**Applied Intelligence**  
**Crime Hotspot Mapping Using the Crime Related Factors--A Spatial Data Mining Approach**  
--Manuscript Draft--

<table>
<thead>
<tr>
<th>Manuscript Number:</th>
<th>APIN-D-12-00188</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full Title:</strong></td>
<td>Crime Hotspot Mapping Using the Crime Related Factors--A Spatial Data Mining Approach</td>
</tr>
<tr>
<td><strong>Article Type:</strong></td>
<td>SI: Frontiers of Applied Intelligence</td>
</tr>
<tr>
<td><strong>Keywords:</strong></td>
<td>Crime Hotspot; Hotspot Optimization Tool; Spatial Data Mining; Geospatial Discriminative Pattern</td>
</tr>
</tbody>
</table>
| **Corresponding Author:** | Dawei Wang  
University of Massachusetts Boston  
Boston, MA UNITED STATES |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author’s Institution:** | University of Massachusetts Boston |
| **Corresponding Author’s Secondary Institution:** | |
| **First Author:**  | Dawei Wang |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Dawei Wang  
Wei Ding, PhD  
Henry Lo  
Tomasz Stepinski  
Josue Salazar  
Melissa Morabito |
| **Order of Authors Secondary Information:** | |
Crime Hotspot Mapping Using the Crime Related Factors--A Spatial Data Mining Approach

Dawei Wang¹, Wei Ding¹, Henry Lo¹, Tomasz Stepinski², Josue Salazar³, and Melissa Morabito⁴

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 mining 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 den-
sity 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) tool-
box 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 statis-
tical 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 intro-
duce a new hotspot mapping tool, Hotspot Optimization Tool (HOT)(Fig. 1), to
improve the identification of crime hotspot through the mining of spatial pat-
terns composed of crime related factors. In particular, HOT initializes a hotspot
map using a given threshold of target crime density, and then adaptively optimi-
izes 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.](image)

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 \(G^*_s\) 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.

\[
X = \frac{\sum_{j=1}^{n} x_j}{n}
\]

\[
S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\overline{X})^2}
\]

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

(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).

\[
\text{sup}(X) = \frac{\text{suppcount}(X)}{\tau}
\]  

(2)

where \(\text{sup}(X)\) is the support of pattern \(X\) and \(\tau\) is the number of transactions.

For example, in Table 1 given a transaction \(T_1\) \{\(AR=\text{high}, POP=\text{low}, IC=\text{low}\)\}, patterns \(X_1\) \{\(AR=\text{high}, POP=\text{low}\)\} and \(X_3\) \{\(AR=\text{high}\)\} are supported by \(T_1\), though pattern \(X_2\) \{\(AR=\text{high}, IC=\text{high}\)\} 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 \(\rho = 70\%\).

<table>
<thead>
<tr>
<th>Transactions</th>
<th>(T_1) : {(AR=\text{high}, POP=\text{low}, IC=\text{low})}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T_2)</td>
<td>{(AR=\text{high}, POP=\text{low}, IC=\text{high})}</td>
</tr>
<tr>
<td>(T_3)</td>
<td>{(AR=\text{high}, POP=\text{low}, IC=\text{medium})}</td>
</tr>
<tr>
<td>(T_4)</td>
<td>{(AR=\text{medium}, POP=\text{low}, IC=\text{medium})}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>(\text{sup}(X_1) = \frac{3}{4} = 75% (T_1, T_2, T_3))</td>
</tr>
<tr>
<td>(X_2)</td>
<td>(\text{sup}(X_2) = \frac{1}{4} = 25% (T_2))</td>
</tr>
<tr>
<td>(X_3)</td>
<td>(\text{sup}(X_3) = \frac{3}{4} = 75% (T_1, T_2, T_3))</td>
</tr>
</tbody>
</table>

**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.

\[ \delta = \frac{\text{sup}(X, D_h)}{\text{sup}(X, D_n)} \]  

where \( \delta \) is the growth ratio; \( \text{sup}(X, D_h) \) is the supports of pattern \( X \) in class \( D_h \) and \( \text{sup}(X, D_n) \) 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(\( \delta \)) 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.](image-url)
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 $\rho$ of closed frequent pattern, a growth ratio threshold $\delta$, and returns a new set of hotspots $D_h$, a set of GDPatterns $G$, and their footprints $\psi$. 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 ($\psi$);
- 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
\( \rho \): a support threshold of closed frequent pattern
\( \delta \): a growth ratio threshold

Result:
\( D_h \): a new set of hotspots
\( G \): a set of GDPatterns
\( \psi \): GDPattern footprints

1. \( \text{count} = 1 \);
2. Generate \( D_h, D_{h'}, \) and \( D_n \);
3. While \( \text{count} \neq 0 \) do
4. \( \text{count} = 0; \)
5. \( \mu = \emptyset; \)
6. \( G = \text{Mine GDPatterns using } D_h, \rho \) and \( \delta; \)
7. \( \psi = \text{footprints}(G); \)
8. For cell \( c \in D_{h'} \) do
9. If \( c \) adjacent to some cell in \( D_h \) and \( c \in D'_{h} \) then
10. \( \mu = \mu \cup c; \)
11. end
12. end
13. For cell \( c \in \mu \) do
14. If \( c \in \psi \) then
15. \( D_h = D_h \cup c; \)
16. \( \text{count}++; \)
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.

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

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 × 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 (19,873 m²) 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 ≥ 18, Fig.5a). We use the threshold of 9 RB incidents (18 > h’ ≥ 9), half of the initial value used for hotspots, to define the “potential hot” area (Dh’). The support threshold is set as 0.001. Also, growth ratios of GDPatterns are set as more than twenty (δ > 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 × 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.

<table>
<thead>
<tr>
<th>Hotspots Method</th>
<th>Total</th>
<th>Cells classified as Residential</th>
<th>Cells classified as non-residential</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>301</td>
<td>257(85.4%)</td>
<td>44(14.6%)</td>
</tr>
<tr>
<td>HOT</td>
<td>429</td>
<td>367(85.5%)</td>
<td>62(14.5%)</td>
</tr>
<tr>
<td>HSA</td>
<td>1094</td>
<td>901(82.4%)</td>
<td>192(17.6%)</td>
</tr>
</tbody>
</table>

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.](image)

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


Optimization of Criminal HotSpots Based on Underlying Crime Controlling Factors Using Geospatial Discriminative Pattern

Dawei Wang1, Wei Ding1, Tomasz Stepinski2, Josue Salazar3, Henry Lo1, and Melissa Morabito4

1. Department of Computer Science, University of Massachusetts Boston 
2. Department of Geography, University of Cincinnati 
3. Department of Computer Science, Rice University 
4. College of Liberal Arts, University of Massachusetts Boston

Abstract. Criminal activities are unevenly distributed over space. The concept of hotspots is widely used to analyze the spatial characters of crimes. But existing methods usually identify hotspots based on an arbitrary user-defined threshold with respect to the number of a target crime without considering underlying controlling factors. In this study we introduce a new data mining model – Hotspots Optimization Tool (HOT) – to identify and optimize crime hotspots. The key component of HOT, Geospatial Discriminative Patterns (GDPatterns), which capture the difference between two classes in spatial dataset, is used in crime hotspot analysis. Using a real world dataset of a northeastern city in the United States, we demonstrate that the HOT model is a useful tool in optimizing crime hotspots and it is also capable of visualizing criminal controlling factors which will help domain scientists further understanding the underlying reasons of criminal activities.

Keywords: Crime Hotspot, Hotspots Optimization Tool, Geospatial Discriminative Pattern, Footprint

1 Introduction

The use of crime hotspots—spatial locations of high crime concentration [3]—is a key component in the study of criminal related problems. The existence of hotspots is due to the nature that criminal activities are unevenly distribution over space. The reasons driving the distribution of crime incidents have been explained in relation to the interaction of target and offender and the strength of guardianship [5]. An accurately identified crime hotspot map will significantly benefit police practise such as threat visualization, police resources allocation, and crime prediction, etc. [4].

However, commonly used hotspots identification methods such as point mapping, thematic mapping, and kernel density estimation (KDE) rely on a user-defined threshold and none of them have taken the underlying controlling factors of crimes into account. There is a potential error when using user-specified
thresholds because the contrast between hotspots and normal areas may be ill-defined. For example, if a block with more than ten crime incidents a year is identified as a hotspot, then is there a large difference between this hotspot and the blocks that have nine crime incidents a year? A better way to accurately locate hotspots is to identify them not only by the criminal density, but also considering the underlying controlling factors.

In this paper, we introduce a new data mining model, *Hotspots Optimization Tool (HOT)* (Fig. 1), to improve the identification of hotspot by optimizing its boundary through the spatial footprints of patterns of crime driving factors. In the proposed method, a pattern means a combination of values of relevant variables. And patterns capable of identifying hotspots out of non-hot (normal) areas from the spatial perspective are called Geospatial Discriminative Patterns (GDPatterns) [7]. The HOT method adaptively optimizes the crime hotspots while searching for GDPatterns between crime hotspots and normal areas. Using a real world six-year dataset of a northeastern city in the United States, we demonstrate that the HOT model is a useful tool in optimizing crime hotspots, and it is also capable of visualizing criminal controlling factors which will help domain scientists further understanding the underlying reasons of criminal activities.

![Fig. 1. The framework of Hotspots Optimization Tool (HOT). The boundaries of hotspots are updated using GDPatterns according to the optimization rules.](image)

The rest of the paper is organized as follows. In Section 2 related works are discussed. Section 3 introduces the data representation and formal definition of the research problems. The *Hotspots Optimization Tool* is also presented in section 3. Our experimental results are discussed in Section 4. And in Section 5 we conclude the paper and discuss future research directions.

## 2 Related Work

Classic criminal theories, such as the Routine Activities Theory [5], 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”[10] and “disorder”[17] explain why adjacent areas of crime hotspots are at higher risk. A recent work done by [16] also discusses how an area is affected by the activity scope of offenders.

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 [9] 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 [12]. 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) [18] 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 [4]. All the above methods focus only on the target crime data and none of them consider underlying controlling factors of crime incidents.

Geospatial Discriminative Pattern applies emerging patterns to the spatial content. Emerging patterns are firstly introduced in [8] and further systematically studied in [14]. In the work of [7] 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 statistically significant difference between target crime hotspots and normal areas with respect to its underlying related factors. It is the first time that GDPatterns have been used in the field of crime hotspot study.

3 Methodology

In this section, we will formally define the research problem and then present the HOT algorithm. To find GDPatterns of a target crime and its associated variables, a transaction-based geospatial database needs to be built. A widely used method for representing spatial distribution of entities is grid thematic mapping [11]. In this work we firstly generate a grid mask to cover the studied area. Variable data (both target crime and explanatory variables that contain information about underlying controlling factors 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 explanatory variables 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 [6]. 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 group similar values, and the differences between categories are maximized.

Finally, with a user-specified threshold, the cells of the target crime grid can be classified into two classes: hotspot and normal area. A transaction-based geospatial dataset \( D \) is built.

**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 explanatory variables, and \( C \) is the class label of target crime. \( C = 0 \) if the area is not a hotspot (or normal area) and \( 1 \) if the area is a hotspot. Using \( C \), objects in \( D \) are labelled into the class of \( D_h \) (hotspots) if \( C = 1 \), or \( D_n \) (normal area) if \( C = 0 \).

**3.1 Geospatial Discriminative Patterns**

Here we give a brief introduction of *Closed Frequent Patterns* [15], GDPatterns and related definitions.

**Definition 2** Transaction and pattern: In a geospatial database, a transaction \( T \) is the group of explanatory variables \( \{ V_1, V_2, ..., V_n \} \) in an object. An pattern \( X \) is a set of values of explanatory variables (e.g. \( V_1 = 1 \), \( V_3 = 4 \)). For example, disregarding the class label \( C \), in dataset \( D \) each object can be viewed as a transaction in location \( (x, y) \) with a fixed-number of variables.

**Definition 3** Support: A pattern is said to be supported by a transaction when it is a subset of the transaction. For example, given a transaction \( T \{ \{ V_1 = 1, V_2 = 1, V_3 = 2, V_4 = 2, V_5 = 3, V_6 = 5 \} \) \}, patterns \( X_1 \{ V_1 = 1, V_2 = 1, V_5 = 3 \} \) and \( X_2 \{ V_1 = 1, V_3 = 2, V_4 = 2 \} \) are supported by \( T \), though \( X_3 \{ V_1 = 1, V_5 = 5, V_6 = 3 \} \) is not because it is not a subset of \( T \). The number of transactions that support an pattern \( X \) is called the support count \( \text{suppcount} \) of \( X \). The support of \( X \) is the ratio of \( X \)’s suppcount and the total number of transactions in a geospatial database (Formula 1).

\[
\text{sup}(X) = \frac{\text{suppcount}(X)}{\tau}
\]

where \( \text{sup}(X) \) is the support of pattern \( X \) and \( \tau \) is the number of transactions.

**Definition 4** Closed frequent patterns: An pattern \( X \) is said to be a closed pattern when none of its immediate super-sets has exactly the same support as \( X \). 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. Furthermore, \( X \) is a closed frequent pattern if the support of \( X \) is greater than a user-defined minimum support threshold \( (\rho) \). We are only interested in closed frequent patterns because infrequent patterns are likely to be insignificant and may happen by chance.

The patterns we are looking for should meet two requirements: (1) to significantly represent the situation or conditions of explanatory variables in objects
in $D$: (2) to significantly distinguish classes ($D_h, D_n$) from dataset $D$. A closed frequent pattern can satisfy the first requirement. To capture the difference of classes, the patterns should be more frequent in one class than in another.

**Definition 5 Geospatial Discriminating Patterns (GDPattern):** In a geospatial database, a closed frequent pattern $X$ is also a GDPattern if the growth ratio ($\delta$) of $X$ is larger than a user defined threshold. Here, growth ratio of a pattern is defined as the ratio of its supports in different classes.

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

where $\delta$ is the growth ratio; $\text{sup}(X, D_h)$ is the supports of closed frequent pattern $X$ in class $D_h$ and $\text{sup}(X, D_n)$ is supports of closed frequent pattern $X$ in class $D_n$.

**Definition 5 Footprint of a GDPattern:** The footprint of a GDPattern $X$ is the objects that support $X$ in geospatial dataset $D$ (Fig. 2). 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.

![Fig. 2. A example map of GDPatterns Footprints. By selecting Residential Burglary (RB) data as the target crime, nine other variables are used as explanatory 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.](image)

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 hotspots.
Algorithm 1: The Hotspot Optimization Tool takes as input a geospatial dataset \( D \), a hotspot threshold \( h \), a hotspot candidate threshold \( h' \), a closed frequent pattern threshold \( \rho \), a growth ratio threshold \( \delta \), and returns a new set of hotspots \( D_h \), a set of GDPatterns \( G \), and their footprints \( \psi \).

**Data:** \( D, h, h', \rho, \delta \)

**Result:** \( D_h, G, \psi \)

1. \( \text{count} = 1; \)
2. Generate \( D_h, D_{h'} \) and \( D_n; \)
3. while \( \text{count} \neq 0 \) do
   4. \( \text{count} = 0; \)
   5. \( \mu = \emptyset; \)
   6. \( G = \text{Mine GDPatterns using } D_h, \rho \) and \( \delta; \)
   7. \( \psi = \text{footprints}(G); \)
   8. for cell \( c \in D_{h'} \) do
      9. if \( c \) adjacent to some cell in \( D_h \) and \( c \in D_h' \) then
         10. \( \mu = \mu \cup c; \)
      end
   end
   11. for cell \( c \in \mu \) do
      12. if \( c \in \psi \) then
         13. \( D_h = D_h \cup c; \)
         14. \( \text{count}++; \)
      end
   end
end

3.2 Hotspot Optimization Tool

As mentioned above, locating hotspots with a user defined threshold 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”[10] or the source of “disorder”[17] 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 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 explanatory variables are the same.

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 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 ($\psi$);
- 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.

4 Experiment Results

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 $130.1 \text{ km}^2$ 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. In addition to RB, total of eight social/criminal features are selected in this study as explanatory variables with the help of a domain expert. 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 data (AR) are related criminal data that pictured the level of activity of crimes. The rates of CB, MV, and ST reflect the strength of guardianship in the area. Arrest rate 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 number of houses units (HU). A hotspot map of RB may simply be displaying locations of high housing density 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.

The original criminal dataset comes as vector maps (points and polygon). A grid 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 100 m × 100 m, 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 (19,873 m²) 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.

4.2 Hotspots Optimization

An initial threshold of RB hotspots is needed to set the initial classes before the HOT algorithm is used. From the study of [16], 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 \( h \geq 18 \).

Using a support threshold of 0.001, 6,327 patterns are mined out of which top 1,500 are selected with a growth ratio more than twenty \( \delta > 20 \), which indicate with an at least 95% confidence level (1:20) that these GDPatterns will reveal the difference between hot spots and normal area. 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 \( (D_{h'}) \). In the 6th loop OHS reaches the final condition and stops (Fig. 3). A final version of the set of patterns is extracted and the growth ratios of top 1,500 GDPatterns are all greater than 50, which is at least twice the initial version.

The new hotspot grid map is projected with satellite images of the studied city and a figure of sample site is extracted and shown in Fig. 4. Using an arbitrary 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, cells which are mostly covered by natural land, parking lots, roads and highways identified and are not classified into hotspots using our methods.
5 Conclusion and Future Work

In this paper we present a data mining model –Hotspots Optimization Tool – to optimize crime hotspots using GDPatterns. It is a first time attempt of using GDPatterns in crime hotspots analysis. Using a real world dataset we have proved that our model is capable of identifying crime hotspots by considering the controlling factors of criminal activities. This is important in criminal analysis because we can visualize areas that are in danger of becoming unstable and changing into a pool of criminal activity.

The GDPatterns mined in the process is an information-rich dataset and from which more details of crime driving factors can be extracted. The optimization process is not only a visualizing of crime itself but also an visualization
of controlling factors and will help our understanding of the underlying reasons of criminal activities. In our future work, we will focus on rational structured and re-organized GDPatterns.

6 Acknowledgement

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

References

This paper introduces a new data mining model to identify and optimize crime hotspots. This work could have a significant impact on analysis of criminal activities.
7: Three strong points of this paper (please number each point)

- The work is innovative since the concept of GDPattern is used for the first time in the study of crime hotspots.
- The algorithm is explained clearly.
- The proposed technique is validated through experimental results that clearly show the efficiency of the optimization method.

8: Three weak points of this paper (please number each point)

- There are some grammatical errors that should be corrected (see below for details)

9: Best paper candidate?

[X] Yes
[ ] No

10: Recommending this paper to expand for international journals?

[X] Applied Intelligence (SCI)
[ ] International Journal of Advancements of Computing Technology (EI)
[ ] No

11: Detailed comments for the authors

- On page 4, second line, part of the sentence should be rewritten. Suggestion:
  … cell in the grid is assigned as the count of incidents falling into the cell.
- On page 5, right below formula (1), part of the sentence should be rewritten. Suggestion: … growth ratio; sup(X, Dh) is the support of closed frequent pattern X in class Dh.
- On page 6, fifth line of the first paragraph, part of the sentence should be rewritten. Suggestion: … adjacent areas (cells) of a hotspot cell are more likely to fall into the active range of the same criminals.
- On page 6, the algorithm code line 6 should be better justified so that the variable Dh is not below the line number. If possible, align Dh so that it is positioned below the word “Generating”.
- On page 8, the caption of figure 3 could be rewritten: Growth Ratio graphs of GDPatterns: On the left, mined from original hotspots; on the right, from optimized hotspots using HOT.
- On page 9, the third sentence of the first paragraph should be rewritten. Suggestion: … have been left out. It is reasonable that houses located in the same block have a similar risk of being broken into.

----------------------- REVIEW 2 ---------------------

PAPER: 184
TITLE: Optimization of Criminal HotSpots Based on Underlying Crime Controlling Factors Using Geospatial Discriminative Pattern
AUTHORS: Dawei Wang, Wei Ding, Tomasz Stepinski, Josue Salazar and Melissa Morabito

OVERALL RATING: 3 (Accept as a regular paper)
REVIEWER'S CONFIDENCE: 4 (expert)

1: Is the paper relevant to IEA/AIE?
[ ] No
[x] Yes

2: How innovative is the paper?
[ ] 5 (Very innovative)
[x] 4 (Innovative)
[ ] 3 (Marginally)
[ ] 2 (Not very much)
[ ] 1 (Not)
[ ] 0 (Not at all)

3: How would you rate the technical quality of the paper?
[ ] 5 (Very high)
[ ] 4 (High)
[ ] 3 (Good)
[x] 2 (Needs improvement)
[ ] 1 (Low)
[ ] 0 (Very low)

4: How is the presentation?
[ ] 5 (Excellent)
[ ] 4 (Good)
[ ] 3 (Above average)
[x] 2 (Below average)
[ ] 1 (Fair)
[ ] 0 (Poor)

5: Is the paper of interest to IEA/AIE users and practitioners?
[x] 3 (Yes)
[ ] 2 (May be)
[ ] 1 (No)
[ ] 0 (Not applicable)

6: Summary of the paper's main contribution and impact

7: Three strong points of this paper (please number each point)

1- The paper is a novel application of data-mining techniques in offering a solution to a real world problem.

2- The idea of benefiting from factors which are catalysts of criminal activities in detecting hot-spots is interesting.

8: Three weak points of this paper (please number each point)

1- There are terms in the article which have been used very loosely. The discussions could have been made more rigorous if precise mathematical definitions would have been offered for data mining related terms.
2. There are many grammatically incorrect or unclear sentences in the article. Also there are instances of punctuation errors.

9: Best paper candidate?
[ ] Yes
[ ] No

10: Recommending this paper to expand for international journals?
[ ] Applied Intelligence (SCI)
[ ] International Journal of Advancements of Computing Technology (EI)
[ ] No

11: Detailed comments for the authors
The new contents of the extended manuscript:

1. A new compare study is done between our HOT method and exist hotspot mapping methods.
2. A new section for the discussion of the experimental results and the pros and cons of our method is included.
3. The Introduction section is refined and more materials are added in the Related Work section.
4. The figures in the paper are refined and three new figures and two new tables are included.