Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 16-031

Crime Hotspot Detection: A Computational Perspective

Emre Eftelioglu, Xun Tang, Shashi Shekhar

September 1, 2016
Crime Hotspot Detection
A Computational Perspective

Emre Eftelioglu, Xun Tang, Shashi Shekhar

ABSTRACT

Given a set of crime locations, a statistically significant crime hotspot is an area where the concentration of crimes inside is significantly higher than outside. The motivation of crime hotspot detection is twofold: detecting crime hotspots to focus the deployment of police enforcement and predicting the potential residence of a serial criminal. Crime hotspot detection is computationally challenging due to the difficulty of enumerating all potential hotspot areas, selecting an interest measure to compare these with the overall crime intensity, and testing for statistical significance to reduce chance patterns. This chapter focuses on statistical significant crime hotspots. First, the foundations of spatial scan statistics and its applications (i.e. SaTScan) to circular hotspot detection are reviewed. Next, ring-shaped hotspot detection is introduced. Third, linear hotspot detection is described since most crimes occur along a road network. The chapter concludes with future research directions in crime hotspot detection.

INTRODUCTION

Analyzing crime locations and using the spatial information associated with them is a fundamental task of environmental criminology since the main goal of a crime investigation is to locate the criminal and/or prevent more crimes from occurring. Two important spatial theories in environmental criminology are routine activity theory and crime pattern theory. Routine activity theory states that a crime location is related to a serial criminal’s frequently visited areas (Burn, 1982). Crime pattern theory extends routine activity theory on a spatial model (Brantingham & Brantingham, 1993). When crime analysts study the locations of crime sites and make inferences that help track down a serial criminal or that identify areas where extra police presence is needed, they are using these theories. Thus, the motivation of crime hotspot detection can be summarized as (1) to detect current/emerging crime hotspots to focus the deployment of police forces and (2) to predict the location of a serial criminal’s residence.

Previously, most analysis of crime locations was done manually. However, the task of enumerating potential hotspots by hand and detecting meaningful patterns in the data is arduous, even for experienced analysts. Rising crime numbers in larger cities increase analysts’ workload and compound their stress. For example, a typical crime dataset of the continental United States includes around $10^7$ crimes annually (FBI, 2015). Therefore, law enforcement agencies need computational tools that can automate hotspot detection. Moreover, these tools should eliminate the chance patterns (false positives). Most police departments and similar entities have limited resources for crime mitigation and prevention. Therefore, false positive results, that is, hotspots that occur only by chance, risk causing real harm. If a police department diverts money and manpower to a chance hotspot location that don’t need it, may leave other areas less protected. Chance hotspots can have other harmful consequences as well. When police increase their presence in a location identified as a crime hotspot, people naturally begin to avoid the area out of fear for their safety, and the neighborhood may become stigmatized.

Hence there is a growing need for crime hotspot analysis tools which can handle large crime datasets as well as eliminate chance patterns. Such tools may help crime analysts accomplish their law enforcement goals with optimal resource allocation (Eck & others, 2005).
Many computational techniques (Spencer Chainey & Ratcliffe, 2013; Eck & others, 2005; Levine, 2006) are used for hotspot detection. One of the most common methods is to enumerate hotspots with clustering techniques. These techniques are highly scalable and widely used in many societal applications (Ester & others, 1996; Ilango & Mohan, 2010; Liao, Liu, & Choudhary, 2004; Sarmah, Das, & Bhattacharyya, 2007; Wu et al., 2008). Nevertheless, clustering has two important limitations for hotspot detection. First, these techniques require the user to input the number of clusters beforehand. This does not conform well to hotspot detection, whose aim is to search for an abnormality that may or may not exist in the data. In other words, the user does not know beforehand whether a hotspot exists, or if so, how many. A second limitation is that most clustering techniques lack a statistical significance test to eliminate chance occurrences of hotspots.

This chapter focuses on statistically significant hotspot detection techniques that aim to remove chance patterns. The shape of the desired hotspot, determines the enumeration technique used to generate candidate hotspots and spatial scan statistics is used to assess the statistical significance of each candidate.

Three representative techniques for discovering significant crime hotspots are presented and each technique is illustrated using sample case studies on real crime datasets collected from different states in the U.S. However, it is important to note that there are many statistically significant hotspot detection techniques with different outputs that are designed to be used in different application domains, such as epidemiology, animal foraging, medical imaging etc. (Agarwal, McGregor, Phillips, Venkatasubramanian, & Zhu, 2006; Coleman et al., 2009; Janeja & Atluri, 2005a; Kulldorff, Huang, Pickle, & Duczmal, 2006; McCullagh, 2006; Neill & Moore, 2004; Neill, 2009; Patil & Taillie, 2004; Prates, Assunção, & Costa, 2012; Shi & Janeja, 2011; Walther, 2010; Zhang, 2010).

Formally, a crime hotspot can be defined as an area where the concentration of crimes inside is significantly higher than the number of activities in any other place in the study area. The hotspot can be circular (partial/full), rectangular, ring-shaped or linear depending on the goal of the hotspot analysis, the geographic features associated with the study area and the crime type. Circular hotspots occur when the crimes are spread around an anchor point of a criminal. For example, narcotics crimes tend to occur in a circular pattern causing a circular hotspot around the criminal narcotics manufacturer (Belenko & Spohn, 2015). Similarly, linear hotspots can be seen for some crimes that are committed along a road network. For example, street robbery cases can be seen on the road network and crime activities tend to be on specific road segments where a criminal selects as his base crime location. Finally, ring-shaped hotspots can be seen for serial criminals, who try not to commit crimes too close to home and try to spend least effort to travel causing a ring-shaped crime pattern around their residence.

SCOPE

This chapter provides and overview of statistically significant crime hotspot detection and outlines its capabilities for aiding in law enforcement with the primary intent of reducing chance patterns in crime hotspot detection. Hence this chapter does not focus on presenting computational performance evaluation. Also there are a number of clustering techniques that are used for crime hotspot detection, but this chapter just focuses on statistically significant crime hotspot detection using spatial scan statistics.

BACKGROUND

This section reviews basic concepts related to statistically significant hotspot detection and formally define the crime hotspot detection problem.
Basic Concepts

**Activity Set:** An activity set \( A \) is a collection of geo-located points (i.e. crime locations). Each activity \( a \in A \) is associated with a pair of coordinates \((Lat, Lon)\) representing its location in two-dimensional Euclidean space.

**Study Area:** Study area \( S \) is territorial area of responsibility of a public safety agency. \( S \) is assumed to be the minimum orthogonal bounding rectangle of the activity set \( A \) in this chapter.

**Spatial Bi-Partitioning:** Spatial bi-partitioning is the division of an area into two non-overlapping subsets. The key point of spatial bi-partitioning is the notion of “inside” and “outside”. Any activity point \( a \in A \) is either inside or outside but not both.

**Spatial Comparison:**

Spatial comparison is the comparing of two populations after spatial bi-partitioning, i.e., the population “inside” and the population “outside” which have been partitioned in the spatial bi-partitioning phase. This comparison can be done by different metrics such as difference or ratio; by different interest metrics such as disease rate, areal density or linear density; or by different significance tests such as parametric (Poisson) test or non-parametric test.

**Spatial Scan Statistics**

Spatial scan statistics is a widely used tool to detect hotspots in a point process and it’s widely used in epidemiology to determine disease outbreaks (Neill & Moore, 2004). It was first studied for one-dimensional data and then extended to multi dimensional data (M. Kulldorff, 1999; Kulldorff, 1997). Spatial scan statistics has proved to be the most powerful statistical test to detect hotspots in an activity set (M. Kulldorff, 1999).

Suppose there is a point process \( N \) in a study area \( S \). The aim is to find a hotspot in this point process. First, candidate hotspot regions are enumerated using a window \( W \), which moves around the study area and creates candidate hotspot regions referred to as zones \( Z \). For each of these zones, the number of observed and expected observations is noted. For example, Figure 1 A circular window enumerating zones in a study area shows a circular window moving in the study area \( S \) and creating zones. Note that each created zone will be a candidate hotspot to be evaluated for being an actual hotspot.

![A circular window enumerating zones in a study area](image)

Figure 1 A circular window enumerating zones in a study area

The fundamental question asked for each of the zone \( Z \) is “Is there any difference between inside the zone and outside?” In order to answer this question, a Hypothesis Test is constructed with two outcomes. The null hypothesis \( (H_0) \) states that the activity points are distributed randomly according to a
homogeneous Poisson process over the study area $S$. The alternative hypothesis ($H_1$) states that the inside of a zone $Z$ has higher activities than outside (Martin Kulldorff, 2014).

The hypothesis test has two important requirements. The first is a metric to give a score for each zone $Z$. If the window had a fixed size and thus all zones $Z$ had the same size, then the test statistic would be the count of the points inside the zone. Since the actual size of the hotspot in a crime scenario is not known beforehand, the window should have a variable size and thus created zones $Z$ will not have the same size. Therefore, a log likelihood ratio test is the metric used to determine the test statistic. The second requirement of hypothesis testing is a distribution of the test statistic under the null hypothesis of no clusters. This distribution will show what one should expect when there is no hotspot in the dataset. This distribution is acquired using a randomization test. Finally, the test statistic score of each zone is compared with the distribution of the test statistic to determine the statistical significance ($p$-value). Given a required significance level ($\alpha$), if a zone $Z$ has a $p$-value lower than $\alpha$, it can be concluded that the zone $Z$ is a significant hotspot (Kulldorff, 1997b). These concepts are described in more detail next.

**Likelihood Ratio Test**

The likelihood ratio test determines the test statistic of a candidate zone $Z$. It is computed by dividing the likelihood of the activities in zone $Z$ under the alternative hypothesis ($H_1$) by the likelihood of the zone $Z$ under the null hypothesis ($H_0$). This comparison gives the following equation:

$$LR_Z = \frac{L(Z)_{H_1}}{L(Z)_{H_0}}$$

In order to find the maximum likelihood of a zone $Z$, the parameters in $LR(Z)$ are selected to maximize the likelihood ratio. These terms are used as the supremum (least upper bound) when likelihood ratios are computed. Thus, the Likelihood Ratio is derived as follow. Note that $\log$ values are used to make the interpretation easier.

$$\log LR_Z = \log \left( \left( \frac{c}{B} \right)^c \left( \frac{A-c}{A-B} \right)^{A-c} I() \right), \text{ where } B = \frac{|A| \text{area}(Z)}{\text{area}(S)} \text{ and } I() = \begin{cases} 1, & \text{if } c > B \\ 0, & \text{otherwise} \end{cases}$$

where $B$ is the expression of the expectation of number of activities in a particular zone $Z$. The observed number of activities is denoted as $c$. Total number of activities is the cardinality $|A|$ of set $A$. The final term $I()$ denotes the indicator function. $I() = 1$ when a zone $Z$ has more activity points than expected ($c > B$) and otherwise it is set to 0 to prevent the detection of low activity areas (Martin Kulldorff, 2014).

**Randomization Test**

Once the enumeration of the candidate hotspots (zone $Z$) is done and the test statistic of each candidate is computed, a randomization test is done to determine the distribution of the test statistic under the null hypothesis ($H_0$). This distribution later helps to determine the statistical significance of a candidate hotspot and will help determine whether the candidate hotspot occurred by chance or is truly anomalous. Although other methods exist, in this chapter the randomization test with Monte Carlo simulations will be explained.

For Monte Carlo simulation, $m$ random activity sets $A_{1...m}^{\text{random}}$ are created in the same study area $S$. The locations of the activities are determined using complete spatial randomness (CSR). In other
words, points inside each activity set are randomly thrown into the study area. For each of these random activity sets $A_{1\ldots m}^{\text{random}}$, new zones $Z$ are generated and their test statistics (log likelihood ratio) are computed using almost the same method as for the original activity set $A$. The only difference is that only the maximum log likelihood ratio is stored. Thus, at the end of this process a list of $m$ test statistic values are obtained from $m$ random activity sets ($A_{1\ldots m}^{\text{random}}$). These values are used to determine the statistical distribution of the log likelihood ratios.

**Hypothesis Test**

The null hypothesis states that activity points are distributed randomly according to a homogeneous Poisson process over the study area $S$ (Kulldorff, 1997a). In order to accept or reject the null hypothesis (accept $H_1$), we need to determine the significance level ($p$-value) of the candidate hotspots. We do this by comparing a candidate’s test statistic with the distribution of test statistics obtained from the randomization test, and then dividing its position by the number of Monte Carlo simulation trials $m$. If the computed $p$-value is higher than the desired significance level ($\alpha_p$), the $H_1$ hypothesis is rejected; otherwise, it is said that $H_1$ can’t be rejected, indicating the detected hotspot is significant.

With this background, we are ready to consider three representative approaches to statistically significant hotspot detection. Case studies on real crime data are included to illustrate each approach. We also briefly explain each method’s limitations.

**CIRCULAR HOTSPOT DETECTION**

Criminal activity in an area tends to spread in a circular process similar to that described by the diffusion model in physics and chemistry. Diffusion means that molecules or heat will move away from their sources once they are discharged. The diffusion model provides a natural way to describe the circular spreading of cases (i.e. crimes). This gives rise to the notion of circular footprints of hotspots in isotropic geographies. Therefore, the ability to detect a circular hotspot is an important task in environmental criminology.

In formal terms, circular hotspot detection can be defined as follows: Given a set of geo-located points (e.g. crime reports) and a significance threshold, circular hotspot detection finds circular shaped areas where the concentration of points inside a circle is significantly higher than the number of points outside. Circular hotspot detection requires the following inputs specified by an analyst:

(a) A spatial crime dataset with a collection of crimes, each associated with geo-location information (i.e., latitude, longitude),

(b) A statistical significance threshold ($\alpha_p$).

Based on the above inputs, the goal of circular hotspot detection is to report a collection of circular hotspot areas with $p$-values lower than the specified statistical significance threshold ($\alpha_p$).

**Case Study on a Real Crime Data**

A typical input to circular hotspot detection is a spatial crime dataset that contains crime reports each associated with geo-location. For example, Figure 2 First Degree Murder cases in Chicago between March 2013 and March 2014. Dataset includes 384 Homicide Cases shown in blue dots. shows a crime dataset from Chicago, Illinois (City of Chicago, 2013). The dataset includes 384 homicide (by first degree murder) crimes committed in Chicago, Illinois between March 2013-March 2014. Each crime location is shown as blue dots and the map is created using QGIS Software (Quantum GIS Development Team, 2012), and the Open Street Plugin (Haklay & Weber, 2008). Study area is selected as the minimum
orthogonal bounding rectangle of the crime dataset. Note that the study area could be the territorial area of responsibility for a public safety agency. The aim is to find circular hotspots with a statistical significance level of $\alpha_p = 0.001$. Figure 3 Output of SaTScan with two circular hotspots shown in blue. shows the output of SaTScan with two hotspot areas having log likelihood ratio of 107.10 and 48.26 respectively. These hotspots have p-value = 0.001 which indicates statistical significance at the 99.9 % confidence level.

Figure 2 First Degree Murder cases in Chicago between March 2013 and March 2014. Dataset includes 384 Homicide Cases shown in blue dots.
SaTScan Circular Hotspot Detection Algorithm

SaTScan is a state-of-the-art software package that uses spatial scan statistics to detect circular hotspots. It is used extensively in epidemiology to detect circular disease hotspots.

In its simplest form SaTScan uses a three-step approach to detect statistically significant circular hotspots: (1) Enumeration of Circles, (2) Log likelihood ratio test and (3) Monte Carlo simulation and Hypothesis test. Although indexing and optimization techniques exist for the circle enumeration and Gumbel Approximation (Martin Kulldorff, 2014) and Early Monte Carlo termination for the Monte Carlo simulation step can be used to improve the scalability of the SaTScan algorithm, this chapter describes these three steps without any algorithmic refinements.

**Step 1 - Enumeration of Circles:** When enumerating candidate circular hotspots, SaTScan uses activity points in the activity set \( a \in A \) as the centers. In other words, SaTScan searches only the circles centered on each of the activity points in the dataset. Whenever a point is selected as a circle center, distances from the rest of the points to this center are used as radii to enumerate the circles. For example given an activity set with \( |A| = 100 \) activity points, \( |A|(|A| - 1) = 10099 = 9900 \) circles are enumerated.

**Step 2 - Log Likelihood Ratio Test:** The computation of \( Log \ LR_z \) requires the area and the number of points inside a zone \( Z \). Suppose the zones are enumerated as circles in Step 1. Using the area of each circle and number of activity points inside, Log Likelihood Ratio of circles are computed. Note that among the overlapping circles, SaTScan returns only the circle with the highest Log Likelihood Ratio since most clusters of this type provide little additional information, but their existence means that while it is possible to pinpoint the general location of a cluster, its exact boundaries must remain uncertain (Kulldorff, 2010).

For example the crime dataset set in Figure 2 First Degree Murder cases in Chicago between March 2013 - March 2014. Dataset includes 384 Homicide Cases shown in blue dots. includes 384 crime
locations and the study area S is a 0.09 square degrees. The area of the circular hotspot in the north is \( \pi r^2 = 3.140.047^2 = 0.0069 \) and thus \( B = 29.39 \) and likelihood ratio is 48.268

**Step 3 - Monte Carlo Simulation and Hypothesis Test:** For the circles created in Step 1, a p-value is computed by doing a Monte Carlo simulation. First, \( m \) random datasets with Poisson distribution are generated. For each random dataset, new circles are enumerated and the maximum log likelihood ratio of each random dataset is stored in decreasing order. Then, log likelihood ratio of each circle enumerated in Step is compared with this list and its corresponding position is determined. This position is divided into the number of Monte Carlo simulations \( m \) to determine the p-value of a circle. Finally, all non-overlapping circles with p-value lower than the specified significance threshold \( \alpha_p \), are returned by the algorithm.

**Limitations:**

As demonstrated, SaTScan relies on points as the centers. However, in cases where the center is sparse, SaTScan may not recognize a hotspot. Therefore, in cases when the contiguities of circular hotspots are disturbed by geographic features (i.e. rivers, roads, jurisdiction boundaries), these hotspots may not be detected by SaTScan. In addition, elliptical or imperfect circular (i.e. quarter/half circle) hotspots caused by geographic features may also be missed by this approach.

Another issue arises when two points occur too close to each other. SaTScan may detect them as significant hotspots since the log likelihood ratio function depends on the number of activity points and the area of the hotspot.

Finally, a hotspot detected by SaTScan has strict boundaries. However, most crimes tend to show a diffusion process where moving further from the center results in fewer observations of activities. Therefore, strict boundaries may not fully represent a crime hotspot. For these reasons, SaTScan should not be thought of as the sole tool to determine crime hotspots.

**RING-SHAPED HOTSPOT DETECTION**

In this section, ring shaped hotspot detection with a statistical significance will be defined and it will be illustrated with an example based on a crime dataset from San Diego, California. Formally, given a set of geo-located points (e.g. crime reports) and a significance threshold, ring-shaped hotspot detection aims to find ring-shaped hotspot areas where the concentration of points inside is significantly higher than the number of points outside. Ring-shaped hotspot detection requires the following inputs specified by an analyst:

- (a) A spatial crime dataset with a collection of crimes each associated with geo-location information (i.e. Latitude, Longitude),
- (b) A statistical significance threshold \( \alpha_p \).

Based on the above inputs, the goal of ring-shaped hotspot detection is to report a collection of ring shaped hotspot areas with p-value lower than the specified statistical significance threshold \( \alpha_p \).

In geographic profiling, two concepts are widely used to predict a potential location for a criminal. These concepts are distance decay and buffer zone as shown in the Figure 4 Distance Decay and Buffer Zone as described in Environmental Criminology(Burn, 1982). Distance decay is the result of a basic behavior caused by the requirements of time, effort and money to travel. Distance decay explains why most crimes occur relatively close to the offender’s home (Brantingham & Brantingham, 1993). For example, 70% of arson crimes occur within two miles of a serial arsonist’s home (Estepp, 1987). Second concept is the buffer zone, which is the basic result of the anonymity principle. In other words, buffer zone is an area where crimes are less likely, because of the increased risk of recognition by the neighbors. Geographic profiling uses the opposing effects of distance decay and buffer zone to predict the potential
residence (location) of a criminal. Ring-shaped hotspots are also known as doughnut hole patterns or distance-decay zones.

Previous work on doughnut hole patterns focused on serial crime and used a formulation known as Rossmo’s formula (Rossmo, 1995). Although it is a widely accepted method to determine the search area for a serial criminal’s residence, Rossmo’s formula has some limitations. It assumes a single crime source (e.g. serial criminal) and it lacks a statistical significance test.

![Diagram of doughnut hole pattern](image)

Figure 4 Distance Decay and Buffer Zone as described in Environmental Criminology (Burn, 1982)

**Case Study on a Real Crime Data**

Similar to the Circular Hotspot Detection, Ring-shaped hotspot detection (RHD) uses an activity set (i.e. crime locations) as input to detect ring-shaped hotspots in the study area. For example, Figure 5 Methamphetamine possession (48 cases shown in red dots) and Methamphetamine Manufacture/Deliver Crimes (6 cases shown in yellow stars) committed in Chicago between March 2013 - March 2014 shows a methamphetamine crime dataset collected from Chicago, Illinois (Chicago, 2014). The study area is selected as the minimum bounding orthogonal rectangle of the activity set. Activity set includes a sample of 48 methamphetamine possession and 6 methamphetamine manufacture/deliver crimes committed in Chicago, Illinois between March 2013 and March 2014. Note that the study area could be the territorial area of responsibility for a public safety agency. The aim is to find ring-shaped hotspots with a statistical significance level of $\alpha_p = 0.001$. Figure 6 Output of Ring Shaped Hotspot Detection with 5 ring-shaped hotspots enumerated from the Methamphetamine crime dataset in Figure 5 shows the output of ring shaped hotspot detection with 5 hotspot areas with high log likelihood ratios. Four of these hotspots have p-value = 0.001 which indicates statistical significance at 99.9 % confidence level.
Figure 5 Methamphetamine possession (48 cases shown in red dots) and Methamphetamine Manufacture/Deliver Crimes (6 cases shown in yellow stars) committed in Chicago between March 2013 between March 2014.

Figure 6 Output of Ring Shaped Hotspot Detection with 5 ring-shaped hotspots enumerated from the Methamphetamine crime dataset in Figure 5

Ring-Shaped Hotspot Detection Algorithm

A ring shaped hotspot can be defined by a variety of techniques; here a ring is defined as a concentric ring (c-ring) (Efthelioglu et al., 2014) in this section.

Basic Concepts

C-Ring: Any three non-collinear activities $a_1, a_2, a_3 \in A$ can uniquely identify a circle $circle_i$. For any activity point $a \in A$ outside $circle_i$, we can identify another circle $circle_o$ where $a$ is on the circle
and \( \text{circle}_o \) is concentric with \( \text{circle}_i \). The shape between any pair of such concentric circles \( \text{circle}_i \); and \( \text{circle}_o \), is defined as a C-Ring, denoted as \( R \). Each C-Ring \( R \) has four parameters: the \( x \) and \( y \) coordinate of the center, the outer circle radius \( r_o \), and the inner circle radius \( r_i \). The area of the shape \( R \) is \( \text{area}(R) \).

Note that, for a given inner circle \( \text{circle}_i \) with \((x, y, r_i)\), there are many outer circles \( \text{circle}_o \) (each defined by a distinct activity point \( a \) outside inner circle) and many c-rings \( R \). (Eftelioglu et al., 2014)

**Naïve Ring Shaped Hotspot Detection Algorithm**

A naïve approach to ring shaped hotspot detection uses a three-step algorithm to detect statistically significant hotspots: (1) Enumeration of C-Rings, (2) Log likelihood ratio test and (3) Monte Carlo simulation and Hypothesis test. This chapter describes these three steps without any algorithmic refinements. In addition to the naïve algorithm, a faster approach with prune and refine steps will be shown.

**Step 1-Enumeration of Rings:** Every three non-collinear points can define a unique circle. Using this notion, RHD enumerates every three non-linear points to generate the inner circles \( \text{circle}_i \) of c-rings. Once the inner circles are enumerated, using every activity point out of the inner circle, the outer circles are enumerated by using the distances from the inner circle center as outer circle radii. Using inner circles and outer circles c-rings \( R \) are generated.

In RHD, since every three activity point creates an inner circle, the enumeration space for candidate c-rings is much larger than the enumeration space of circular hotspot detection.

**Step 2 - Log Likelihood Ratio Test:** Similar to the circular hotspot detection, the computation of \( \log LR_z \) requires the area and the number of points inside a zone \( Z \). Suppose the zones are enumerated as ring-shaped hotspots in Step 1. Using the area of each c-ring and number of activity points inside, Log Likelihood Ratio of c-rings are computed.

**Step 3 - Monte Carlo Simulation and Hypothesis Test:** Monte Carlo simulation uses as similar algorithm to get the distribution of the log likelihood ratios under the null hypothesis. However, in Monte Carlo simulation, instead of the original input activity set \( A \), random activity sets are generated under the null hypothesis and then Step 1 and Step 2 of the algorithm are run on these datasets. Among all c-rings enumerated using the random activity sets, only the c-ring with the highest likelihood ratio is stored for each random activity set. Then, the log likelihood ratio of each c-ring enumerated from the original activity set \( A \) is compared with this list and its corresponding position is determined. This position is divided into the number of Monte Carlo simulations \( m \) to determine the p-value of a c-ring. Finally, all non-overlapping c-rings with a p-value lower than the specified significance threshold \( \alpha_p \), are returned by the algorithm.

In a naïve approach to RHD problem, c-rings are enumerated by using circumcircle of all possible triplets of activities. Since each inner circle is created by three activities, outer circles are enumerated starting from the fourth activity. Thus the total number of inner circles is \( \binom{|A|}{3} \) where \(|A|\) is the cardinality of the activity set \( A \). For example given an activity set with \(|A|=100\) activity points, \( \binom{100}{3} = 161700 \) inner circles are enumerated. Additionally, outer circles are enumerated using every point outside every inner circle. Therefore, the total number of candidate c-rings is high. For example using a activity set of 1000 crime incidents, this number increases to approximately \(10^{12}\) c-rings and computation time becomes exorbitant.

**Dual Grid Based Pruning Algorithm**
In order to address scalability limitation caused by the high number of c-rings generated by a naïve approach, Dual Grid based Pruning (DGP) algorithm namely DGP is proposed (Eftelioglu et al., 2014).

Simply, DGP algorithm defines an upper bound on the likelihood ratio of the c-rings. The **pruning phase** uses a dual grid approach, which keeps the number of activities in a geometric grid and enumerates c-rings in a parametric grid. For each ring enumerated in the parametric grid space an upper bound likelihood ratio is computed. Note that a c-ring in the parametric space represents a collection of actual c-rings in Euclidean space. Once the upper bound likelihood ratios are computed, the parametric grid cells which survive a given log likelihood ratio threshold are saved and the associated activity points are sent to the refine phase. **Refine phase** enumerates the actual c-rings and computes their likelihood ratios using the same enumeration technique described in the naïve approach. Finally, the candidate c-rings are tested for statistical significance using **Monte Carlo simulation phase**.

**Limitations:**

The enumeration technique used in ring-shaped hotspot detection is costly. Although the DGP algorithm reduces the cost significantly, its refine phase still uses the same enumeration technique whereby triplets of points are used to enumerate inner circles and a fourth point to enumerate outer circles. An algorithm which does not do extensive enumeration of rings would be beneficial for efficient detection of c-rings.

A second limitation of RHD is that rings are assumed to be concentric. This is a very restrictive assumption. Imperfect rings and non-concentric rings which have occurred due to geographic features should be taken into account.

Finally, a ring shaped hotspot detected by Ring Shaped Hotspot Detection has strict boundaries. However, the diffusion of crimes described in circular hotspot detection holds in ring-shaped hotspots as well. Therefore, most crimes tend to show a diffusion process where moving further from the inner circle will cause less observation of activities.

Finally, a hotspot detected by Ring Shaped Hotspot Detection has strict boundaries. However, the diffusion of crime activity described in circular hotspot detection holds for ring-shaped hotspots as well. Since most crimes tend to show a diffusion process, moving further from the inner circle will cause less observation of activities. However, ring shaped hotspot detection does not take this into account.

**LINEAR HOTSPOT DETECTION**

The problem of linear hotspot detection (LHD) focuses on hotspots on network space. This problem is motivated by the fact that many types of crime only happen on road networks (e.g., street robbery). This section is organized as follows: First, a formal definition to the linear hotspot detection is introduced. Second, a motivating example for Linear Hotspot Detection is shown by a Case study on real data. Next, algorithms to solve the linear hotspot detection problem are introduced. Finally, the limitations of the proposed approaches are discussed.

Formally, given a spatial network (i.e. road network), a set of geo-located activities (i.e. crime reports), and a significance threshold, linear hotspot detection aims to find linear hotspots where the concentration of activities is significantly higher than other road segments. Detection of linear hotspots requires the following inputs specified by an analyst:

(a) A spatial network (i.e. road network)
(b) A spatial crime dataset with a collection of crimes each associated with a road segment,
(c) A statistical significance threshold ($\alpha_p$).

Based on the above inputs, the goal of linear hotspot detection is to report a collection of linear hotspot areas with p-value lower than the specified statistical significance threshold ($\alpha_p$).
Case Study on a Real Crime Data

Linear hotspot detection will be illustrated with an example based on a crime dataset collected from Orlando, Florida in May 2015. Linear hotspot detection (LHD) uses activity locations and road network as input to detect the paths (road segments) where the number of activities is significantly higher than other road segments. For example, Figure 7 Street Robbery cases in Orlando, Florida in May 2015. Crime dataset includes 44 reported theft cases as shown in red shows 44 street robbery crimes occurred in Orlando, Florida in May 2015. In this example, crime activities are shown by red dots (Orlando & FL, 2014). The input spatial network is the road network of the Orlando area. The aim is to find the linear hotspot areas with a statistical significance level of $\alpha_p = 0.03$. Figure 8 Output of Linear Hotspot Detection shows the output of Linear Hotspot detection with two significant paths. These paths are shown in blue and green and p-value is 0.02 and 0.01 respectively, which indicates statistical significance at 97% confidence level.

Figure 7 Street Robbery cases in Orlando, Florida in May 2015. Crime dataset includes 44 reported theft cases as shown in red

Figure 8 Output of Linear Hotspot Detection
Linear Hotspot Detection Algorithm

Since linear hotspot detection is slightly different from hotspot detection in Euclidean space, related basic concepts will be introduced first, then a naïve and enhanced algorithm will be presented.

Basic concepts:

Spatial network: $G = (N, E)$ consists of a node set $N$ and an edge set $E$, where each element $u$ in $N$ is associated with a pair of real numbers $(x, y)$ representing the spatial location of the node in a Euclidean plane. Edge set $E$ is a subset of the cross product $N \times N$. Each element $e = (u, v)$ in $E$ is an edge that joins node $u$ to node $v$. Each edge has a weight $w$, which may represent its length or traffic density, etc. Note that each activity in the activity set $A$ is associated with a specific edge in the spatial network $G = (N, E)$.

Distance: Distance between any nodes in the spatial network is defined by the shortest path between the nodes. The triangle inequality of distance metric holds for this definition. Note that, the shortest path between two nodes could be not unique because there may be several paths with the same weight. However, the distance between two nodes is unique.

Figure 9 An example spatial network shows an example of spatial network, which contains 7 nodes denoted by $1, N2, \ldots, N7$ and 7 edges. Each edge is associated with 2 numbers $(a, w)$, where $a$ represents the number of activities associated with that edge, and $w$ represents the weight of the edge. For the simplicity, the edge weights are set 1. For example, edge $(N1, N2)$ has 6 activities and a weight of 1.

![Figure 9 An example spatial network](image)

Likelihood Ratio of a Path: In the linear hotspot detection, the likelihood ratio function is used as described in Linear Semantic Likelihood Ratio computation (Janeja & Atluri, 2005b). The linear semantic likelihood ratio ($LR_p$) function is described as following:

$$LR_p = \frac{\frac{a_p}{w_p}}{\frac{A - a_p}{W}}$$
As can be seen from the equation, the likelihood ratio of a path \( p \), is the ratio of the activity density inside path \( p \) to the activity density outside \( p \). In Figure 9 An example spatial network, likelihood ratio of the path \( N_1, N_2, N_3 \) can be computed as \( LR_p = \frac{11}{2} = 3.05 \).  

Note that although likelihood ratio of a path described in this section is different that the log likelihood ratio function described in Circular and Ring-Shaped Hotspot Detection, both functions can be used to assess the test statistic of an enumerated path.

**Naïve Linear Hotspot Detection Algorithm** (Oliver et al., 2014)  

The basic idea behind linear hotspot detection algorithm is to find all statistically significant shortest paths in the spatial network whose likelihood exceeds \( \theta \). Note that the shortest paths returned are constrained so that they are not sub-paths of any other path in the output. This constraint aims to improve solution quality by reducing redundancy in the paths returned. The output is also constrained such that the shortest paths returned start and end with active nodes. This constraint also aims to improve solution quality by ignoring edges at the start and/or end of a path that do not have any activities.

Similar to the previous algorithms Naïve Linear Hotspot Detection algorithm (Oliver et al., 2014) consists of three steps.  

**Step 1-Enumeration of Road Segments:** In its first step naïve linear hotspot detection algorithm gets the spatial network as input and calculates the all pair shortest path for this spatial network.

**Step 2-Log Likelihood Ratio Test:** For each enumerated road segment, this step computes the likelihood ratio if the starting and ending nodes are active nodes, meaning that the road segments should have activities associated with them. If any of the road segments has likelihood ratio higher than the specified threshold, they are saved into a candidate list for hypothesis test.

**Step 3 - Monte Carlo Simulation and Hypothesis Test:** In the Monte Carlo simulations, each activity in the original spatial network graph \( G \) is randomly associated with an edge so that the number of activities on each edge is shuffled, forming a new graph \( G_s \). Note that all the activities in \( G \) are present in \( G_s \), with no activities added or removed; the original activities in \( G \) are now shuffled so they may be on different edges in \( G_s \). Then the highest likelihood of randomized \( G_s \) is compared with the highest likelihood ratio of original \( G \). If the original one is smaller, then \( p = p + 1 \). The above process repeats \( m \) times and after it terminates, the p-value is subsequently \( p/m \). Paths whose p-values are less than or equal to the given p-value threshold are deemed statistically significant and the null hypothesis is rejected.

LHD problem is challenging from computational perspective for several reasons. First, given a spatial network, the computation of all pair shortest paths is a costly operation. Moreover, the randomization test with Monte Carlo simulations (typically with 1000 trials) multiplies this cost. Second, log likelihood ratio function does not have monotonicity, which means given a path, its log likelihood ratio depends on both the number of activities and the weight of the path. Thus, monotonicity based pruning algorithms do not apply in LHD problem.

**Smart Significant Route Miner Algorithm** (Oliver et al., 2014)  

A likelihood ratio pruning based algorithm is proposed to overcome the challenges of LHD problem. SmartSRM proposes an upper bound pruning approach to avoid computing all shortest paths of a graph. The framework of SmartSRM is similar to Dijkstra’s algorithm, which uses a shortest path tree to represent the shortest paths from a source node (root of the tree) to all the other nodes. It stops growing a branch in the tree if the likelihood ratios of any shortest path on that branch must smaller than a specified likelihood ratio threshold \( \alpha \). SmartSRM proposes a novel linear likelihood ratio upper bound to
determine such branches. In addition, a Monte Carlo Simulation (MCS) early stop algorithm is used in SmartSRM. The key idea is stopping a MCS trial once a linear likelihood ratio is found higher than the highest one in the candidate paths.

**Limitations:**

Linear hotspot detection (LHD) has a high computational cost due to the extensive all pair shortest path enumeration. Moreover, Monte Carlo simulation multiplies this cost. Another problem with LHD arises from the definition of a Linear Hotspot. LHD defines linear hotspots as starting and ending with Nodes. However, long paths between two nodes may cause missing a dense linear hotspot if the activities are in a specific part of a path. For instance in Figure 9, although activities are dense in a part of the returned path LHD returned a longer path which starts and ends with Nodes in the Spatial Network. Finally, some road crimes occur when a criminal travels from one location to another, causing a linear hotspot, which moves in the network over time. However, LHD described in this section can not evaluate linear hotspots that takes the temporal dimension into account.

**CURRENT HOTSPOT DETECTION TOOLS**

This chapter focused on three representative techniques that use spatial scan statistics for crime hotspot detection. However, these are just the tip of the iceberg and there are many tools that modern police enforcement uses for crime mitigation as well as research on hotspot detection that are going on. In the following paragraphs, those tools will be presented. More detailed information on spatial aspect of crime can be found in (Brimicombe, 2005; Cohen & Felson, 1979; Gonzales, Schofield, & Hart, 2005; Levine, 2006; Ratcliffe & McCullagh, 1999; Wang et al., 2013; Yu, Ward, Morabito, & Ding, 2011).

Today, public security officials use crime analysis software for to improve their efficiency and reaction time (Spencer Chainey & Ratcliffe, 2013; Gonzales et al., 2005). Software, which are used by public security officials, depends on the crime type (i.e., fraud, cyber crime, gun related crime). For example, a fraud detection software may audit the stock exchange movements and determine fraudulent activity on stock market whereas a cyber crime software may focus on the detection of cyber crimes. Since this chapter is focused on crime hotspot detection, the following examples will focus on the “spatial” aspect of crime.

GIS tools (ESRI ArcGIS, QGIS, Python GDAL/OGR, etc.) can be used to determine locations for new police deployment, detect suspicious activities, integrate 911 emergency call locations, perform resource (i.e. patrol) allocation to incidents, and analyze the crime locations (Mitchell, 1999). Modern GIS software integrates the statistical analysis tools in order to determine hotspots of crime as well as the days and times of high crime activities. Although some crime specific features may be added to those GIS software (Goetz, n.d.), they are not specialized in crime activities. Therefore, several softwares (i.e. CrimeStat) were developed to analyze the spatial aspect of crime. CrimeStat (Lavine, 2013), a software package for spatial analysis of crime locations, incorporates several methods to determine the crime hotspots in a study area. CrimeStat package has k-means tool, nearest neighbor hierarchical (NNH) clustering, Risk Adjusted NNH (RANNH) tool, STAC Hot Spot Area tool, and a Local Indicator of Spatial Association (LISA) tool that are used to evaluate potential hotspot areas (Leong K, 2015). Urban Growth Simulator (UGS), which is developed at Kent State University, aims to estimate changes in crime rates as induced by urban growth (Lee, Brody, Zhang, Kim, & Bradac, 2002). Rigel software, which depends on Rossmo’s Formula, uses crime locations, suspect information, case details and investigator details to analyze a series of linked crimes and determine the most probable locations of the offender's residence (Beauregard, Proulx, & Rossmo, 2005; Devlin & Lorden, 2007). Hotspot Detective, which is (was) an add-on for MapInfo Software, uses a grid based approach together with kernel density estimation to determine hotspot locations (S Chainey, Tompson, & Uhlig, 2008).
Apart from the software above, several spatial statistical analysis tools are also used for the analysis of the spatial aspect of crime. For example, GeoDa, a spatial data analysis software, can be used to do multivariate exploratory data analysis, global and local spatial autocorrelation analysis, and basic linear regression with crime data (Anselin, Syabri, & Kho, 2006). In addition, several R software packages (spatstat, geoR, spdep, etc.) are widely used for spatial analysis of point referenced data (e.g. crime locations) (R Development Core Team, 2011). Finally, it is worth noting that there are many web-based crime analysis and visualization tools thanks to the spatial abilities of HTML5 (Holdener, 2011).

**FUTURE RESEARCH DIRECTIONS**

Although there are a variety of techniques (Ester & others, 1996; Ilango & Mohan, 2010; Karypis, Han, & Kumar, 1999; Levine, 2006; Wu et al., 2008) which are developed to be used in different domains, the techniques described in this chapter are representative approaches to detect circular, ring-shaped and linear hotspots given an activity set. In this section, some important research topics that include methodological advancements that may be needed to analyze crime hotspots are reviewed.

Crime occurrences are always associated with a population. However, the techniques described here do not take the population into account. For example, in most of the crime datasets used in this chapter, crime activities are dense near downtowns since these locations are dense in population. When detecting crime hotspots, the underlying population should be taken into account to reflect this.

In addition, the circular and ring-shaped hotspot detection techniques described here are based on the Euclidean space. However, people tend to use road networks to travel. Similarly, criminals tend to use the road network to travel to the crime sites. Thus, even a hotspot detection technique which focuses on circular hotspots should take road network distances into account. Detecting hotspots on a road network may provide additional insights that crime analysts can use to identify patrolling districts after taking a variety of other factors into consideration. Linear hotspot detection technique described in this section considers spatial network nodes as the starting and ending nodes of a significant road path. In order to detect partial road segments with high number of activities, activities on a spatial network can be considered as dynamic nodes, and the edges can be segmented accordingly. Linear hotspot detection on this new “dynamic segmented” spatial network may be able to give more precise location of hotspots. Techniques described in this chapter focus on the spatial dimensions of crime hotspots in either in Euclidean Space or Spatial Networks. However, new hotspot detection techniques are required which can evaluate the temporal space and detect hotspots using a constant flow of new crime activities.

Furthermore, fuzzy statistically significant hotspot detection techniques are needed to better represent the diffusion of crimes. Techniques described in this chapter detect crime hotspot areas with strict boundaries. Since the crime activities are similar to the diffusion processes in Physics and Chemistry, the crime activities become sparser when moving further from the center. Therefore, new fuzzy hotspot detection techniques are needed.

Given the current rate of crimes each year (e.g. annual 10^7 crimes U.S.), any crime hotspot detection technique should also consider the scalability of the designed algorithms. For example, the circular hotspot detection technique (i.e. SaTScan) described in this chapter is limited to a couple of thousands of activity points (i.e. crimes) due to the extensive circle enumeration using points as centers. Therefore, there is a need for new scalable algorithms to handle the high number of crime activities.

In this chapter, hotspot detection using the locations of crime incidents (geo-located crimes) is discussed. However, in the light of the technological advances, digital crimes (i.e. cyber crimes) are starting to get prevalent among crimes (Fei, Eloff, Olivier, & Venter, 2006). Such crimes include child abuse, hacking, financial theft, stealing private and/or confidential data etc. (Robert W. Taylor, Tory J. Caeti, D. Kall Loper, Eric J. Fritsch, 2006). Although these crimes do not have a specific location information (i.e., longitude, latitude), an approximate location can be captured from suspects’ internet address (i.e. IP address) (Kao & Wang, 2009). However, since the locations of these crimes will be approximate (unlike a crime location in geographic space), new hotspot detection techniques are required
to capture these events. Moreover, social media posts (Twitter, Facebook etc.) and geo-tags can be leveraged as well as the social network connections (i.e. friends, buddies) to have a sense of a criminal’s location (Surette, 2010). Such opportunities can be new frontiers for hotspot detection.

CONCLUSION

This chapter explored crime hotspot detection methods, which is an important task in environmental criminology. When evaluating hotspot detection techniques, this chapter focused on the statistically significant hotspot detection techniques, to reduce false positive rates (i.e. hotspot areas which are occurred by chance). The chapter started with the foundations of Spatial Scan Statistics and the applications using Spatial Scan Statistics to find statistically significant circular hotspots. Next, ring-shaped hotspot detection is described. Finally, since most crimes occur and diffuse along a road network, linear hotspots are illustrated. Finally, chapter is concluded with future directions and opportunities for the crime hotspot detection task.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 1029711, IIS-1320580, 0940818 and IIS-1218168, USDOD under Grant Number HM1582-08-1-0017, HM0210-13-1-0005, and University of Minnesota via U-Spatial. We would like to thank Kim Koffolt and the members of the University of Minnesota Spatial Computing Research Group for their comments.

REFERENCES


