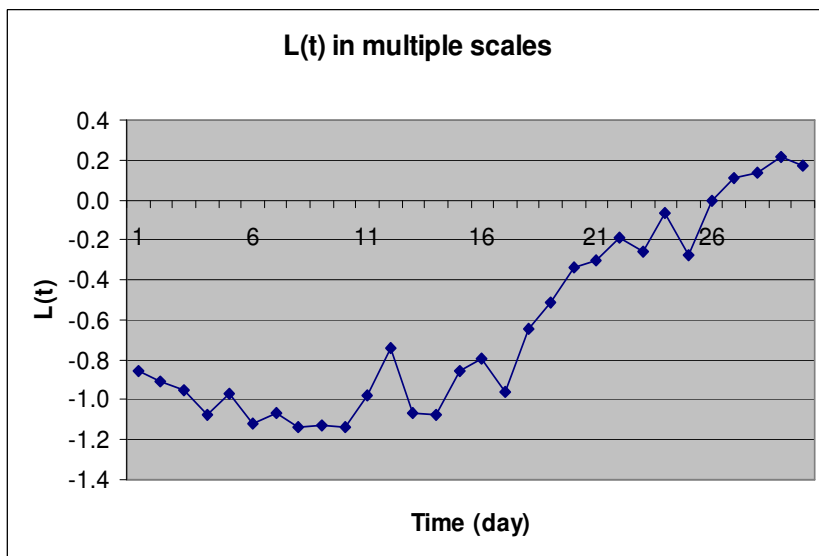


Figure 9. The graph of estimated $L(t)$ across t for study area 2

To find out any tendency for spatial patterns of fatal crashes that may vary by months, we disaggregate data into 12 groups by month. Compared to study area 1, the estimated value of $L(h)$ gives a rather spiky impression, which is not surprising due to the reduced number of observations in the detailed scale considered. In Figure 10, November turns out to be most pronounced, and thus it allows us to conclude there is a highly spatially concentrated cluster of fatal crashes in November compared to other months. When it comes the optimal scale, the estimated value of $L(h)$ is the most

pronounced in the scale of 0.36 kilometer.

4.3. Visualizing clusters

This section is designed to present kernel density maps of fatal crashes whenever they turn out to exhibit significant clustering pattern in space and time. It will help readers ascertain the visual evidence of clustered patterns. Kernel density maps are created in a way that the bandwidth is determined by the methods shown in the section 4.2. Kernel density maps are chosen as a visualizing tool because they give a better sense of what is going on by surrendering relatively trivial details to general trends.

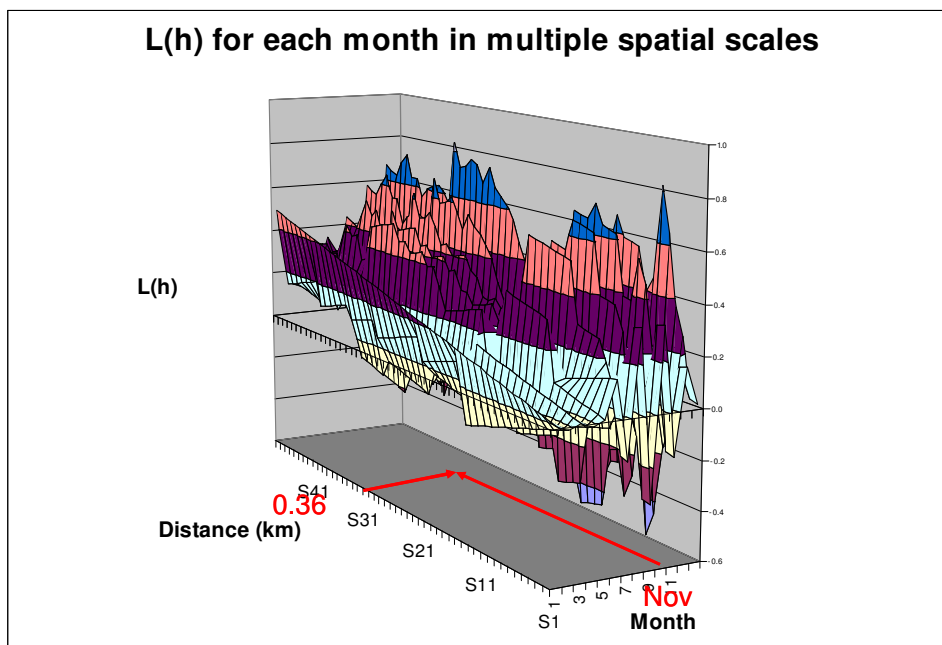


Figure 10. The graph of estimated $L(b)$ for space K function extended to temporal attributes across space (b) and time (months) (obtained from study area 2)

Figure 11 shows the kernel density map of fatal crashes where bandwidth is set to 16 kilometers. The evidence of clustered spatial pattern confirmed in Figure 5 can be visually ascertained – highly clustered pattern of fatal crashes in Albany, Syracuse, and corridor linking Buffalo and Rochester.

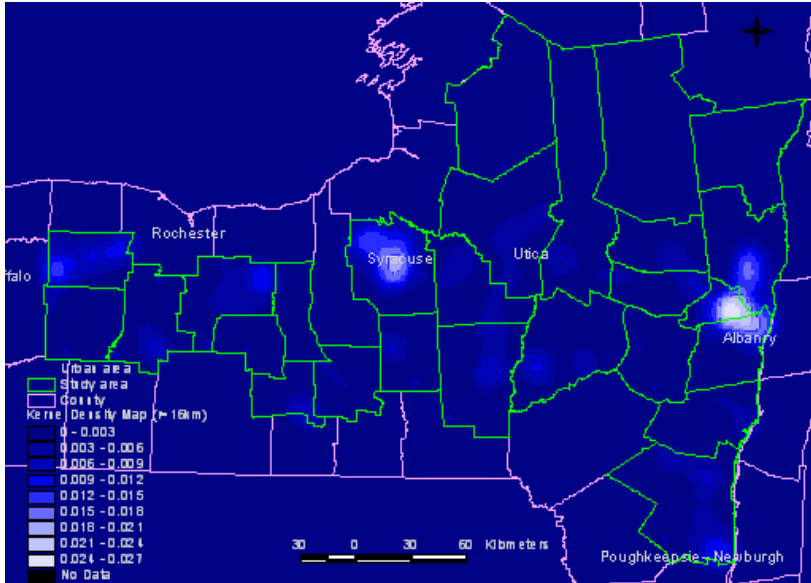


Figure 11. New York State kernel density map for total fatal crashes ($r = 16$ km)

Figure 12 depicts the kernel density map of fatal crashes that occurred on May. Reduced number of observations combined with larger bandwidth (30 km) smooths the distribution rather too much while it is hard to find any significant difference in spatial variations of clusters.

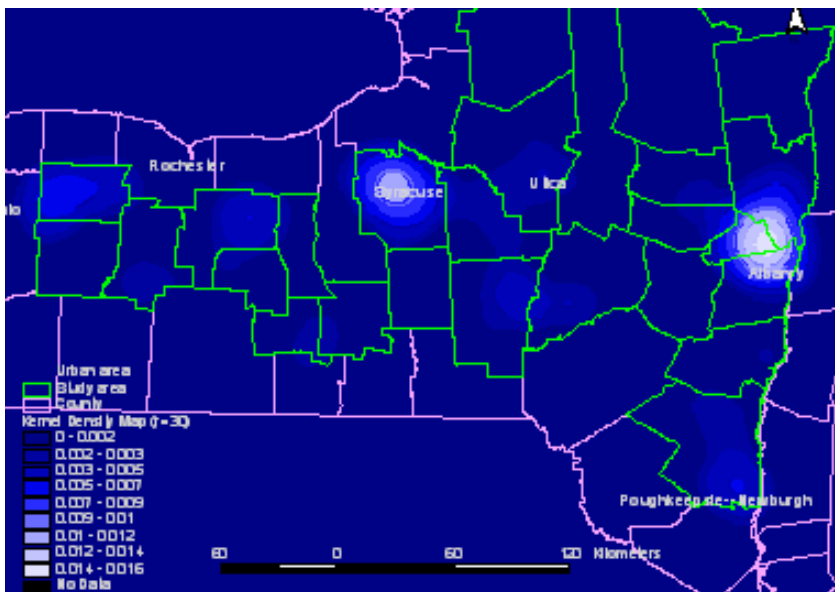


Figure 12. New York State kernel density map for fatal crashes on May ($r=30$ km)

Fatal crashes in New York City are mapped as a kernel density estimate with bandwidth 0.18 km in Figure 13. Detailed spatial scale gives a rather spiky impression of spatial distribution, but it still

provides a better sense of where hot spots are than the map of crude count (Figure 4).

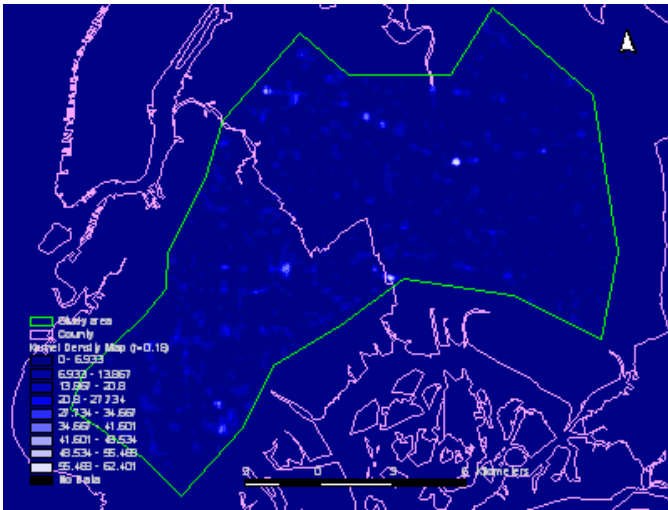


Figure 13. New York City (King, Queens County) kernel density map of total fatal crashes ($r=0.18$ km)

What about fatal crashes that occurred only in November for the recent six years? As shown in Figure 14, the spatial variation of fatal crashes on November shows a different pattern compared to total fatal crashes shown in Figure 13. It may suggest these hot spots are involved in different cause of crashes (e.g. possibly winter driving in combination with unique topography) than that of crashes in general.

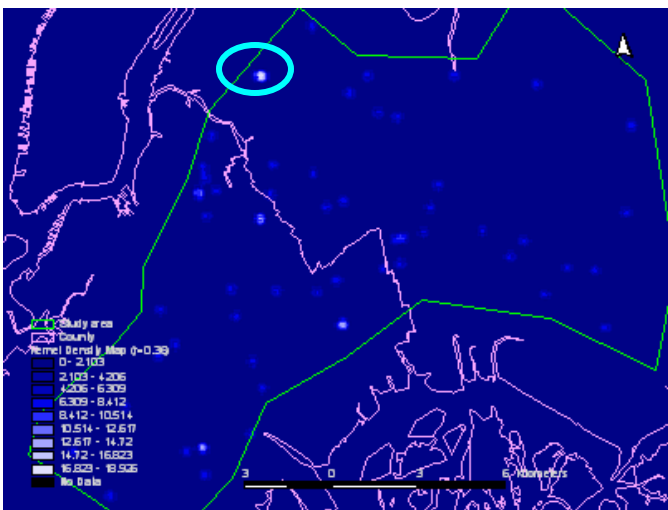


Figure 14. New York City kernel density map of fatal crashes on November ($r=0.36$)

5. Conclusions

We devised a spatial statistical method for detecting hot spots of point events in space and time. With K function as a base model, temporal extension to K function has been formulated. Modified versions of K function are designed to analyze pattern in space and time. More specifically, several distinct cases for the temporal extension to K function are identified: (1) time K function detects temporal clusters (2) space-time K function detects (a) spatial clusters of point events stratified by categorical temporal attributes (b) temporal clusters of point events stratified by categorical spatial attributes (c) space-time interaction.

The reformulated K functions combined with the method for the selection of optimal scale are written as a macro language to fully integrate the functionalities in GIS. The K function-based multi-scale spatio-temporal clustering algorithms are implemented in the application of fatal crashes analysis. A case study does not show any evidence for temporally concentrated pattern of fatal crashes in contrast to highly clustered pattern in space. More important, temporal extension of space K function (i.e. which attempts to examine spatial pattern of observed events disaggregated by temporal attributes such as month) turns out to be useful in discovering pattern that would have been unnoticed if observed events were not disaggregated by temporal types and if the whole range of possible scales were not explored. The noticeable differences in the spatial variations of total fatal crashes versus fatal crashes on November in New York City illustrate the benefit of multi-scale spatio-temporal analysis.

One of the limitations of this study lies in the assumption we made with regard to the behavior of temporal phenomenon. Since we extend space K function to temporal dimension, we were not able to give special treatment of temporal dimension. Rather, time is assumed to behave as if it were space except for the different treatment of dimensional size (i.e. two dimensional versus one dimensional). Recognizing the difference between spatial phenomenon and temporal phenomenon is definitely more fundamental task that precedes algorithms or statistical formula. However, the reality does not preclude the possibility that temporal phenomenon can act like spatial phenomenon. Rather, ontological extension to modified K function will suffice.

Further works associated with more workable limitations are: (1) The study would not cover the complete list of spatiotemporal cases, thus it is necessary to elaborate on space-time K function. (2) The study does not correct for variation of population at risk. It is expected to discover more “interesting” pattern when variation of population at risk is corrected for. (3) Case study limits the application of extended K function to crashes data. Applying other types of space-time K function to other types of data such as epidemics data may yield fresh insights due to different behavior of spatio-temporal phenomenon.

References

Bailey TC, and Gatrell AC, 1995, *Interactive Spatial Data Analysis*, Longman Scientific & Technical, Essex