Crime Hotspot Mapping Using the Crime Related Factors–A Spatial Data Mining Approach

Dawei Wang[1], Wei Ding[1], Henry Lo[1], Tomasz Stepinski[2], Josue Salazar [3], and Melissa Morabito[4]

1. Department of Computer Science, University of Massachusetts Boston
2. Department of Geography, University of Cincinnati
3. Department of Computer Science, Rice University
4. Department of Criminology and Criminal Justice, University of Massachusetts Lowell

Abstract. The technique of *Hotspot Mapping* is widely used in analysing the spatial characteristics of crimes. The spatial distribution of crime is considered to be related with a variety of socio-economic and crime opportunity factors. But existing methods usually focus on the target crime density as input without utilizing these related factors. In this study we introduce a new crime hotspot mapping tool– *Hotspot Optimization Tool* (HOT). HOT is an application of spatial data miming to the field of hotspot mapping. The key component of HOT is the Geospatial Discriminative Patterns (GDPatterns) concept, which can capture the differences between two classes in a spatial dataset. Experiments are done using a real world dataset from a northeastern city in the United States and the pros and cons of utilizing related factors in hotspot mapping are discussed. Comparison studies with the Hot Spot Analysis tool implemented by Esri ArcMap 10.1 validate that HOT is capable of accurately mapping crime hotspots.

Keywords: Crime Hotspot, Hotspot Optimization Tool, Spatial Data Mining, Geospatial Discriminative Pattern

1 Introduction

Criminal activities are believed to be unevenly distributed over space. They tend to concentrate in certain places for reasons that have been explained in relation to the interaction of victims and offenders and the strength of guardianship [4]. Areas of concentrated crime are often referred to as hotspots. An accurately identified and clearly visualized crime hotspot map will significantly benefit police practices by aiding threat visualization, police resource allocation and crime prediction [3].

In practice, the occurrence of crime has been related to a variety of socio-economic and crime opportunity factors, such as population density, economic investment and arrest rate. It is reasonable to take these related factors in to

account when mapping the hotspots of a target crime. However, existing hotspot mapping techniques such as point mapping, thematic mapping, and kernel density estimation (KDE) usually focus only on target crime density. For example, the Spatial and Temporal Analysis of Crime (STAC), one of the earliest and widely used hotspot mapping software applications, uses an iterative search that identified the densest clusters of events on the map and demonstrates hotspots through standard deviational ellipses that fits the clusters. Another relatively more recent hotspot mapping applications is the Hot Spot Analysis (HSA) toolbox implemented by Esri ArcMap 10.1 [12]. HSA calculates a G_i^* statistic for the density of incidents inside spatial areas (polygons) and identifies the statistical significance of each area as a hotspot. To the best of our knowledge, none of the exist hotspot mapping tools implies the criminal related socio-economic and crime opportunity factors during the process of hotspot identification. On the other hand, recently spatial data mining has emerged as an active research tool in the studies of criminology that try to answer the questions of "why" and "where" the crime happens [16, 15]. It has been proven very powerful in identifying the linkage between target crime and its related factor.

In this paper, we combine the ideas from spatial data mining and introduce a new hotspot mapping tool, *Hotspot Optimization Tool* (HOT)(Fig. 1), to improve the identification of crime hotspot through the mining of spatial patterns composed of crime related factors. In particular, HOT initializes a hotspot map using a given threshold of target crime density, and then adaptively optimizes the hotspot boundary by mining the Geospatial Discriminative Patterns (GDPatterns) [6]–patterns that are capable of distinguishing hotspots and non-hot (normal) areas. We examine our tool using a real world crime dataset from a northeastern city in the United States. We also compare our tool with the HSA and discuss the pros and cons of utilizing related factors in hotspot mapping.

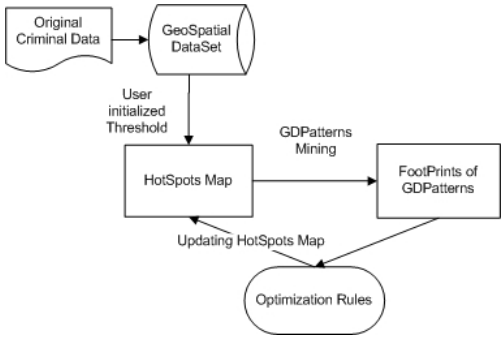

Fig. 1. The framework of *Hotspot Optimization Tool* (HOT). The boundaries of hotspots are updated using GDPatterns according to the optimization rules.

The rest of the paper is organized as follows. Related works, including the concept of HSA, are discussed in Section 2. Section 3 introduces the data repre-

sentation and formal definition of the research problems. Our HOT is also presented in section 3. Our experimental results and compared study are discussed in Section 4. In Section 5 we conclude the paper and discuss future research directions.

2 Related Work

Classic criminal theories, such as the Routine Activities Theory [4], conclude that three concepts contribute to crime: accessible and attractive targets, a pool of motivated offenders, and lack of guardianship. The concepts of "tipping point"[9] and "disorder"[19] explain why adjacent areas of crime hotspots are at higher risk. A recent work done by [18] also discusses how an area is affected by the activity scope of offenders.

GDPatterns [6] apply emerging patterns to the spatial content. Emerging patterns are first introduced in [7] and further systematically studied in [14]. In the work of [6] they adopted the relative risk ratio as the measure of pattern emergence and use the method in vegetation remote sensing datasets. In our work GDPatterns are used as a tool to spatially mine the significant difference between target crime hotspots and normal areas with respect to its underlying related factors.

The Spatial and Temporal Analysis of Crime (STAC) program [2] is one of the earliest and widely used hotspot mapping applications. STAC uses "standard deviational ellipses" to display crime hotspots on a map and does not pre-define spatial boundaries. But some studies [8] show that STAC may be misleading because hotspots do not naturally follow the shape of ellipses. Another popular hotspot representation method is thematic mapping, in which boundary areas (geographic boundaries like census blocks or uniform grids) are used as the basic mapping elements [11]. Compared to point mapping, thematic mapping uses aggregate data, and spatial details within the thematic areas are lost. Also, the identified hotspots are restricted to the shape of thematic units. Kernel density estimation (KDE) [20] aggregates point data inside a user-specified search radius and generates a continuous surface representing the density of points. It overcomes the limitation of geometric shapes but still lacks statistical robustness that can be validated in the produced map [3].

Esri ArcGIS is the most widely used Geographic Information System (GIS) and its newest component, ArcMap 10.1, includes a Hot Spot Analysis (HSA) toolbox, which provides users the ability to analyse the hotspot existed in the input spatial dataset (usually a polygon map with interested attributes). In particular, HSA will calculate a G_i^* statistic and output z-scores and p-values for the spatial areas (polygons in the map) that tell the statistically significance of the polygons. To be a statistically significant hotspot, a polygon will have a high value of the target attribute and be surrounded by other polygons with high values as well. The local sum of the attribute values for a polygon and its neighbours are compared proportionally to the sum of attribute values of all polygons. When the local sum is very different from the expected local sum (very

high z-score), and that difference is too large to be the result of random chance (very small p-value), the polygon is considered as a hotspot.

$$\begin{aligned} \overline{X} &= \frac{\sum_{j=1}^{n} x_j}{n} \\ S &= \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\overline{X})^2} \\ G_i^* &= \frac{\sum_{j=1}^{n} w_{i,j}x_j - \overline{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2)\right]}{n-1}}} \end{aligned} \tag{1}$$

where x_j is the value of the attribute (amount of incidents) for spatial polygon j, $w_{i,j}$ is the spatial weight between polygon i and j (generally, spatial weights can be calculated using different distance methods), n is the total number of polygons.

The value of the G_i^* statistic is considered as the z-score of the polygon, which in fact is the standard deviation. After calculating z-scores for all the polygons, a p-value, the probability distribution of the z-scores, is calculated for each polygon. Both very high or very low (very high absolute value) z-scores will associate with very small p-values. In summary, a polygon with a high z-score and a p-value less or equal to 0.05 will be considered as having a high enough attribute value to be statistically significant, and thus be considered a hotspot.

3 Methodology

The key insight behind our methods is searching and utilizing patterns in a geospatial space. To find GDPatterns of a target crime and its associated variables, a transaction-based geospatial database needs to be built (thereafter we use database or D refer to the transaction-based geospatial database). A widely used method for representing spatial distribution of entities is grid thematic mapping [10]. In this work we firstly generate a grid mask to cover the studied area. Variable data (both target crime and its related factors that contain information related to the occurrence of target crime) in the original spatial dataset is plotted onto a grid map with the same dimension as the mask. The cell in the grid is assigned as the count of incidents falling into it.

Since the related variables (we use the words related variables to represent the target crime related factors) come from very different sources, the range of their values varies. As with most criminal activities, the counts of cells with same values in each grid map follow a power-law distribution [5] (Fig. 4). A better way to fairly represent all the variables in one pattern is to categorize them and change the original values into categorized numbers. Jenks Optimization for Natural Breaks Classification [13], a method that is based on natural groupings inherited in data is used to divide every variable into categories. Using the Nature Break method the categories' breaks are identified that best grouping similar values, and the differences between categories are maximized.

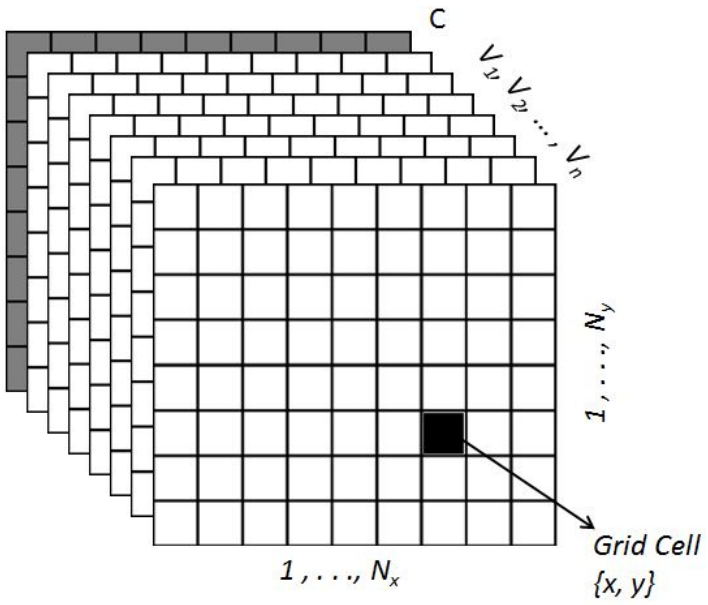

Fig. 2. An illustrative example of a transaction-based geospatial dataset D. x, y indicate the object's spatial coordinates, $V_1, V_2, ..., V_n$ represent the related variables, and C represents the target crime.

Definition 1 *Geospatial database object*: A geospatial database object is a tuple of the form: $\{x, y, V_1, V_2, ..., V_n, C\}$, where x, y indicate the object's spatial coordinates, $V_1, V_2, ..., V_n$ are the categorized values of the related variables, and C is the class label of target crime.

Fig. 2 shows an illustrative example of such a database D. Using C, objects in D can be labelled into the different classes. For example, we say C is 0 if the area is not a hotspot (or normal area) and 1 if the area is a hotspot. Then the geospatial database can be divided into two parts: D_h (hotspots) if $C = 1$, or D_n (normal area) if $C = 0$.

3.1 Geospatial Discriminative Patterns

The patterns we are looking for should meet two requirements: (1) to significantly represent the situation or conditions of related variables in objects of the database D; (2) to significantly distinguish classes (D_h, D_n) from D. Here we give a brief introduction of Closed Frequent Patterns [17], GDPatterns and related definitions.

Definition 2 *Transaction and Pattern*: In a geospatial database D, a transaction T is the group of related variables $(V_1, V_2, ..., V_n)$ in an object. A pattern X is a set of values of related variables (e.g. $\{V_1 = 1, V_3 = 4\}$).

For example, disregarding the location information (x, y) and the class label C, each object in D can be viewed as a transaction of n variable values. The database can be viewed as a set of $N_x \times N_y$ transactions.

Definition 3 *Support and Support Count*: A pattern is said to be supported by a transaction when it is a subset of the transaction. The number of transactions that support a pattern X is called the support count (suppcount) of X. The support of X is the ratio of $X's$ suppcount and the total number of transactions in a geospatial database (Formula 2).

$$sup(X) = \frac{suppcount(X)}{\tau} \tag{2}$$

where $sup(X)$ is the support of pattern X and τ is the number of transactions.

For example, in Table 1 given a transaction T_1 { AR=high, POP=low, IC=low}, patterns X_1 {AR=high, POP=low} and X_3 {AR=high} are supported by T_1, though pattern X_2 $\{AR = high, IC = high\}$ is not because it is not a subset of T_1.

Definition 4 *Closed Frequent Patterns*: A pattern whose support is above a user-defined threshold is considered frequent. A pattern X is said to be a closed frequent pattern when it is frequent and none of its immediate super-sets has exactly the same support as X.

Examples of closed patterns and closed frequent patterns are shown in Table 1. In Table 1 Pattern X_3 is not a closed pattern because X_1, its immediate superset, has exactly the same support. X_1 is a closed frequent pattern if we set the minimum support threshold $\rho = 70\%$.

Transactions	$T_1 : \{AR = high, POP = low, IC = low\}$ $T_2 : \{AR = high, POP = low, IC = high\}$ $T_3 : \{AR = high, POP = low, IC = medium\}$ $T_4 : \{AR = medium, POP = low, IC = medium\}$
Patterns	Support
$X_1 : \{AR = high, POP = low\}$	$sup(X_1) = \frac{3}{4} = 75\%(T_1, T_2, T_3)$
$X_2 : \{AR = high, IC = high\}$,	$sup(X_2) = \frac{1}{4} = 25\%(T_2)$
$X_3 : \{AR = high\}$,	$sup(X_3) = \frac{3}{4} = 75\%(T_1, T_2, T_3)$

Table 1. Examples of transactions, patterns and patterns' supports. In the examples AR, POP and IC stand for arrest rate, population density and income respectively.

A closed pattern can represent a set of non-closed patterns without losing any support information. Because the support of non-closed patterns can be calculated directly from the closed pattern. Using closed patterns will effectively reduce the total number of patterns. We are only interested in closed frequent patterns because infrequent patterns are likely to be insignificant and may happen by chance.

A closed frequent pattern can satisfy of representing the situation or conditions of related variables. To further capture the difference of classes, the patterns should also be more frequent in one class than in another.

Definition 5 *Growth Ratio*: The growth ratio of a pattern is defined as the ratio of its supports in different classes.

$$\delta = \frac{sup(X, D_h)}{sup(X, D_n)} \tag{3}$$

where δ is the growth ratio; $sup(X, D_h)$ is the supports of pattern X in class D_h and $sup(X, D_n)$ is the supports of pattern X in class D_n

Definition 6 *Geospatial Discriminating Patterns (GDPattern)*: In a geospatial database D, a closed frequent pattern X is also a GDPattern if the growth ratio(δ) of X is larger than a user defined threshold.

Hence, with a rational threshold of growth ratio the GDPatterns mined from D are significantly different between classes and are capable of digging out the meaningful information underlying the spatial distribution of target crime.

Definition 7 *Footprint of a GDPattern*: The footprint of a GDPattern X is the objects that support X in the geospatial database D. It is the set of cells whose correspondent objects support X in the grid map of study area.

Footprints of GDPatterns provide a way to measure the spatial distribution of those patterns in studied area. Examples of footprints are shown in Fig. 3.

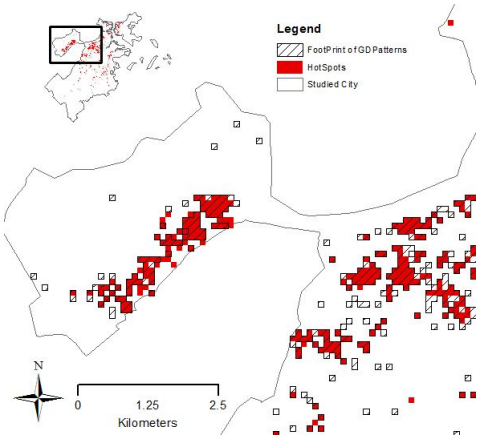

Fig. 3. An example map of GDPatterns Footprints. By selecting Residential Burglary(RB) data as the target crime, nine other variables are used as related variables from the experiment dataset and 1,500 GDPatterns are mined with a growth ratio larger than twenty. The red area are RB hotspots with a user defined threshold and hallow squares with slash lines are footprints of the 1,500 GDPatterns.

3.2 Hotspot Optimization Tool

As mentioned above, locating hotspots only with target crime density is not sufficient. Here we introduce a model, *Hotspot Optimization Tool* (HOT), to emphasize the identification of hotspots by optimizing user-specified hotspot boundaries. The practicality of HOT is based on two concepts: firstly, a hotspot can be considered as a "tipping point"[9] or the source of "disorder"[19] of its adjacent blocks, which means the adjacent areas have the possibility of being affected by crimes happening in hotspots. Also, from the point of view of spatial correlations [1], adjacent areas (cells) of a hotspot cell are more likely to fall into the active range of the same criminals. Therefore these areas (adjacent cells) are potential hotspots, especially those with a relatively high crime density. Secondly, according to the definition, GDPatterns which carry the information of related variables are much more frequent in hotspots than in normal area. Normal areas located in the footprints of GDPatterns are more likely to be hotspots because in these areas the values of relate variables are the same as in hotspots.

With a target crime being selected, to find hotspots (D_h) we firstly initialize a threshold of target crime rates. Then we optimize the boundaries of hotspot using HOT (Algorithm 1) with the intrinsic discriminative information embedded in the GDPatterns:

This algorithm takes as input a geospatial dataset D, a hotspot threshold h, a hotspot candidate threshold h', a support threshold ρ of closed frequent pattern, a growth ratio threshold δ, and returns a new set of hotspots D_h, a set of GDPatterns G, and their footprints ψ. It does the following:

- Identify areas with a relatively high crime density ($D_{h'}$, areas with high target crime density that are close to the density in hotspots, line 2);
- Mine GDPatterns based on current hotspot boundaries and draw the footprints of GDPatterns (lines 6 and 7);
- Generate candidate cells(lines 8-12): cells located in $D_{h'}$ and adjacent to some cell in D_h.
- Test the hypothesis for candidate cells (line 14): a candidate cell is inside the footprints of GDPatterns (ψ);
- If the hypothesis is true, the boundaries of the hotspot are modified by changing the current cell into a hotspot cell (from $D_{h'}$ to D_h) (line 15);
- Iterate until all hypothesis tests are fault (line 3 and line 19).

When the boundaries of a hotspot are changed, a new set of GDPatterns will be generated based on the modified hotspots, followed by the change of footprints. If in the current loop the set of GDPatterns is the same as the former loop, it means there are no new footprints and there will be no "true" from the hypothesis test (lines 4-10 in Algorithm 1). The HOT will stop and a new optimized hotspot map is generated.

Algorithm 1: *The Hotspot Optimization Tool*

Data:
h : a hotspot threshold
h' : a hotspot candidate threshold
ρ : a support threshold of closed frequent pattern
δ : a growth ratio threshold
Result:
D_h : a new set of hotspots
G : a set of GDPatterns
ψ : GDPattern footprints

```
1  count = 1;
2  Generate D_h, D_h' and D_n;
3  while count ≠ 0 do
4  |  count = 0;
5  |  μ = ∅;
6  |  G = Mine GDPatterns using D_h, ρ and δ;
7  |  ψ = footprints(G);
8  |  for cell c ∈ D_h' do
9  |  |  if c adjacent to some cell in D_h and c ∈ D'_h then
10 |  |  |  μ = μ ∪ c;
11 |  |  end
12 |  end
13 |  for cell c ∈ μ do
14 |  |  if c ∈ ψ then
15 |  |  |  D_h = D_h ∪ c;
16 |  |  |  count++;
17 |  |  end
18 |  end
19 end
```

4 Case Study

A case study of using HOT for locating and optimizing the crime hotspots is discussed in this section. Also, with the purpose of compare study, hotspot maps are drawn using HSA with the same data.

4.1 Data Preprocessing

The experiments are done using historical data with a time span of six years (2004-2009) from a northeastern city in the United States. The size of study area is 130.1 km^2 and the approximate population is 600,000. As one of the most frequently reported and resource-demanding crimes in the studied city (according to the city police department report), *Residential Burglary* (RB, burglaries target at residential houses) is selected as the target crime (Fig. 4). In addition to RB, total of eight social/criminal features are selected in this study (Table 2) as related variables with the help of domain experts. Among those are:

- *Commercial Burglary* (CB, burglaries that target at commercial sites), *Street Robbery* (SR), *Motor Vehicle Larceny* (MV, crimes against possession inside vehicles) and *Arrest Rate* (AR) are related criminal data that pictured the level of activity of crimes. The rates of CB, MV, and SR reflect the strength of guardianship in the area. AR is a good indicator for the pool of offenders.
- *Foreclosed Houses* (FC, houses that are redeemed by mortgage lender) reflect the house vacancy conditions and a vacant house has a higher risk of being broken into than an inhabited one. It is also an indicator of guardianship.
- The spatial density of RB is affected by the *Density of Population* (POP) and *Density of Houses Units* (HU). A hotspot map of RB may simply be displaying locations of high housing density [8] because such areas have a potential higher RB rate than areas with fewer houses.
- The studied city is a hub of higher education and a significant amount of houses near universities or colleges are usually rented by students or scholars, which make them easy targets of burglars during semester breaks. The variable of *Distance to Colleges* (DC) is used to address this concern.

Variables	Total Records (2005-2009)
Residential Burglary (RB)	12,020
Street Robbery (SR)	18,321
Commercial Burglary (CB)	4,438
Motor-Vehicle Larceny (MV)	29,685
Arrest (AR)	254,309
Foreclosed Houses (FC)	11,671
Population (POP)	—
Number of Houses Units (HU)	—
Distance to Colleges (DC)	—

Table 2. Crime related variables for the case study.

The original criminal dataset comes as vector maps (points and polygon). A grid map (raster map) is made as a mask to cover the whole study area and acts as the background map for data preprocessing. The cell size selected is $100m \times 100m$, which results in a number of 12,984 cells in the study area. There are two concepts to consider when choosing an appropriate cell size. Firstly, the cell is approximately half the size of average city block size ($19,873m^2$) in the studied city, which will be a good representative of reality. Secondly, with this cell size the number of cells which fall into the study area is at the same order of magnitude with the number of RB incidents, which minimizes the loss of spatial information during aggregation. Both the target crime and related variables data are converted to grid maps (rasters) with the same dimension as the mask and the values of each cell in the grids are assigned as the count of incidents falling into the cell. On the other hand, HSA needs to be conducted using polygon maps

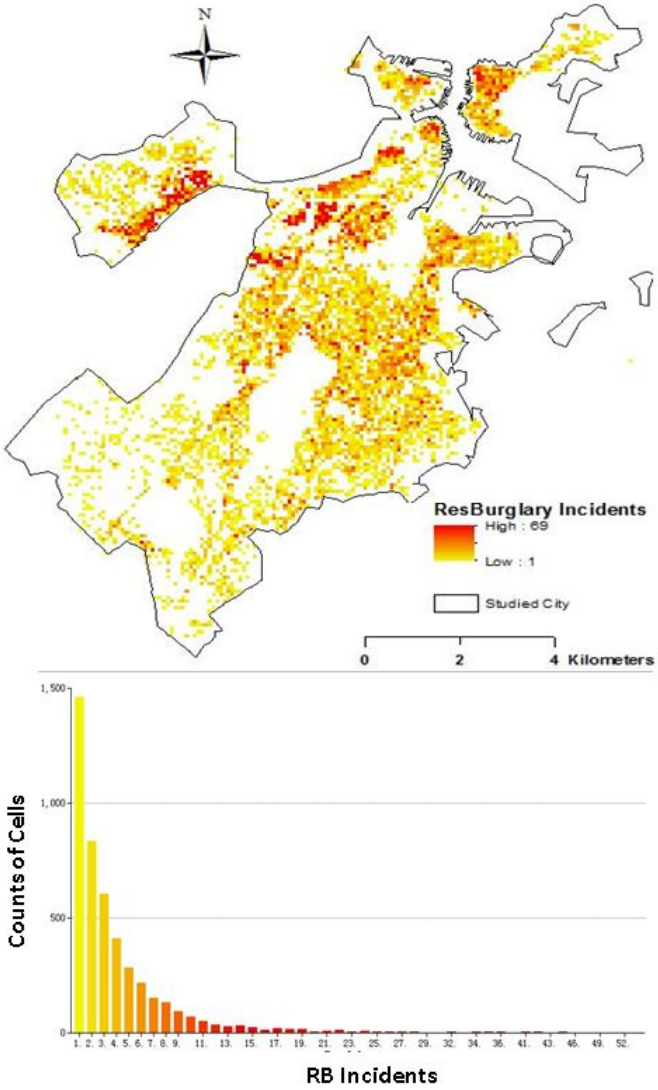

Fig. 4. Residential burglary rates in the studied city. Top is the grid density map of RB. On the bottom it is a graph showing the frequency of cell values.

instead of rasters. So the raster of RB is converted into a fishnet map with the same dimension as the mask. Each polygon in the fishnet map has an attribute of "RB Counts" indicating the amount of RB incidents happened in the area. In order to facilitate the discussion, we call the polygons in the fishnet map cells as well.

4.2 Hotspots Identification Using HOT and HSA

An initial threshold of RB hotspots is needed to set the initial classes before the HOT algorithm can be conducted. From the study of [18], a house is under a relatively higher risk if a burglary happened in the nearby area in the past four months. Relatively, if three or more burglary incidents happened in the block in one year, the area is likely a hotspot of burglary. Because the time span of our RB data is six years, we set an area (cell) to be a hotspot if there are eighteen or more burglary incidents ($h \geq 18$, Fig.5a). We use the threshold of 9 RB incidents($18 > h' \geq 9$), half of the initial value used for hotspots, to define the "potential hot" area ($D_{h'}$). The support threshold is set as 0.001. Also, growth ratios of GDPatterns are set as more than twenty ($\delta > 20$), which indicate that with an at least 95% confidence level (1:20) the mined GDPatterns will reveal the difference between hotspots and normal area.

With the about inputs, HOT is run and in the 6th loop it reaches the final condition and stops. The optimized hotspot map is drawn in (Fig.5b).

For the HSA method, we choose inverse distances as the spatial weights and *Euclidean Distance* as the distance method. A cell with positive z-value and p-value less than 0.05 is considered as a statistical significant hotspot. With the RB fishnet map as the input, a hotspot map of RB for the studied city is drawn in (Fig.5c)

4.3 Discussion

A land cover map of the studied city is draw (Fig. 6) with the purpose of evaluating the accuracy of our hotspot maps. In Table 3 we calculated the cell statistics for each map. All the three hotspot maps in Fig.5 are based on grid thematic mapping, which restricts the demonstration of hotspots. This is an intrinsic defect when using grid thematic mapping for hotspots identification. Because by converting points representing crime incidents into cells with crime counts, spatial details within and across the cells boundaries can be lost. This limitation is reflect by the fact that cells considered as non-residential areas (Fig. 6) are classified as hotspots of RB in all the three maps. The hotspot map using the user-specified threshold (HT, Fig.5a) can be considered as a benchmark for the case study. In other word, using the current grid resolution($100m \times 100m$), the accuracy for identifying residential areas when mapping hotspots in the studied city is 85.4% (Table 3). HSA does not achieve this accuracy and our HOT method outperforms HSA. Because by using the informative GDPatterns, only the areas with similar background as HT hotspots are considered. The use of

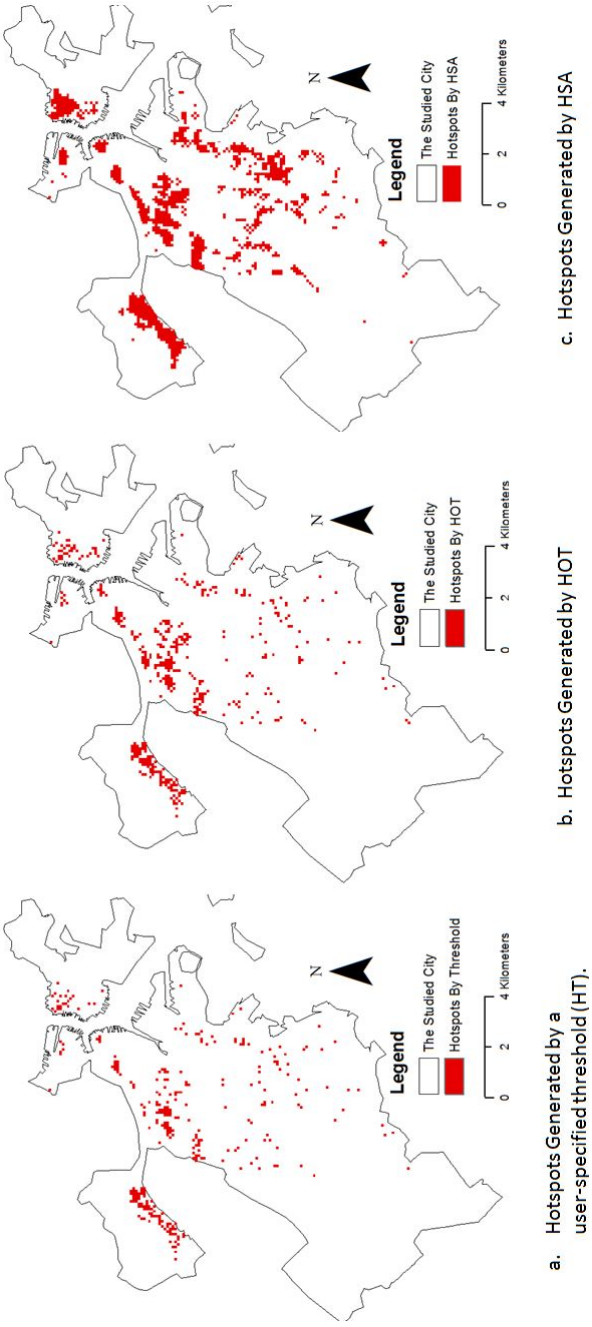

Fig. 5. RB hotspot maps of the studied city.

GDPatterns ensures that the accuracy of the generated hotspot map will consist with the original inputs.

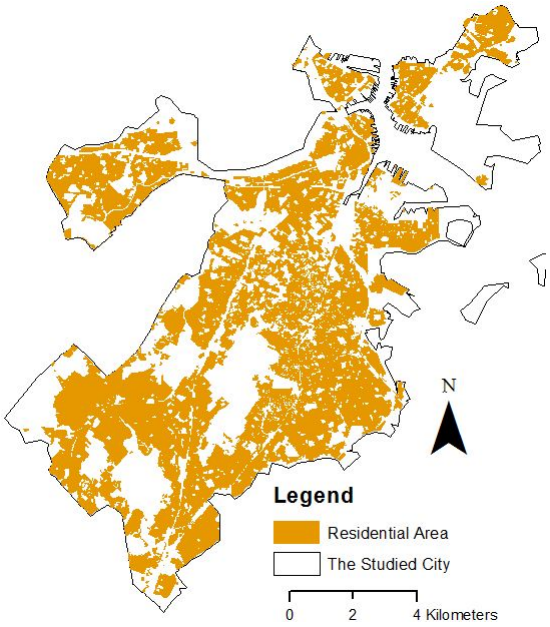

Fig. 6. A land cover map showing the residential areas in the studied city.

Hotspots Method	Total	Cells classified as Residential	Cells classified as non-residential
HT	301	257(85.4%)	44(14.6%)
HOT	429	367(85.5%)	62(14.5%)
HSA	1094	901(82.4%)	192(17.6%)

Table 3. Cell Statistic of Hotspot Maps

On the other hand, 13% of the hotspot cells identified by the hard threshold (h) is not considered as hotspot by HSA. HSA hotspots are areas with high crime density that surrounded by other cells with high values as well. The 13% cells can be seen as areas with abrupt high crime densities compared to their surrounding cells and HSA takes these 13% cells as random events. However, this may not be true for the practice of RB hotspot mapping. Because the surrounding cells with relatively low RB densities may just be areas with very few residents, like a public park. Also, the longer the studied period is, the more unlikely that those high value in the cells are happened by chance. This is the built-in limitation of

HSA because it does not consider any crime related factors when generating the hotspots. Our HOT method overcome this limitation. The hard threshold (h) in this case study is identified using experiences from previous study [18] and domain expert's advice and HOT takes the HT hotspots as a starting point. All the HT hotspots are included in the HOT hotspots.

To give an intuitive view of HOT's performance, we project a sample site extracted from the HOT hotspot map with satellite images of the studied city (Fig. 7). In Fig. 7, using the user specified threshold h the red cells are classified into hotspots and cells in same blocks (in the colour of blue) have been left out. It is reasonable that houses located in the same block have a similar risk of being broken into. Our optimization method successfully captures these cells and modifies the hotspot boundaries rationally. Also, adjacent cells mostly covered by natural land, parking lots, roads and highways are identified and have been left out of hotspots by HOT.

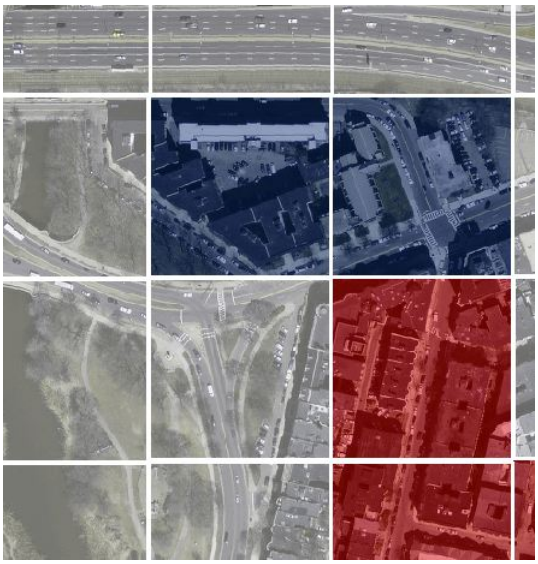

Fig. 7. A re-projection example of hotspots with satellite images. The red cells are hotspots defined by the user-specified threshold. HOT modified the original hotspot boundary and add the blue cells into hotspots.

5 Conclusion and Future Work

In this paper we present a new crime hotspot mapping tool—Hotspot Optimization Tool. Unlike existing hotspot mapping methods, HOT not only utilizes the target crime density, but also take the informative target crime related factors into account. The information inside the crime related factors are mined using spatial data mining algorithm and represented as GDPatterns. The GDPatterns

mined in the process is an information-rich dataset and from which more details of crime related factors can be extracted. Based on a user-specific threshold, HOT generating new hotspot map by optimizing the current hotspot boundaries. The hotspot mapping process is not only a visualizing of crime itself but also an visualization of those factors and will help our understanding of the underlying reasons of criminal activities. Using a real world dataset, compare studies with HSA are done and we have proved that HOT is capable of identifying crime hotspots accurately, especially for long-term studies.

6 Acknowledgement

The work was partially funded by the National Institute of Justice (No.2009-DE-BX-K219).

References

1. T.C. Bailey and A.C. Gatrell. *Interactive spatial data analysis*. Longman Scientific & Technical Essex, 1995.
2. S. Bates. Spatial and temporal analysis of crime. *Research Bulletin, April*, 1987.
3. S. Chainey, L. Tompson, and S. Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.
4. L.E. Cohen and M. Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.
5. W. Cook, P. Ormerod, and E. Cooper. Scaling behaviour in the number of criminal acts committed by individuals. *Journal of Statistical Mechanics: Theory and Experiment*, 2004:P07003, 2004.
6. W. Ding, T.F. Stepinski, and J. Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *Proceedings of SIAM*, 2009.
7. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD*, pages 43–52. ACM, 1999.
8. J.E. Eck, S. Chainey, J.G. Cameron, M. Leitner, and R.E. Wilson. *Mapping crime: Understanding hot spots*. National Institute of Justice, 2005.
9. M. Gladwell. *The tipping point: How little things can make a big difference*. Little, Brown and Company, 2000.
10. K.D. Harries. *Mapping crime: Principle and practice*. US Dept. of Justice, Office of Justice Programs, Crime Mapping Research Center, 1999.
11. A. Hirschfield. *Mapping and Analysing Crime Data: Lessons from research and practice*. CRC, 2001.
12. http://www.esri.com/software/arcgis.
13. G.F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967.
14. J. Li, G. Liu, and L. Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *Proceedings of the 13th ACM SIGKDD*, pages 430–439. ACM, 2007.
15. J. Mennis and D. Guo. Spatial data mining and geographic knowledge discovery–an introduction. *Computers, Environment and Urban Systems*, 33(6):403–408, 2009.

16. A.T. Murray, I. McGuffog, J.S. Western, and P. Mullins. Exploratory spatial data analysis techniques for examining urban crime. *British Journal of Criminology*, 41(2):309–329, 2001.
17. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *Database Theory*, pages 398–416, 1999.
18. MB Short, AL Bertozzi, and PJ Brantingham. Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression. *Journal on Applied Dynamical Systems*, 9:462, 2010.
19. W.G. Skogan. *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ of California Pr., 1992.
20. M.P. Wand and M.C. Jones. *Kernel smoothing*, volume 60. Chapman & Hall/CRC, 1995.