A Generic Dual Grid Pruning Approach for Significant Hotspot Detection

Emre Eftelioglu
Department of Computer Science
University of Minnesota
Minneapolis, MN
emre@cs.umn.edu

ABSTRACT

Given a set of points in two dimensional space, statistically significant hotspot detection aims to detect locations where the concentration of points inside the hotspot is much higher than outside. Statistically significant hotspot detection is an important task in application domains such as epidemiology, ecology, criminology, etc. where it may reveal interesting information for domain experts. However, significant hotspot detection is challenging because of a lack of a generic technique for different hotspot patterns (i.e. shapes) and thus a large number of candidate hotspots to be enumerated and tested. Previous hotspot detection techniques focus on specific shapes (e.g. circles, rectangles, ellipses) to identify hotspot areas, but they cannot be used interchangeably which cause a vast variety of complicated and sometimes confusing techniques for each individual hotspot pattern. For example, a circular hotspot detection technique can not be used to discover a rectangular or an elliptical hotspot. In this paper, we propose a generic dual grid based pruning approach for hotspot detection that can be used for different hotspot patterns. We also present a cost analysis, a simplified experiment on the dataset size and a case study on a synthetic dataset to show the applicability of our proposed approach to circular hotspots.

CCS Concepts

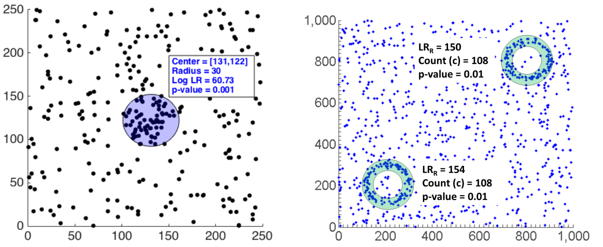
•Information systems → Geographic information systems; Data mining; Clustering;

Keywords

Spatial Data Mining, Hotspot Detection, Statistical Significance, Spatial Scan Statistics

1. INTRODUCTION

Informally, given a set of geo-located points (e.g., crime/disease locations), statistically significant hotspot detection aims to detect hotspots of points where the concentration of points inside is significantly higher than the number of points in any other place in a study area. Depending on the type of analysis and the region of interest, hotspot detection methods may focus on regions with different shaped hotspots such as circles [5, 2], ellipses, rectangles [6], rings [1], etc. For example, disease hotspots tend to be circular or elliptical as described in diffusion theory; some types of crime hotspots (e.g. serial crimes) tend to be ring-shaped. In other cases, jurisdictional or municipial boundaries may mean a hotspot of interest is rectangular-shaped. Figure 1 shows an example of detected hotspots from synthetic datasets with two different shapes (e.g. circles, rings).



(a) A sample output of circular hotspot detection [5]

(b) A sample output of ring shaped hotspot detection [1]

Figure 1: Sample output of hotspot detection techniques with different shapes for different point sets (best in color).

Application Domains: Statistically significant hotspot detection is widely used in application domains such as criminology, epidemiology, ecology, biology, etc. where detection of regions that have unexpectedly highly numbers of activities/points may reveal interesting information for domain experts [7]. For example, in epidemiology the ability to detect circular hotspots may help officials to take required precautions to prevent further diffusion of an infectious disease.

Challenges: Statistically significant hotspot detection is challenging since it is hard to enumerate all possible candidate hotspots due to the fact that the location, size and shape of hotspots are not known beforehand. Furthermore, each hotspot shape (e.g., ring, circle, rectangle, star etc.) may require an extensive shape-specific candidate enumeration which makes the hotspot detection problem more complicated and exorbitant with increasing point set sizes.

Related Work: Previous approaches for statistically significant hotspot detection are using shape-specific approaches to enumerate candidates in different domains (epidemiology with circles [5], epidemiology with rectangles [6], criminology with rings [1]). These approaches lack generality (independence of a shape) and they cannot be used interchange-

3rd ACM SIGSPATIAL PhD Workshop'16, October 31-November 03 2016, Burlingame, CA, USA

DOI: http://dx.doi.org/10.1145/3003819.3003821

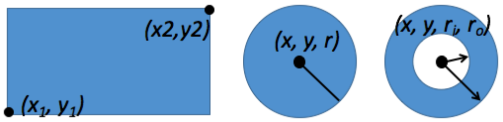


Figure 2: Example geometric shapes (i.e. rectangle, circle, ring) with their parameters.

ably for different situations. For example, SaTScan [5] uses points as centers for circular hotspots and uses the distances from the rest of the points to the centers as radii. Similarly, Ring-Shaped Hotspot Detection [1] uses every three points to create inner circles and uses the rest of the points to enumerate outer circles, thus rings. For domains where multiple hotspot patterns (i.e. different shapes) are interesting, these techniques not only get confusing and complicated but also time consuming due to their lack of generality.

2. CONTRIBUTIONS

In this paper, we present opportunities to use our dual grid based pruning approach (proposed for Ring Shaped Hotspots [1]) for different geometric shapes. Dual Grid Based Pruning Approach uses geometric and parametric grids to enumerate hotspots in a parameter space and prunes the ones that do not contribute to an actual hotspot. Since hotspots are defined in parameter space, any geometric shape (circle, rectangle, ring, ellipse, etc.) can be represented by their parameters as illustrated in Figure 2. Next, we first define basic concepts, formally introduce the problem statement and then we briefly show the steps of our algorithm.

3. PROBLEM STATEMENT

3.1 Basic Concepts

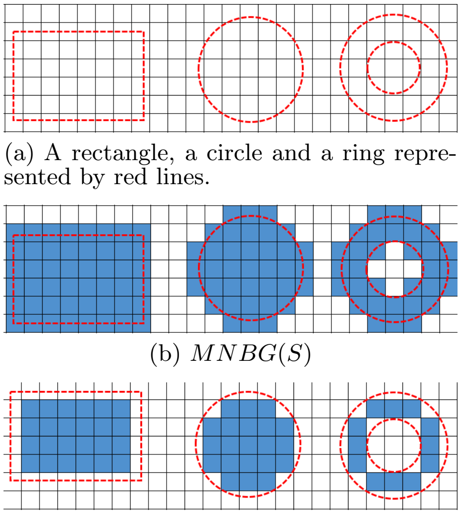
Definition 1. ***A point set*** *P is a collection of geo-located activities. A* ***point*** *$p \in P$ is associated with a pair of coordinates (x, y) representing its spatial location in 2D Euclidean space.*

Definition 2. *From mathematical perspective, a* ***geometric shape*** *(S) can be defined by n parameters. For example, a circle can be defined by $n = 3$ parameters, namely the coordinates of its center ($cent = [x, y]$) and its radius r.*

Mathematical definition of geometric shapes may introduce infinite number of candidates which is computationally infeasible to enumerate. Therefore, for implementation reasons, many approaches use assumptions to reduce the candidate enumeration to a finite space. Nevertheless, those implementation specific enumeration techniques are not central to this paper and the proposed technique can be applied to any geometric shape that can be defined with parameters.

Definition 3. *In spatial scan statistics, a candidate hotspot is evaluated by its* ***Log Likelihood Ratio*** *(Log LR). It is the interest measure used to determine the test statistic for a candidate hotspot [4] and then this test statistic is compared with the test statistic distribution (obtained by Monte Carlo simulation) to determine the significance of a candidate. The equation can be shown as:*

$$Log\ LR = Log\left(\left(\frac{c}{B}\right)^{c} \times \left(\frac{|P| - c}{|P| - B}\right)^{|P|-c} \times I()\right) \qquad (1)$$



(a) A rectangle, a circle and a ring represented by red lines.

(b) $MNBG(S)$

(c) $MXBG(S)$

Figure 3: Given a shape S, Figure 3(b) shows $MNBG(S)$ and Figure 3(c) shows $MXBG(S)$ used for $\widehat{LR}_{param}$ computation.

$$B = \frac{|P| \times area(R)}{study\ area}. \quad I() = \begin{cases} 1, & if\ c > B \\ 0, & otherwise \end{cases}$$

where B denotes the expected number of points, c denotes the observed number of points in a candidate hotspot, $|P|$ denotes the number of points of set P. The term $I()$ is an indicator function. $I() = 1$ when the candidate hotspot has more points than expected ($c > B$), otherwise it is set to 0 to prevent the detection of low activity areas [5].

Definition 4. ***A geometric space grid (geo-grid)*** *with grid cell size lg is a partitioning of the study area into a 2-dimensional geometric grid where each geo-grid cell is a square with an area of $lg \times lg$. Each geo-grid cell $cell_{geo}$ is represented by its coordinate interval $\left([x^{min}, x^{max}], [y^{min}, y^{max}]\right)$ and the count of the points inside the cell.*

Definition 5. ***Minimal bounding geo-sub-grid*** *($MNBG$): Given a shape S (e.g. rectangle, circle, ring, etc.) and a geo-grid, $MNBG(S)$ is defined as a collection of $cell_{geo}$ which overlap with S. For example, blue cells in Figure 3(b) represent $MNBG$ for the shapes shown in Figure 3(a).*

Definition 6. ***Maximal bounded geo-sub-grid*** *($MXBG$): Given a shape S and a geo-grid, $MXBG(S)$ is defined as a collection of $cell_{geo}$ that are completely inside S. For example, blue cells in 3(c) represent $MXBG$ for the shapes shown in Figure 3(a).*

Definition 7. ***A parametric space grid (param-grid)*** *is an n-dimensional grid space where each param-grid cell $cell_{param}$ represents a collection of shapes S which contain $MXBG(S)$ but are contained by $MNBG(S)$. Since shape S is enumerated in a grid space, n is the number of parameters that define the shape S. For example, a parameter space for a circle can be defined by $n = 3$ parameters: two dimensional center coordinates (x and y) and a radius (r) [3].*

Definition 8. *In order to determine an upper bound on the log likelihood ratio of a collection of candidates in parametric space (to prune candidates in parametric space), we* *created a new function namely the* ***upper bound log likelihood ratio*** *(Log $\widehat{LR_{param}}$ [1]) as follows:*

$$Log\ \widehat{LR_{param}} = Log\left(\widehat{LR_{int}} \times \widehat{LR_{ext}} \times \widehat{I()}\right),\ where$$

$$\widehat{LR_{int}} = \left(\frac{U(c)}{L(B)}\right)^{U(c)},\ and$$

$$\widehat{LR_{ext}} = \begin{cases}\left(\frac{|P|-L(c)}{|P|-U(B)}\right)^{(|P|-U(c))}, & if\ L(c) \geq U(B)\\ 1, otherwise & \end{cases}$$

$$\widehat{I()} = \begin{cases}1, & if\ U(c) > L(B)\\ 0, otherwise & \end{cases}$$

where $U(c)$ is an upper bound of c, $L(c)$ is a lower bound of c, $U(B)$ is an upper bound of B, and $L(B)$ is a lower bound of B. Those upper and lower bounds of c and B can be determined as follows:
$U(c)$ = number of points in $MNBG(S)$,
$L(c)$ = number of points in $MXBG(S)$,
$U(B) = \frac{area(MNBG(S))\times|A|}{area(S)}$ and
$L(B) = \frac{area(MXBG(S))\times|A|}{area(S)}$.

3.2 Problem Statement

The Dual Grid Based Pruning for Hotspot Detection problem can be formally stated as follows:

Given:

1. A set of points P where each point $p \in P$ has x, y coordinates in a Euclidean space,
2. A specific geometric shape S with n parameters,
3. A log likelihood ratio threshold (θ),
4. A cell length lg

Find:

prunedSets $\in P$ with $Log\ \widehat{LR_{param}} \geq \theta$.

Objective: Computational efficiency.

Constraints:

1. Correctness of the result set.
2. Hotspots should have a specific geometric shape S with n parameters,

Point set P and points $p \in P$ are defined in Definition 1, θ indicates the minimum desired log likelihood ratio and $Log\ \widehat{LR_{param}}$ is defined in Definition 8. The outputs are *prunedSets* that exceed the $Log\ \widehat{LR_{param}}$ in parameter space which can be used to create the exact candidate hotspots.

4. PROPOSED APPROACH - DUAL GRID BASED PRUNING [1]

Our approach has three phases:

- *The dual grid based pruning phase* first discretizes the space into a geo-grid by using the input cell length lg. Then the parametric grid is created by using the n parameters of a geometric shape S. Each param-grid cell represents collections of hotspots with a specific shape S in the parameter space which ensures that the test statistic (i.e. log likelihood ratio) of any candidate hotspot in this collection will be less than or equal to the computed upper bound likelihood ratio for that cell. Finally, those param-grid cells and associated points that exceed the user-defined test statistic threshold (i.e. log likelihood ratio threshold θ) are sent to a refine phase as *prunedSets*.
- *The refine phase* takes the subsets of points (i.e. *prunedSets*) that are returned by the prune phase and creates actual candidate hotspots. A test statistic (i.e. log likelihood ratio) is then computed for each candidate. It's important to note that creating actual candidate hotspots in this phase can be accomplished by using any enumeration strategy and it's outside the scope of our proposed approach.
- *Monte Carlo simulation phase* generates random point sets and runs the previous two phases to determine the highest log likelihood ratio of each. This list is then used to determine p-values of candidate hotspots.

5. COMPLEXITY ANALYSIS

The presented pruning technique aims to reduce the complexity of enumerating all candidates in a study area. The proposed algorithm's running time depends on the size of the cell length (lg) placed on the area of interest as opposed to the number of points ($|P|$) in the input.

Table 1: Computational complexity analysis

Algorithm	Complexity (best case)	Complexity (worst case)
Naive Circular Hotspot Detection	$\Omega(m \times \|P\|^3)$	$O(m \times \|P\|^3)$
Circular Hotspot Detection with the Proposed Pruning Approach	$\Omega(m \times N^3)$	$O(m \times (N^3 + \|P\|^3))$

For illustrative purposes, we provide an example cost analysis for circular hotspot detection by using a naive technique which uses one point in its circumference and one point on its center similar to [5].

The computational cost of the naive approach will be $O(m \times |P|^3)$ where m is the number of Monte Carlo simulation trials, $|P|^2$ will be enumeration cost and $|P|$ will be to count the points inside each candidate.

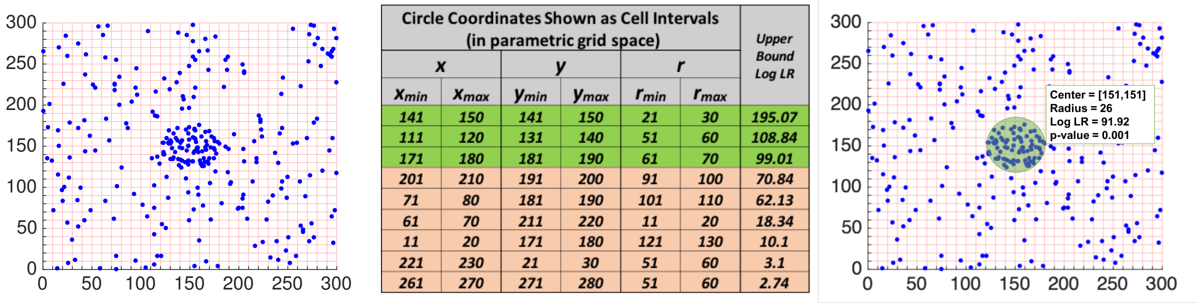
The proposed prune phase uses a dual grid based enumeration on a geometric grid space of N^2 and a parametric space of N^3, where N^2 is the number of geo-grid cells created using the grid cell size lg. The worst case cost of the proposed approach is $O(m \times (N^3 + |P|^3))$, if no pruning occurs and all points are returned to the refine phase. In the best case, the cost will be $O(m \times N^3)$ as the prune phase will not return any *prunedSet* and the algorithm will terminate.

In conclusion, it is clear that the prune phase has an extra cost and it is sensitive to the selected cell length lg. However, if $|P| > N$, the proposed approach will have the same asymptotic complexity in the worst case and it will perform faster in the best case.

6. EVALUATION

To illustrate the performance improvement over a naive approach by using the dual grid based pruning technique, we provided an experiment on the point set size $|P|$. Note that more detailed comparisons can be found in [1] and [2].

The goal of the following experiment was twofold: to evaluate the performance of the proposed approach algorithm under varying number of points and to compare its performance with SaTScan [5]. To achieve the goals, the following questions are asked: (1) How is the scalability compared to its rivals (e.g. SaTScan)? (2) How effective is the pruning



Circle Coordinates Shown as Cell Intervals (in parametric grid space)						Upper Bound Log LR
x		y		r		
x_{min}	x_{max}	y_{min}	y_{max}	r_{min}	r_{max}	
141	150	141	150	21	30	195.07
111	120	131	140	51	60	108.84
171	180	181	190	61	70	99.01
201	210	191	200	91	100	70.84
71	80	181	190	101	110	62.13
61	70	211	220	11	20	18.34
11	20	171	180	121	130	10.1
221	230	21	30	51	60	3.1
261	270	271	280	51	60	2.74

(a) Input point set with 300 points. (b) Parametric grid cells shown as rows (c) Output of DGP approach

Figure 4: Example input and output of DGP approach for a given point set P, test statistic threshold ($\theta = 90$), cell size $lg = 10$ and the parameters for a desired hotspot shape (i.e., $circle = [x, y, r]$).

step in reducing the cost of a naive algorithm? (3) How is the result quality of the proposed algorithm?

The first two questions is answered by providing a simple illustrative experiment from [2]. The last question is answered by providing a case study on a synthetic dataset.

6.1 How are the scalability and effectiveness of the proposed approach?

In this experiment, synthetic point sets were created (with cardinalities ranging from 20K to 60K) to compare SaTScan with a circular hotspot detection algorithm that use the dual grid based pruning technique (i.e. CGC [2]) and θ is selected as 10^4. Figure 5 shows that there is at least two orders of magnitude difference between SaTScan and CGC execution times. Also in Figure 5, it can be seen that CGC prune phase performs faster and most of the execution time is spent on the refine phase. Overall it can be stated that the proposed approach reduced the cost of a naive approach and CGC algorithm with dual grid based pruning performs faster than SaTScan and the trend is that the savings increase when the point set size increases.

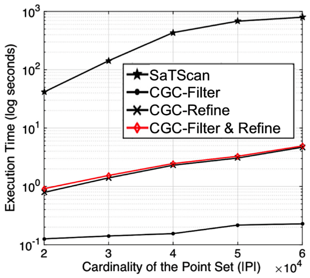


Figure 5: Scalability of an algorithm with dual grid based pruning (i.e. CGC [2]) compared with SaTScan [5] with increasing number of points in the point set [2].

6.2 How is the result quality of the proposed approach?

Consider the input point set with 300 points in a 300 × 300 study area in Figure 4(a). Since a circle can be defined by three parameters (i.e. two dimensional center coordinates and a radius), a three dimensional parametric grid with grid cell size $lg = 10$ is created for this point set as shown in Figure 4(b). Using an upper bound test statistic function (i.e. upper bound log likelihood ratio), candidate hotspots on parametric space are evaluated against a user defined log likelihood ratio threshold $\theta = 90$. Each cell in parametric grid space is shown as a row in Figure 4(b) and each parametric grid cell represents a collection of candidate hotspots whose parameters are completely inside the intervals of the parametric grid cells that represents them. Finally, only the points associated with the top three rows in Figure 4(c) which exceed the θ are sent to the refine phase (green rows). Refine phase enumerates actual candidate hotspots and their test statistic and p-values are determined. Figure 4(c) shows the output of Dual Grid Based Pruning approach with a naive refine Approach. Finally, experiments and case study on synthetic data shows that the proposed approach is a valid approach to be used for hotspot detection.

7. FUTURE WORK

In this paper, we focused on generalizing the prune phase of our dual grid based prune and refine approach for hotspot detection. In future, we plan to create a generalized refine phase. In addition, we are planning to extend our work for spatio-temporal datasets. Finally, we are planning to explore new techniques for hotspot detection in spatial networks.

8. REFERENCES

[1] E. Eftelioglu, S. Shekhar, et al. Ring-shaped hotspot detection: a summary of results. In *2014 IEEE International Conference on Data Mining*, pages 815–820. IEEE, 2014.

[2] E. Eftelioglu, X. Tang, and S. Shekhar. Geographically robust hotspot detection: A summary of results. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1447–1456. IEEE, 2015.

[3] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer vision, graphics, and image processing*, 44(1):87–116, 1988.

[4] M. Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26:1481–1496, 1997.

[5] M. Kulldorff. Satscan user guide for version 9.0, 2011.

[6] D. B. Neill and A. W. Moore. A fast multi-resolution method for detection of significant spatial disease clusters. In *Advances in Neural Information Processing Systems*, 2003.

[7] S. Shekhar et al. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193, 2011.