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

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

H. Jiang et al. (Eds.): IEA/AIE 2012, LNAI 7345, pp. 553–562, 2012.

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.

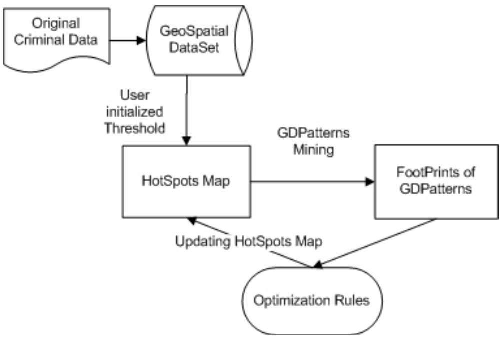

Fig. 1. The framework of *Hotspots Optimization Tool* (HOT). The boundaries of hotspots are updated using GDPatterns according to the optimization rules.

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 statically 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: hotspots and normal area and 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 is 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 (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).

$$sup(X) = \frac{suppcount(X)}{\tau} \tag{1}$$

where $sup(X)$ is the support of pattern X and τ 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 (ρ). 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(δ) 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.

$$\delta = \frac{sup(X, D_h)}{sup(X, D_n)} \tag{2}$$

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

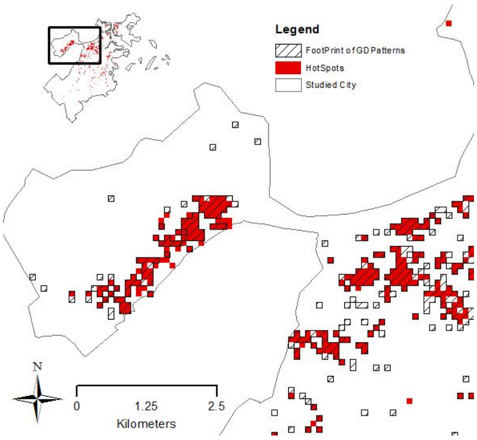

Fig. 2. A example map of GDPatterns Footprints. By selecting Residential Burglary(RB) data as the target crime, nine other variables are used as explanatory variables from the experiment dataset and 1,500 GDPatterns are mined with a growth ratio larger than twenty. The red area are RB hotspots with a user defined threshold and hallow squares with slash lines are footprints of the 1,500 GDPatterns.

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 ρ, a growth ratio threshold δ, and returns a new set of hotspots D_h, a set of GDPatterns G, and their footprints ψ.

```
Data: $D, h, h', \rho, \delta$
Result: $D_h, G, \psi$
$count = 1$;
Generate $D_h$, $D_{h'}$ and $D_n$;
while $count \neq 0$ do
    $count = 0$;
    $\mu = \emptyset$;
    $G$ = Mine GDPatterns using $D_h$, $\rho$ and $\delta$;
    $\psi = footprints(G)$;
    for cell $c \in D_{h'}$ do
        if $c$ adjacent to some cell in $D_h$ and $c \in D'_h$ then
            $\mu = \mu \cup c$;
        end
    end
    for cell $c \in \mu$ do
        if $c \in \psi$ then
            $D_h = D_h \cup c$;
            $count$++;
        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 (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.

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 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 $100m \times 100m$, 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.

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.

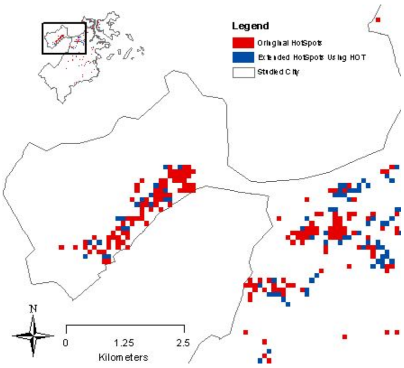

Fig. 3. Optimized hotspots map of the studied city. The purple cells are hotspots initially defined by the user-defined threshold and the blue cells represent hotspots that are added from candidate areas using HOT.

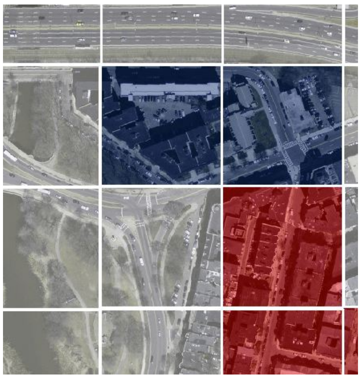

Fig. 4. A re-projection example of hotspots with satellite images. The purple cells are hotspots defined by the original threshold and the red cells are hotspots identified using our optimization method.

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.

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

References

1. Bailey, T.C., Gatrell, A.C.: Interactive spatial data analysis. Longman Scientific & Technical Essex (1995)
2. Bates, S.: Spatial and temporal analysis of crime. Research Bulletin (April 1987)
3. Chainey, S., Ratcliffe, J.: GIS and crime mapping, vol. 6. John Wiley & Sons Inc. (2005)
4. Chainey, S., Tompson, L., Uhlig, S.: The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal 21(1), 4–28 (2008)
5. Cohen, L.E., Felson, M.: Social change and crime rate trends: A routine activity approach. American Sociological Review, 588–608 (1979)
6. Cook, W., Ormerod, P., Cooper, E.: Scaling behaviour in the number of criminal acts committed by individuals. Journal of Statistical Mechanics: Theory and Experiment 2004, 07003 (2004)
7. Ding, W., Stepinski, T.F., Salazar, J.: Discovery of geospatial discriminating patterns from remote sensing datasets. In: Proceedings of SIAM (2009)
8. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the 5th ACM SIGKDD, pp. 43–52. ACM (1999)
9. Eck, J.E., Chainey, S., Cameron, J.G., Leitner, M., Wilson, R.E.: Mapping crime: Understanding hot spots (2005)
10. Gladwell, M.: The tipping point: How little things can make a big difference. Little, Brown and Company (2000)
11. Harries, K.D.: Mapping crime: Principle and practice. US Dept. of Justice, Office of Justice Programs, Crime Mapping Research Center (1999)
12. Hirschfield, A.: Mapping and Analysing Crime Data: Lessons from research and practice. CRC (2001)
13. Jenks, G.F.: The data model concept in statistical mapping. International Yearbook of Cartography 7, 186–190 (1967)
14. Li, J., Liu, G., Wong, L.: Mining statistically important equivalence classes and delta-discriminative emerging patterns. In: Proceedings of the 13th ACM SIGKDD, pp. 430–439. ACM (2007)
15. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering Frequent Closed Itemsets for Association Rules. In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 398–416. Springer, Heidelberg (1998)
16. Short, M.B., Bertozzi, A.L., Brantingham, P.J.: Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression. Journal on Applied Dynamical Systems 9, 462 (2010)
17. Skogan, W.G.: Disorder and decline: Crime and the spiral of decay in American neighborhoods. Univ. of California Pr. (1992)
18. Wand, M.P., Jones, M.C.: Kernel smoothing, vol. 60. Chapman & Hall/CRC (1995)