Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns

Dawei Wang[a], Wei Ding[a,*], Henry Lo[a], Melissa Morabito[b], Ping Chen[e], Josue Salazar[c], Tomasz Stepinski[d]

[a] Department of Computer Science, University of Massachusetts Boston, United States
[b] Department of Criminology and Criminal Justice, University of Massachusetts Lowell, United States
[c] Department of Computer Science, Rice University, United States
[d] Department of Geography, University of Cincinnati, United States
[e] Computer and Mathematical Sciences Department, University of Houston-Downtown, Texas, United States

ARTICLE INFO

Article history:
Received 23 April 2012
Received in revised form 25 January 2013
Accepted 31 January 2013

Keywords:
Crime related variable
Geospatial Discriminative Pattern
Hotspot Optimization Tool
Footprint

ABSTRACT

Crime tends to cluster geographically. This has led to the wide usage of hotspot analysis to identify and visualize crime. Accurately identified crime hotspots can greatly benefit the public by creating accurate threat visualizations, more efficiently allocating police resources, and predicting crime. Yet existing mapping methods usually identify hotspots without considering the underlying correlates of crime. In this study, we introduce a spatial data mining framework to study crime hotspots through their related variables. We use *Geospatial Discriminative Patterns* (GDPatterns) to capture the significant difference between two classes (hotspots and normal areas) in a geo-spatial dataset. Utilizing GDPatterns, we develop a novel model—*Hotspot Optimization Tool* (HOT)—to improve the identification of crime hotspots. Finally, based on a similarity measure, we group GDPattern clusters and visualize the distribution and characteristics of crime related variables. We evaluate our approach using a real world dataset collected from a northeast city in the United States.

1. Introduction

Crime is understood to be related to the interaction of victims and offenders, and to the strength of guardianship (Cornish & Clarke, 1986). In practice, these concepts can be measured using a variety of socio-economic and crime opportunity variables, such as population density, economic investment, and arrest rate.

Geographical studies reveal that crime is often concentrated in clusters, which in the literature are called hotspots. Hotspot mapping techniques for crimes draw continuous attention from researchers and public safety agencies. This is because accurately identified and clearly visualized crime hotspots, and understanding their relation to underlying crime related variables, can significantly benefit crime analysis and police practices by providing a solid basis for threat visualization, police resource allocation, and crime prediction.

Existing hotspot mapping methods can be essentially divided into three main categories: point mapping, choropleth mapping, and kernel density estimation (KDE) (Eck, Chainey, Cameron, Leitner, & Wilson, 2005; Williamson, McGuire, Ross, Mollenkopf, & Goldsmith, 2001; Boba, 2005). Usually, these methods aggregate the density of a target crime, which results in a net loss of information (Van Patten, McKeldin-Coner, & Cox, 2009). For example, in choropleth mapping, incident-level data is first aggregated into arbitrary administrative or political boundary areas. During this step, spatial details within and across the thematic areas are lost. Second, when hotspots are generated based on aggregated data, there is a further decline of precision in the resulting map. Because traditional methods mainly rely on target crime density, particular areas with relatively less crime may be left out of hotspots, even though crime related variables indicate they are under similar risks as those hotspots.

A reasonable way to reduce this accuracy and precision loss in choropleth mapping is to use more related information in the mapping process. Crime related variables can be aggregated and used along with target crime data in the hotspot identification process. Information carried by these variables can provide clues on whether the relatively high crime rate in a certain area happens by chance. Compared to traditional methods, the utilization of related information in hotspot mapping can reduce information loss during analysis.

Additionally, such an approach can benefit further analysis on the characteristics of crime related variables. Instead of just evaluating crime by itself, recent studies also integrate crime related data into a unified framework that assists the analysis

* Corresponding author.
E-mail addresses: ding@cs.umb.edu (W. Ding), chenp@uhd.edu (P. Chen).

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.01.008

and exploration of crime hotspots (Maciejewski et al., 2010). Using related variables in hotspot mapping can additionally benefit such visualization and analyzation processes by providing an intuitive linkage between target crime and its related data.

In this paper, we present a framework that uses spatial data mining concepts to map hotspots and investigate the relationship between socio-economic and criminal variables. Recently, spatial data mining has emerged as an active research area in studies of spatial relationships that try to answer the questions like "why" and "where" (Ester, Kriegel, & Sander, 1997; Mu, Ding, Morabito, & Tao, 2011). It has been proven to be very powerful in identifying the linkage between target objects and its related factors. The components of our method are shown in Fig. 1. In particular, we:

- Introduce a spatial data mining concept, *Geospatial Discriminative Patterns* (GDPatterns), to study the relationship between target crime hotspots and their underlying related variables.
- Introduce a model, *Hotspot Optimization Tool* (HOT), to identify crime hotspots through their related variables.
- Use a similarity based method to cluster the crime related variables that contribute to hotspots into groups.
- Visualize the locations of those clusters in a rational way to assist domain scientists in further analysis, using the footprints of GDPatterns.

Utilizing the proposed framework, a case study is conducted using a 6-year crime dataset from a city in northeast United States. We compare our mapping tool with a widely used hotspot evaluating technique,the G_i^* statistics (Getis & Ord, 2010), and demonstrate the potential in assisting crime analysis using related variable clusters.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 introduces the data representation and formal definition of the research problems. HOT model and the implementation of the similarity measure are also presented in this section. Section 4 evaluates the proposed framework in a real-world case study. We conclude the paper in Section 5.

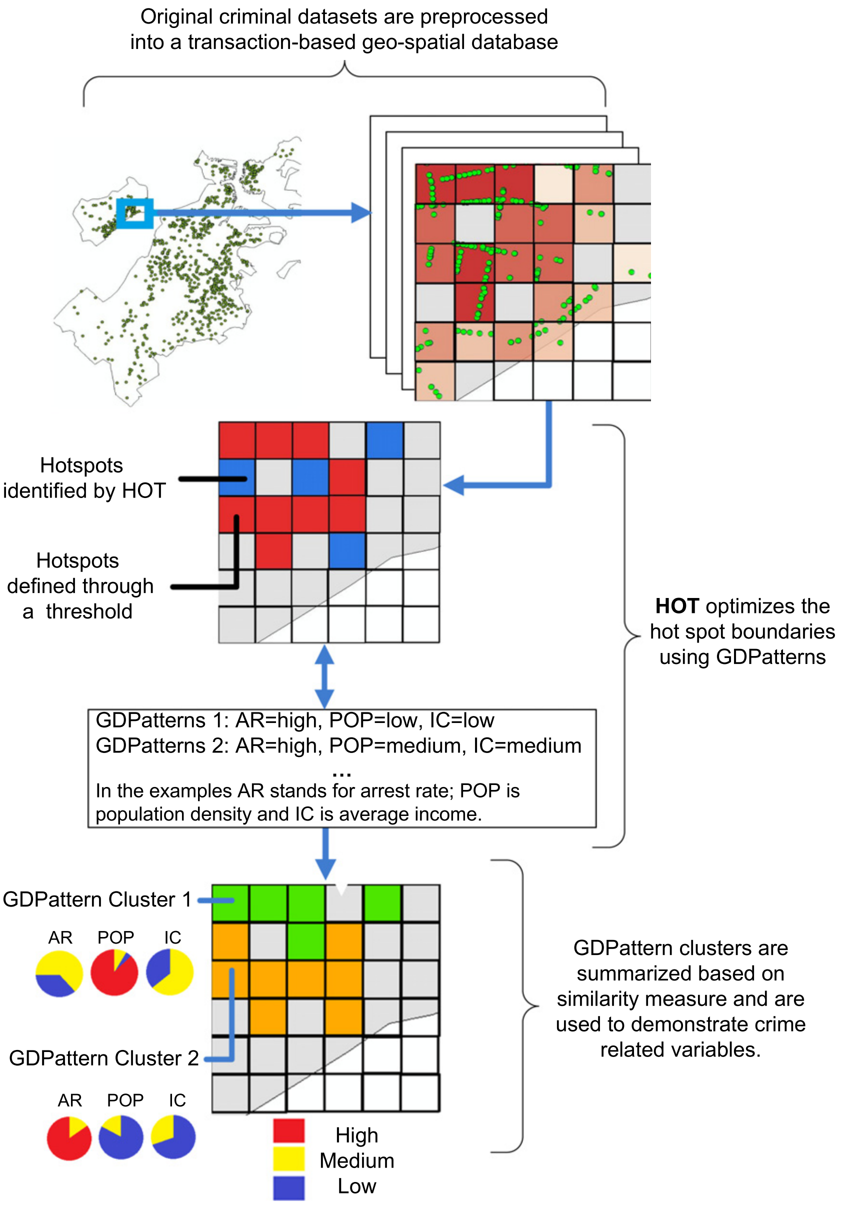

Original criminal datasets are preprocessed
into a transaction-based geo-spatial database
Hotspots
identified by HOT
Hotspots
defined through
a threshold
HOT optimizes the
hot spot boundaries
using GDPatterns
GDPatterns 1: AR=high, POP=low, IC=low
GDPatterns 2: AR=high, POP=medium, IC=medium
...
In the examples AR stands for arrest rate; POP is
population density and IC is average income.
GDPattern Cluster 1
AR
POP
IC
GDPattern Cluster 2
AR
POP
IC
GDPattern clusters are
summarized based on
similarity measure and are
used to demonstrate crime
related variables.
High
Medium
Low

Fig. 1. The framework of our methods. With the help of GDPatterns, criminal hotspot maps are generated using HOT. By applying a similarity measure method, GDPatterns are clustered and visualized for domain scientists.

2. Related work

In this section we briefly present some literatures related to criminology, spatial data mining, and hotspot mapping techniques. Additionally, we give a brief introduction to a choropleth mapping application—the Hotspot Analysis (HSA) tool implemented by Esri ArcGIS (ESRI, 2011).

Occurrence of crime has been linked to a number of different variables. Classic criminology theories, such as Routine Activities Theory (Cohen & Felson, 1979), conclude that three concepts contribute to crime: accessible and attractive targets, a pool of motivated offenders, and lack of guardianship (Brantingham & Brantingham, 1984; Cornish & Clarke, 1986). The concept of "disorder" (Skogan, 1992) explains why adjacent areas of crime hotspots are at higher risk. The probability of arrest or the social penalties for committing crime may be lower in crime hotspots than in other neighborhoods, which leads to the "contagion" of criminal activity in crime hotspots (Ludwig, Duncan, & Hirschfield, 2001; Sah, 1991; Sampson, Raudenbush, & Earls, 1997). Recent work done by Short, Bertozzi, and Brantingham (2010) also discusses how an area is affected by the activity scope of offenders. Criminology theories explain why crime is clustered in particular areas, and why certain victims are selected. They also help in deciding which variables are related to a certain type of crime.

Spatial data mining (Ester et al., 1997) is a knowledge discovery technique for "extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases" (Koperski & Han, 1995). It has been proven to be very powerful and efficient for studying comprehensive relationships in large databases (Miller & Han, 2009; Ester et al., 1997; Qian, He, Chiew, & He, 2012). The GDPattern is an application of integrating spatial association rules (Agrawal et al., 1994; Koperski & Han, 1995) with emerging patterns (Dong & Li, 1999; Herrera, Carmona, González, & del Jesus, 2011; Yu, Ding, Simovici, & Wu, 2012). Applications using association rules have been developed to explore the spatial and temporal relationships among objects using census data (Malerba, Esposito, Lisi, & Appice, 2002). In the work of Mennis (2006) and Mennis and Liu (2005), association rule mining techniques have been used to explore the non-linear relationships among socioeconomic-vegetation variables. In the work of Lin (1998) the authors present a similarity measure method for summarizing large number of emerging patterns. Ding, Stepinski, and Salazar (2009) adopts the relative risk ratio as the measure of pattern emergence and uses spatial data mining techniques in investigating vegetation remote sensing datasets. In our work GDPatterns are used as a tool to discover the statically significant difference between target crime hotspots and normal areas spatially, with respect to the underlying related variables.

The Spatial and Temporal Analysis of Crime (STAC) program (Bates, 1987) is one of the earliest and widely used hotspot mapping applications. Based on point mapping, STAC uses "standard deviational ellipses" to display crime hotspots on a map and does not pre-define any spatial boundaries. But some studies (Eck et al., 2005) show that STAC may be misleading because hotspots do not naturally follow the shape of ellipses. Another popular hotspot representation method is choropleth mapping, in which boundary areas (geographic boundaries like census blocks or uniform grids) are used as the basic mapping elements (Hirschfield, 2001). Unlike point mapping, choropleth mapping uses aggregate data, which removes spatial details within the thematic areas. Also, identified hotspots are restricted to the shape of these areas. The method of Kernel Density Estimation (KDE) (Wand & Jones, 1995) 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. Reviews and comparative studies for the three methods have been done in the works of Chainey, Tompson, and Uhlig (2008), in which authors introduce a "prediction accuracy index" to evaluate the accuracy of the different methods in the content of predicting where crime may occur.

Esri ArcGIS (ESRI, 2011) is the most widely used Geographic Information System (GIS) and its newest component, ArcMap 10.1, includes a Hotspot Analysis (HSA) toolbox, which implements the G_i^* statistics (Getis & Ord, 2010) and provides users the ability to analyze the hotspots existed in the input spatial dataset (usually a polygon map with interested attributes). In particular, HSA calculate the G_i^* statistics and outputs z-scores and p-values for each spatial area (polygon) that tell the statistically significance of the polygon as a hotspot. 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 neighbors 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.

3. Methodology

The key insight behind our methods is identifying hotspots by searching, utilizing, and presenting patterns in geographic space. By preprocessing the crime related data sets into a transaction-based geospatial dataset, we develop a model, called HOT, to map crime hotspots through the related variables. Then we introduce a similarity method to summarize the identified GDPatterns into clusters. Based on these clusters, a relevant report of crime hotspots and related variables is visually presented for domain experts.

3.1. Problem formulation and data representation

To discover GDPatterns from a target crime's related variables, we firstly build a transaction-based geospatial database, which we refer to as the database or simply D. A widely used method for representing spatial distribution of entities in D is through grid mapping (Harries, 1999; Janeja & Palanisamy, 2012). Both target crime and related variables in the original spatial dataset can be plotted onto grid maps with the same dimensions. The cell value in the grid is assigned to be the number of incidents falling into it. An illustrative example of D is shown on the top right of Fig. 1. Additionally, instead of using the original values directly, the way to fairly represent all the variables in one pattern is to categorize them and change the original values into categories. Standard tools (Nguyen & Nguyen, 1998) such as the *Jenks Optimization for Natural Breaks Classification* (or Nature Breaks; Jenks, 1967), a method that is based on natural groupings inherited in data, can be used in the categorization process.

Definition 1 (*Database object*). A object in D is a tuple of the form: $\{x, y, V_1, V_2, \ldots, V_n, C\}$, where x, y indicate the object's spatial coordinates, $V_1, V_2, \ldots, V_n$ are the values of the related variables, and C is the class label of target crime.

Using C, objects in D can be labeled into 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 $C = 0$. 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. For example, in Table 1, T_1, T_2, T_3, and T_4 are transactions with three variable values.

3.2. Geospatial Discriminative Patterns (GDPatterns)

The GDPatterns we are looking for should meet two requirements: (1) to significantly represent the situation or conditions of related variables in objects in database D; (2) to significantly distinguish hotspots D_h from normal areas D_n. GDPatterns are built upon closed frequent patterns. Here we give a brief introduction of relevant concepts.

Definition 2 (*Pattern*). Given a set of related variables, a pattern is a set of values for a subset of those related variables.

For example, Table 1 gives an example of a database that has 3 related variables AR, POP, and IC, which can take the values of low, medium, or high. In the examples AR, POP and IC stand for arrest rate, population density and average income, respectively. A combination of these variables and values constitutes a pattern; e.g., X_1: $\{AR = high, POP = low\}$, or X_3: $\{AR = high\}$.

Definition 3 (*Support and support count (Agrawal et al., 1994)*). A pattern is said to be supported by a transaction when it is a sub set of the transaction. The support count of a pattern X is the number of times X appears in a database D.

$$supportcount_D(X) = |\{T \in D | X \subseteq T\}| \tag{1}$$

where T represents transactions in D.

The support of a pattern X is calculated as the support count of X divided by the total number of transactions in the database D.

$$support_D(X) = \frac{supportcount_D(X)}{|D|} \tag{2}$$

For example, in Table 1 pattern $X_1 = \{AR = high, POP = low\}$ is supported by transactions T_1,T_2 and T_3, then the support count of X_1 is 3 for the database. Since there are totally 4 transactions in this database, the support of X_1 is 3/4 = 0.75.

Definition 4 (*Closed pattern (Pasquier, Bastide, Taouil, & Lakhal, 1999)*). A pattern is closed if none of its supersets has exactly the same support.

For example, in Table 1 X_1 is a closed pattern and X_3 is not, because its immediate superset X_1 has exactly the same support.

Note that if we consider only closed frequent patterns, we can deduce the support of non-closed frequent patterns from their correspondent closed patterns. To see why this is true, note that the supports of patterns exhibit a property called downward closure:

$$\text{If} \quad X \subset X', \quad \text{then} \quad support_D(X) \geqslant support_D(X')$$

Thus, if X is closed, and X' is not, then $support_D(X) = support_D(X')$.

The benefit of considering only closed patterns is a reduction in the set of considered patterns without losing information. In Table 1 both X_3 and X_1 are supported by T_1, T_2 and T_3. In other words, both X_3 and X_1 carry information about the characteristics of these transactions. But X_1 carries more information ($\{AR = high, POP = low\}$) than X_3 ($\{AR = high\}$) does and the information carried by X_3 ($\{AR = high\}$) is fully represented by X_1. There is no information loss if we only consider X_1 in further analysis.

Definition 5 (*Closed frequent pattern (Pasquier et al., 1999)*). A closed pattern whose support is above a user-defined threshold is considered as a closed frequent pattern.

Definition 6 (*Growth ratio*). Let set $\{D_h, D_n\}$ be an exhaustive partition of D. The growth ratio δ of a pattern X is the ratio of X's support in one partition D_h to its support in the other partition D_n.

Table 1
Examples of transactions, patterns and patterns' supports. In the examples AR, POP and IC stand for arrest rate, population density and average income, respectively. 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 ρ = 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: $\{AR = high\}$	$sup(X_3) = \frac{3}{4} = 75\%(T_1, T_2, T_3)$

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

Definition 7 (*Geospatial Discriminative Pattern (GDPattern)*). A closed frequent pattern X whose growth ratio exceeds a user-defined threshold is considered a GDPattern.

With a rational growth ratio threshold, the GDPatterns mined from D carries information that is significantly different between a subset and the remainder in D. For example, if the growth ratio is greater than 20, thus a closed frequent pattern will be considered as a GDPattern when the pattern is 20 times more frequent in hotspots than in normal areas. In other words, this pattern will have a more than 95% (19/20) chance of being found in hotspots. So the locations out of which such a pattern is mined are more than 95% (or "significantly") likely to be a hotspot.

Definition 8 (*Footprint*). The footprint of a GDPattern X is the objects that support X in database D. It is the set of cells in the grid map whose corresponding objects support X.

For example, in Fig. 2 a GDPattern: {Commercial Burglary-"low",Street Robbery-"Average",Motor-Vehicle Larceny-"Average"} is selected from the case study (Section 4) and the hollow squares with slash lines are footprints of this GDPattern. These areas (the footprints) have similar characteristics of the related variables (low in commercial burglary rate and average in street robbery and motor-vehicle larceny rate). The utilizing of footprint provides a way to measure the spatial distribution of the corresponding patterns in studied area.

3.3. Hotspot Optimization Tool

GDPatterns are capable of digging out the meaningful information underlying the spatial distribution of target crime hotspots. Utilizing the informative GDPatterns, here we develop 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 the source of disorder 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 (Bailey & Gatrell, 1995), adjacent areas of a hotspot are more likely to fall into the active range of the same criminals. Therefore these cells can be considered as 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 areas. Normal areas located in the footprints of GDPatterns are more likely to be hotspots because in these areas the values of related variables are the same.

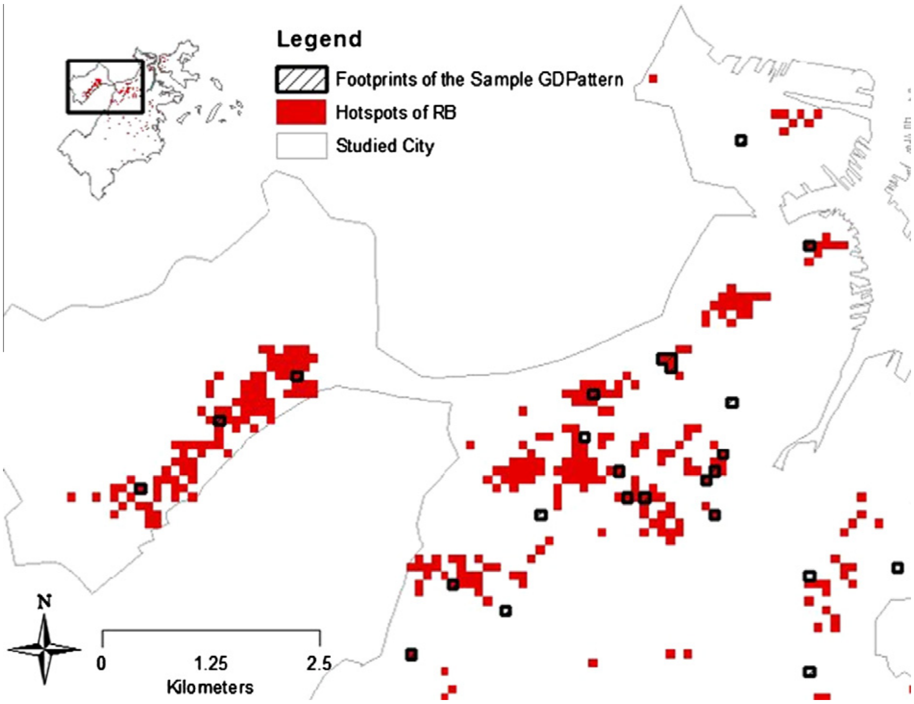


Fig. 2. A example map of GDPattern footprints. By selecting Residential Burglary (RB) data as the target crime, nine other variables are used as related variables from the experiment dataset and GDPatterns are mined with a growth ratio larger than twenty ($\delta \geqslant 20$). The hollow squares with slash lines are footprints of one example GDPattern (Commercial Burglary-"low", Street Robbery-"Average", Motor-Vehicle Larceny-"Average") whose grawth ratio is 67.0. The red area are RB hotspots defined by a user-specific threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In summary, by initializing hotspots of a target crime with a user-specified threshold, HOT considers a normal location as a hotspot if (1) it is adjacent to current hotspots; (2) its crime rate is relatively high compared to the user-specified hotspot threshold; and (3) it is inside the footprints of GDPatterns mined out of current hotspots. The detailed process of HOT is showed in Algorithm 1.

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  $count = 1$;
2  Generate $D_h$, $D_{h'}$ and $D_n$;
3  while $count \neq 0$ do
4      $count = 0$;
5      $\mu = \emptyset$;
6      $G =$ Mine GDPatterns using $D_h$, $\rho$ and $\delta$;
7      $\psi = footprints(G)$;
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             $count{+}{+}$;
17         end
18     end
19 end
```

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 whose corresponding objects belong to $D_{h'}$ and adjacent to some cell whose corresponding objects belong to 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 (moving its corresponding object from $D_{h'}$ to D_h) (line 15).
- Iterate until all hypothesis tests are false (lines 3 and 19).

When hotspot boundaries 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.

3.4. Crime related variables demonstration

Hotspots of target crime extracted using GDPatterns carry a wealth of information. But the GDPattern mining process usually results in an explosive number of possible patterns (Han et al., 2000). It is desirable to organize these GDPatterns in a meaningful way in order to make the information usable to domain analysts. Here we present a pattern summarization method that can cluster GDPatterns into small groups which have similar structures.

Given two patterns X and Y that are mined out of m variables, the function to calculate similarity between X and Y is

$$s'(X,Y) = \frac{\sum_{i=1}^{m} s(X_i, Y_i)}{m} \tag{4}$$

where $s'(X,Y)$ is the similarity between pattern X and Y; $s(X_i,Y_i)$ is the similarity between the i_{th} variables of X and Y; m is the number of variables in each pattern. For example, $s(X_i,Y_i) = 1$ if X_i and Y_i are in the same category and 0 if they are not. We calculate the similarities between every variable and take the mean of the m similarities as the overall similarity between the patterns.

The categories of the crime related variables can be presented using ordinal numbers. For example, the categories of population density can be presented using ordinal numbers: 1 ("low"), 2 ("medium") and 3 ("high"). The similarity between two ordinal values of the ith variable $s(X_i,Y_i)$ can be measured by the ratio between the amount of information needed to state the commonality between X_i and Y_i, and the information needed to fully describe both X_i and Y_i. In practice when we calculate the similarity between patterns X and Y, the ith variable does not always exist in both patterns (Fig. 3). There are three cases according to the presence of X_i and Y_i.

Case 1: Both X_i and Y_i are in the pattern:

$$s(X_i, Y_i) = \frac{2 \times \log P(X_i \vee Z_1 \vee Z_2 \cdots \vee Z_k \vee Y_i)}{\log P(X_i) + \log P(Y_i)} \tag{5}$$

where $P()$ is the probability calculated using the known distribution of the values of i_{th} variable in D and $Z_1,Z_2,\ldots,Z_k$ is the ordinal intervals delimited by X_i and Y_i. For example, in Fig. 3 the ordinal interval between the first variable $X_A$and Y_A is $Z_1 = 2$.

Case 2: Either X_i or Y_i is absent (here we use the case that X_i is absent):

$$s(-, Y_i) = \sum_{k=1}^{n} P_X(Z_k) s(Z_k, Y_i) \tag{6}$$

where n is the amount of different values that the i_{th} variable has, $P_X(Z_k)$ is the probability of the ith variable having value Z_k in all transactions that support pattern X. $P_X(Z_k) = 0$ if Z_k does not exist in the footprint of X at all and $\sum_{k=1}^{n} P_X(Z_k) = 1$. The similarity is a weighted average between Y_i and all ordinal values of the i_{th} variable presented in the footprint of pattern X. Example is shown in Fig. 3 case 2.

Case 3: Neither X_i or Y_i is present:

$$s(-,-) = \sum_{l=1}^{n} \sum_{k=1}^{n} P_X(Z_l) P_Y s(Z_k) s(Z_l, Z_k) \tag{7}$$

In this case the probability of all ordinal values $(Z_1,Z_2,\ldots,Z_n)$ of the ith variable in patterns X and Y are checked and a weighted average pairwise comparisons is calculated (case 3 in Fig. 3).

Using the similarity measurements, we can build a $N \times N$ distance matrix of GDPatterns using $distance = \frac{1}{similarity}$, where N is the number of GDPatterns. Standard clustering tools such as *Hierarchical Agglomerative Clustering* (HAC), which treat each GDPattern as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters according to their distance until all clusters have been merged into a single cluster that contains all GDPatterns, can be used to group the closest GDPatterns into clusters.

These clusters serve as compositions of crime related variables and carry rich information not only about relationships between variables, but also about their spatial distributions. Locations exhibiting certain socio-economic and crime-related characteristics tShat are significantly related with target crime hotspots can be drawn using the clusters' footprints. In Section 4 we present a case study to show how these GDPattern clusters can assist domain experts in criminal studies.

4. Case study

Utilizing the proposed framework, a case study is conducted with real world data from a northeastern city in the United States. We firstly describe the data preprocessing in Section 4.1. Secondly, with the purpose of comparison study, crime hotspot maps are drawn in Section 4.2 using HOT, HSA, and user-specified thresholds, respectively. Kappa Index (Cohen et al., 1960; Rossiter, 2004) and cell statistics are used to compare the results and the pros and cons of HOT are discussed. Finally, we cluster the GDPatterns using the similarity method (Section 3.4) and discuss the potentials of utilizing GDPattern clusters in demonstrating the characteristics of crime related variables in Section 4.3.

4.1. Data preprocessing

The data in the case study includes reported crimes and associated variables in a northeastern city in the United States from 2004 to 2009. 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's police department report), residential burglary (RB) is selected as the target crime (Fig. 4). In addition to RB, total of eight social/criminal features (Table 2) are selected in this study as related variables with the help of a domain expert. Among those are:

- *Commercial burglary* (CB), street robbery (SR), and motor vehicle larceny (MV). These indicate the level of activity of related crimes, and also reflect the strength of guardianship in the area.

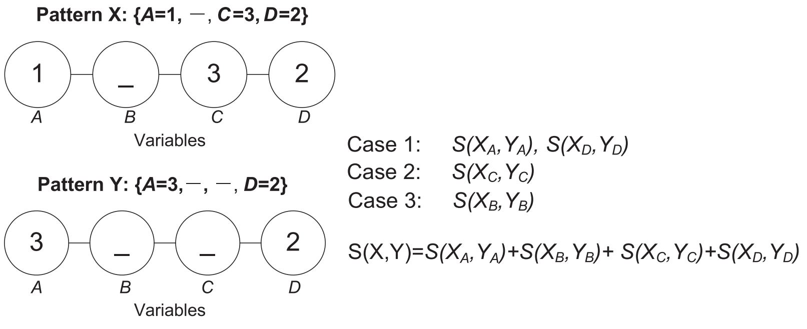

Pattern X: {A=1, −, C=3, D=2}
1
_
3
2
A
B
C
D
Variables
Case 1: S(X_A,Y_A), S(X_D,Y_D)
Case 2: S(X_C,Y_C)
Case 3: S(X_B,Y_B)
Pattern Y: {A=3,−, −, D=2}
3
_
_
2
A
B
C
D
Variables
S(X,Y)=S(X_A,Y_A)+S(X_B,Y_B)+ S(X_C,Y_C)+S(X_D,Y_D)

Fig. 3. An illustrative example showing the similarity measure approach between patterns X and Y.

- *Arrests* (AR). This helps indicate the size of the pool of offenders.
- *Foreclosed homes* (FC). A vacant house has a higher risk of being broken into than an inhabited one, and is also a sign of lack of guardianship.
- *Population* (POP) *and housing density* (HU). A hotspot of RB may simply be a location of high housing density because such areas have a potential higher RB rate than areas with fewer houses.
- *Distance to colleges* (DC). The studied city is heavily populated by college students, which makes many properties easy targets for burglars during semester breaks. DC is calculated as the distance to the geographical center of a university or college.

Table 2
Crime related variables for the case study.

Variables	Number of incidents (2005–2009)
Residential burglary (RB)	18,321
Street robbery (SR)	12,020
Commercial burglary (CB)	4438
Motor-vehicle larceny (MV)	29,685
Arrest (AR)	254,309
Foreclosed houses (FC)	11,671
Population (POP)	–
Number of houses units (HUs)	–
Distance to colleges (DCs)	–

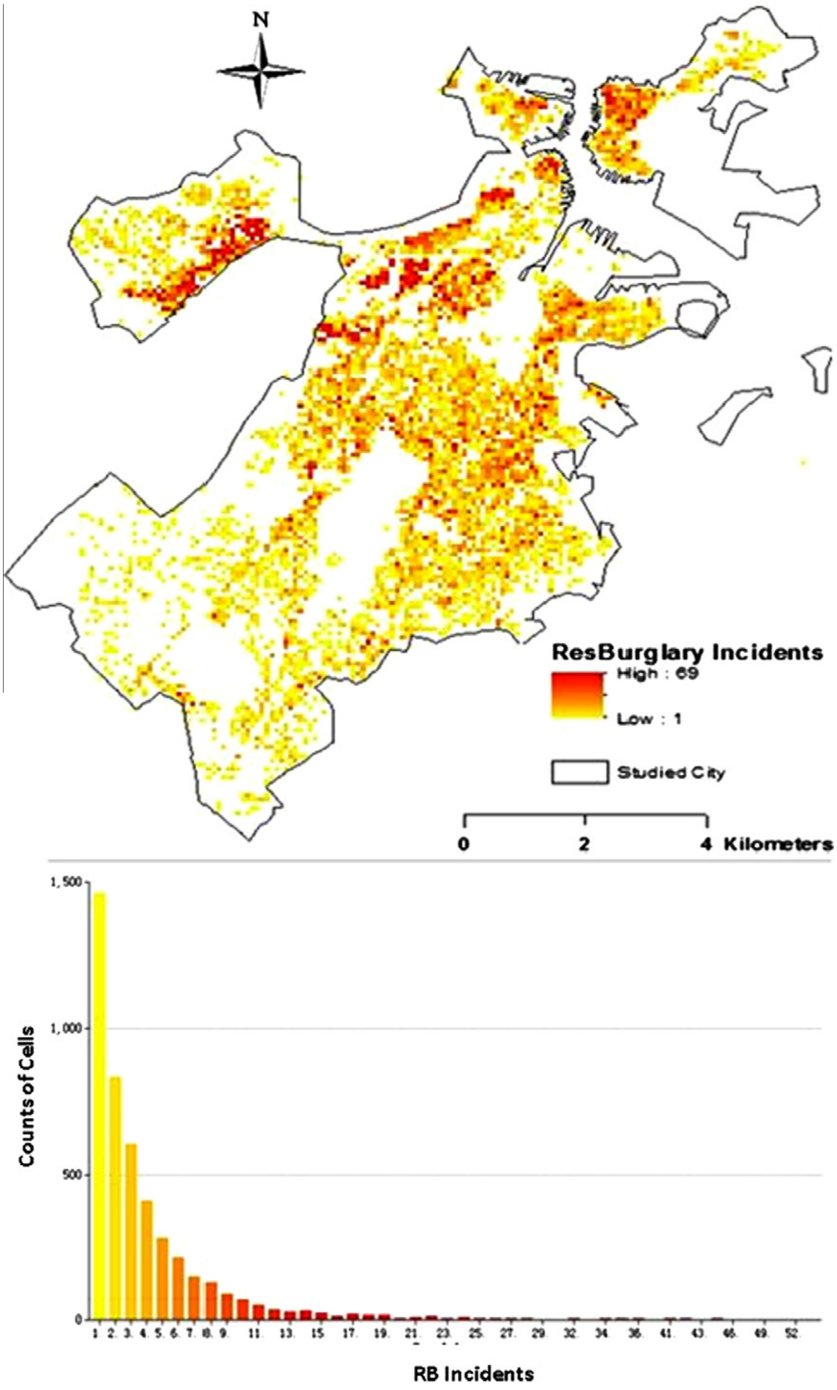


Fig. 4. Residential burglary rates in the studied city. Top is the grid density map of RB. On the bottom it is a graph showing the frequency of cell values.

The original criminal dataset comes as vector maps (points and polygons). We firstly convert all the variable data into grid maps (Fig. 4). The grid 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 the cell size. Firstly, the cell size (10,000 m^2) is approximately half the size of average city block size (19,873 m^2) in the studied city. According to domain experts, this will be a good representative of reality and helpful in police practice. Secondly, at this size, the number of cells covering the study area is the same order of magnitude as the number of RB incidents (Table 2), which minimizes the loss of spatial information during aggregation. On the other hand, HSA needs to be conducted using polygon maps instead of rasters. 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 in the area. In order to facilitate the discussion, we call the polygons in the fishnet map cells as well.

Since the related 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 (Cook, Ormerod, & Cooper, 2004) (Fig. 4). Using *Nature Breaks* (Jenks, 1967), every variable is divided into six categories: 0 – "empty", 1 – "lowest", 2 – "low", 3 – "average", 4 – "high", and 5 – "highest". Using the Nature Break method the categories' breaks are identified with best grouping of similar values, and the differences between categories are maximized.

4.2. Hotspot mapping

An initial threshold of RB hotspots is needed to set the initial classes before utilizing HOT. From the study of (Short et al., 2010), a house is at relatively higher risk if a burglary happened nearby within the past 4 months. Therefore if three or more burglaries happened in the block in one year, the area is likely to be a burglary hotspot. Because the time span of our data is 6 years, we set an area (cell) to be a hotspot if there are eighteen or more burglary incidents ($h \geqslant 18$). We use the threshold of 9 RB incidents ($18 > h' \geqslant 9$), to define the "potential hot" areas ($D_{h'}$). The growth ratio for GDPatterns is set at more than twenty ($\delta > 20$), which insures an at least 95% confidence level (1:20) that these GDPatterns will reveal the difference between hotspots and normal areas. To test the tolerance of HOT, four different support thresholds ($\rho = 0.001, 0.005, 0.01, 0.02$) are used in the experiments.

For comparison, hotspot maps generated by hard thresholds and the HSA method are presented. Three maps using the hard thresholds are generated. Two of them are just using the thresholds of $h \geqslant 18$ and $h \geqslant 9$. The third one is generated using an initial threshold of $h \geqslant 18$ and then locating cells with RB rate $h \geqslant 9$ that are also adjacent to the $h \geqslant 18$ cells.

HSA takes the fishnet map (Section 4.1) as input and calculates a G_i^* (Formula (8)) statistic for each polygon in the map. The G_i^* statistic is considered as the z-score of the polygon. Then a *p*-value, the probability distribution of the *z*-scores, is calculated for each polygon. In summary, a polygon with a high *z*-score and a *p*-value less or equal to 0.05 is considered as having a high enough attribute value to be statistically significant, and thus is considered as a hotspot.

$$\begin{aligned} \bar{X} &= \frac{\sum_{j=1}^{n} x_j}{n} \\ S &= \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \\ G_i^* &= \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \end{aligned} \tag{8}$$

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* (In the case study we use inverse distances as the spatial weights (Deane, Beck, & Tolnay, 1998; Ratcliffe & Taniguchi, 2008; Tita & Greenbaum, 2009) and Euclidean Distance as the distance method.), *n* is the total number of polygons.

We name the maps generated using hard thresholds $h \geqslant 18$ and $h \geqslant 9$ HT18 and HT9, respectively. The map generated using $h \geqslant 18$ cells and their adjacent cells with $h \geqslant 9$ is called HT18_9. The HOT produced maps using the support thresholds $\rho = 0.001$, $\rho = 0.005$, $\rho = 0.01$, $\rho = 0.02$ are called HOT001, HOT005, HOT01, and HOT02, respectively. The map generated by HSA is named the HSA map. All these maps are shown in Fig. 5.

The standard Kappa Index *k* (Formula (9)) (Cohen et al., 1960; Rossiter, 2004) is used to compare the difference between hotspot maps (Table 3). The value of *k* is between −1 and 1, and two maps are considered more similar when the *k* between them is larger (closer to 1).

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{9}$$

where p_0 is the proportion of cells that classified into the same class (agreed) by both maps. p_c is the proportion of units for which the agreement is expected by chance.

From Fig. 5 and Table 3 we can tell that even using different support thresholds, the final HOT hotspot maps are very close to each other (the Kappa indices between them are all larger than 0.94). Although different support thresholds will result in different set of closed frequent patterns, by setting a relatively high growth ratio value, only the most significant patterns are selected as GDPatterns that contribute to hotspot mapping.

The HOT maps and the HT18_9 map are similar to each other (average Kappa Index 0.86) because they all contain the $h \geqslant 18$ cells. On the other hand, there are totally 344 (different hotspot cells between HT18 and HT18_9, Table 4) cells that having RB rate $h \geqslant 9$ and adjacent to the $h \geqslant 18$ cells and around 69.4% of them are considered as hotspots by HOT (calculated by dividing the average value of different hotspot cells between the HOT maps and the HT18 map by 344, Table 4). The difference between them (HOT maps and the HT18_9 map) can be considered as the information gained using HOT.

A land cover map of the studied city is drawn (Fig. 6) with the purpose of evaluating the precision of our hotspot maps. In Table 4 we calculated the cell statistics for each map. The percentages of RB hotspot cells that are actually located in residential areas can be seen as the precisions of the maps (Column 3, Table 4).

All the hotspot maps we generated are based on grid choropleth mapping. There is an intrinsic defect when using grid choropleth mapping for hotspot identification. By converting points representing crime incidents into cells with crime counts, spatial details within and across the cells boundaries can be lost. In the case study, this limitation is reflected by the fact that cells in non-residential areas (Fig. 6) are classified as hotspots of residential burglary (RB) in all the hotspot maps. For example, after the aggregation process a certain cell may contain 20% non-residential areas, like roads or parks, and 80% of residential areas. If during the hotspot analysis process the cell is classified as a residential burglary (RB) hotspot, then the precision of this hotspot is 80%.

The hotspot maps using the user-specified thresholds (HT18, HT9 and HT18_9) can be considered as benchmarks for the case study. In other words, using the current grid map (cell size 100 m × 100 m), the precision for describing residential areas in the studied city is around 85% (percentage of hotspot cells locating in residential area in the hard threshold hotspot maps; Table 4). HSA does not achieve this precision. Because during the hotspot analysis (the G_i^* statistic calculation) process, all the cells are only

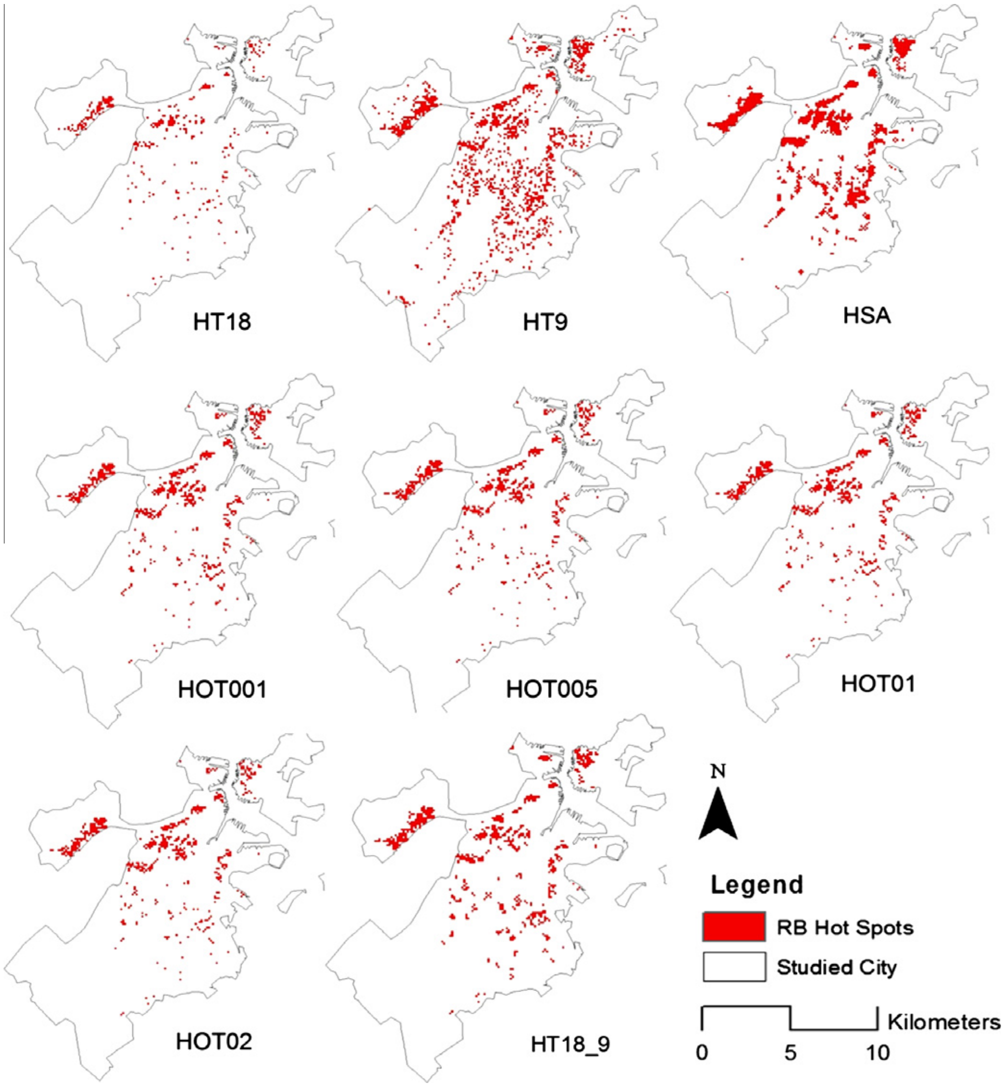


Fig. 5. RB hotspot maps of the studied city. HT18 and HT9 are generated by the thresholds of $h \geqslant 18$ and $h \geqslant 9$, respectively. HSA is the hotspot map generated by the Hotspot Analysis tool in Esri ArcGIS. HOT001, HOT005, HOT01, HOT02 are the HOT generated hotspot maps with the support thresholds equal to 0.001, 0.005, 0.01, and 0.02, respectively. In the map of HT18_9, cells with RB rate $h \geqslant 18$ and cells with RB rate $h \geqslant 9$ that are also adjacent to the $h \geqslant 18$ cells are considered as hotspots.

Table 3
Comparison results of the hotspot maps. The number in front of the brackets is the amount of cells that being classified as hotspots in both maps. The number inside the brackets is the Kappa Index between the two maps.

	HT18	HT9	HSA	HOT001	HOT005	HOT01	HOT02	HT18_9
HT18	301(1.00)							
HT9	301(0.38)	1245(1.00)						
HSA	262(0.39)	668(0.74)	1094(1.0)					
HOT001	301(0.69)	561(0.61)	456(0.61)	561(1.0)				
HOT005	301(0.73)	523(0.58)	428(0.58)	509(0.95)	523(1.0)			
HOT01	301(0.69)	567(0.62)	457(0.61)	546(0.98)	511(0.95)	567(1.0)		
HOT02	301(0.74)	508(0.57)	416(0.57)	504(0.95)	487(0.96)	507(0.94)	508(1.0)	
HT18_9	301(0.63)	645(0.67)	523(0.66)	496(0.87)	475(0.85)	501(0.88)	466(0.84)	645(1.0)

considered as areas with or without RB rates. There is not enough information for HSA to tell if a cell contains 80%, or only 20% residential areas. This results in a further precision lost (82%). The HOT model outperforms HSA under current setting of parameters because not only the target crime rate, but also the related variables have been taking into account in HOT. By using the informative GDPatterns, only the areas with similar background (or similar characteristics of related variables) as the hard threshold hotspots are considered. The use of GDPatterns ensures that the precision of the HOT hotspot maps (86% in average, Table 4) will consist with the original inputs.

To give an intuitive view of HOT's performance, two of the hotspot maps, HT18 and HOT001 (Fig. 5) are projected with satellite images of the studied city and a figure of sample site is extracted (Fig. 7). Using an initial threshold ($h \geqslant 18$) the red cells are classified into hotspots and cells in same blocks (in the color of blue)

Table 4
Cell statistic of the hotspot maps. The number in front of the brackets is the amount of cells located in the corresponding area. The number inside the brackets shows the percentage.

	Total hotspot cells	Cells in residential areas	Cells in non-residential areas
HT18	301	257(85.4%)	44(14.6%)
HT9	1245	1056(84.8%)	189(15.2%)
HSA	1094	901(82.4%)	192(17.6%)
HOT001	561	484(86.3%)	77(13.7%)
HOT005	523	451(86.2%)	72(13.8%)
HOT01	567	488(86.1%)	79(13.9%)
HOT02	508	435(85.6%)	73(14.4%)
HT18_9	645	548(85.0%)	97(15.0%)

have been left out. Understandably, houses in the same block are at similar risk of being broken into. Our optimization method successfully captures these cells. Other than a choropleth mapping tool, the HOT performs a dasymetric mapping by modifying the hotspot boundaries rationally. Also, locations covered by natural land, parking lots, roads, and highways are identified and are classified out of hotspots using our method (Fig. 7).

4.3. Demonstrating crime related variables

One thousand five hundred GDPatterns in the experiment satisfying a support threshold of 0.001 are selected for further analysis. These GDPatterns (H-GDPatterns) are sorted by growth ratios from high to low. All 1500 patterns have a growth ratio greater than 50 ($\delta > 50$). For comparison, a set of GDPatterns (N-GDPatterns) based on normal areas are also mined using HOT. Specifically, we set cells with $h \geqslant 18$ as D_n, cells with $18 > h' \geqslant 9$ into $D_{h'}$ and other cells into D_h ($h < 9$). In order to facilitate the comparative analysis, 1500 top N-GDPatterns are selected after running HOT. The growth ratios of these N-GDPatterns are all larger than 30 ($\delta > 30$).

Using the similarity method discussed in Section 3.4, the distance between each pair of GDPatterns is calculated. We use the cluster heat map tool (Wilkinson & Friendly, 2009) to visualize the clusters in sorted distance matrices (Fig. 8). In sorted distance matrices, the value of a_{ij} represents the distance between GDPattern i and GDPattern j, where GDPattern j is the $|i - j|$th closest to GDPattern i by distance. The heat maps use different colors to represent the different values in the sorted distance matrices.

After locating all the clusters, the footprints of these clusters are drawn (Fig. 9), which demonstrate the spatial distribution of GDPatterns. Moreover, we use pie-chart to explore the structure of GDPatterns in the same clusters (Fig. 10), in which the values of variables are shown using different colors.

A lot of information can be revealed from these figures. For example, when we look at the H-GDPattern clusters in the studied city,

- High residential burglary (RB) rates are associated with high population density only in areas with few foreclosures (FC), commercial burglaries (CB), motor-vehicle larcenies (MV),

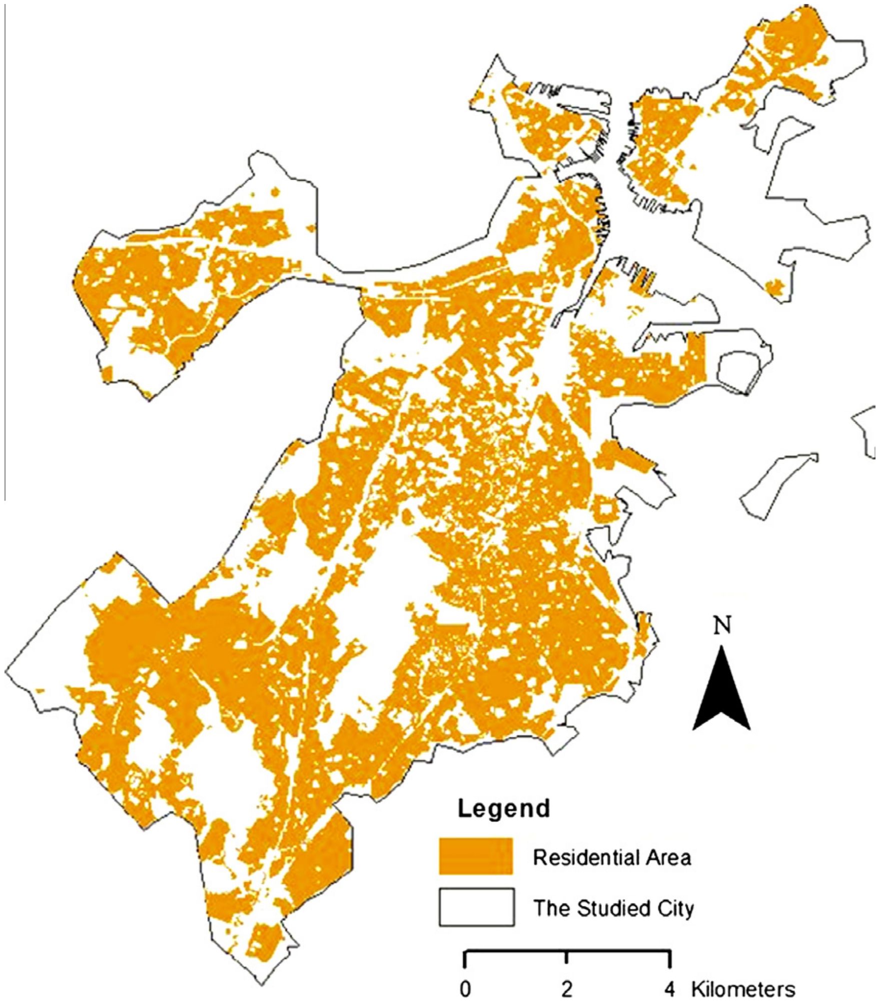


Fig. 6. A land cover map showing the residential areas in the studied city.

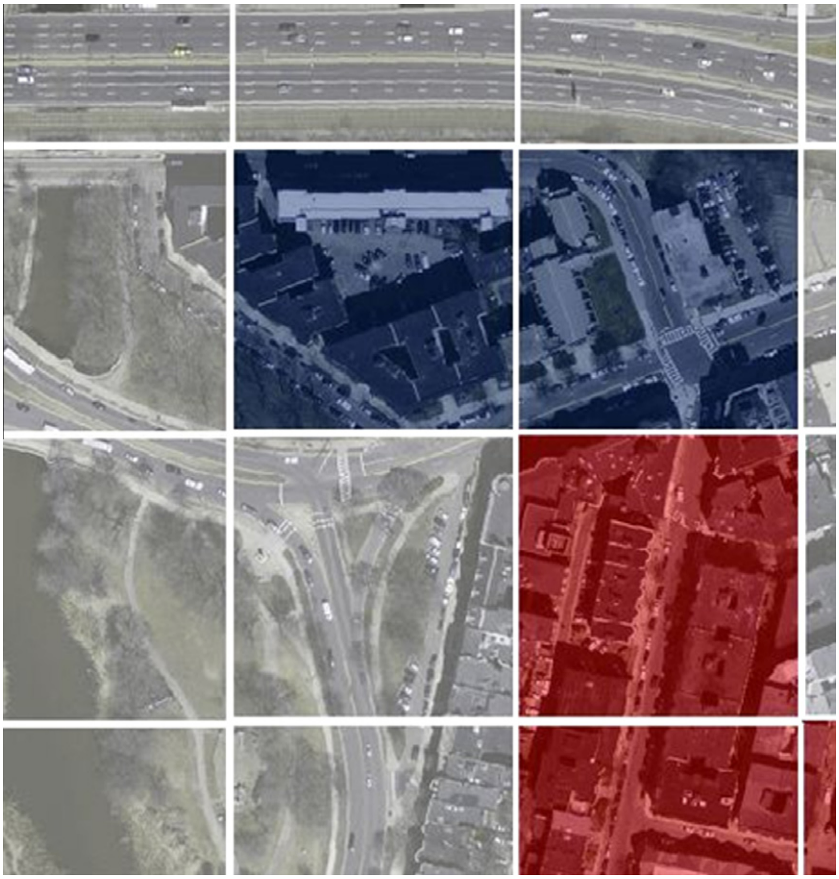


Fig. 7. An example of re-projected hotspots with satellite images. The blue cells are hotspots defined using a threshold of $h \geqslant 18$ (HT18 in Fig. 5). Both the blue and red cells belong to the hotspots identified using HOT with a support threshold of 0.001 (HOT001 in Fig. 5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

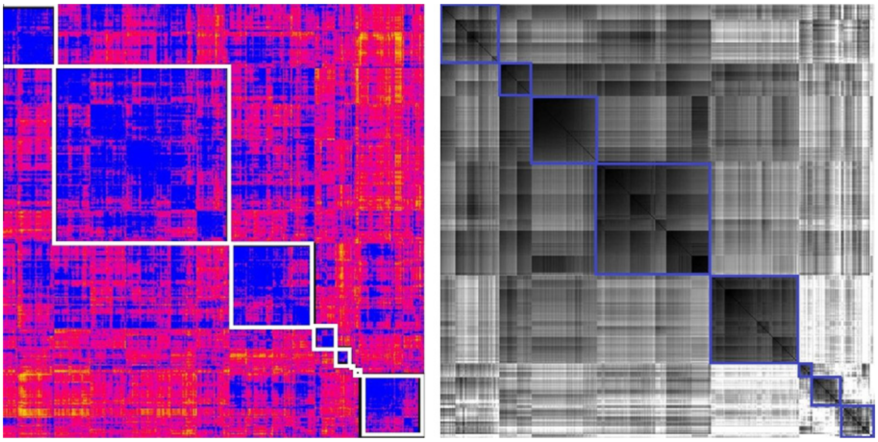
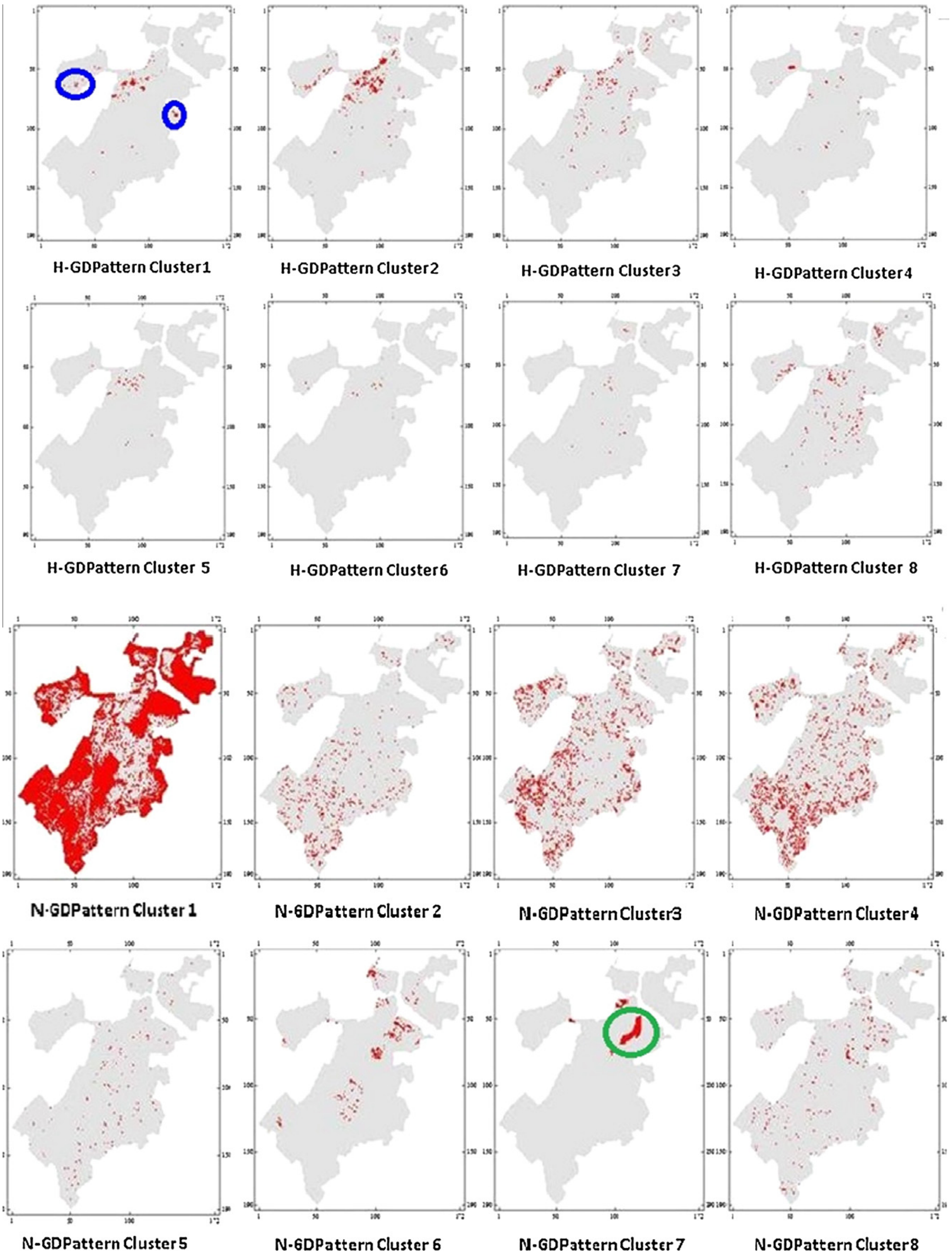


Fig. 8. Heat maps for distance matrices of GDPatterns. On the left side a heap map based on distance matrix of H-GDPatterns is drawn by using the color ramp from blue to red representing distances between H-GDPatterns from small to great. GDPattern clusters that identified using HAC (Section 3.4) are marked with white frames. On the right is the heat map for the distance matrix of N-GDPatterns with color ramp from black to white representing distances from small to great and GDPattern clusters are marked with blue frames. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

street robberies (SR), and very low arrest rates (AR) (Cluster 1). These areas also have high residential density (HU) and are close to universities or colleges (DC). Such locations are shown in the footprint map of H-GDPattern Cluster 1 in Fig. 9.

- High residential burglary (RB) rates are associated with very low foreclosure rate (FC) in most instances (Cluster 1–7). The only locations with many residential burglaries (RB) and a moderate number of foreclosures (FC) are shown in Fig. 9, H-GDPattern Cluster 8. These areas are usually far from universities or colleges, have average population and house density, and low to moderate arrest (AR), commercial burglary (CB), motor-vehicle larceny (MV) and street robbery (SR) rates (Cluster 8).
- Areas with high residential burglary rates and not close to any colleges or universities (low in DC) can be mainly considered in two categories (Clusters 4 and 7 in Fig. 10). One of them is



H-GDPattern Cluster1

H-GDPattern Cluster2

H-GDPattern Cluster3

H-GDPattern Cluster4

H-GDPattern Cluster 5

H-GDPattern Cluster6

H-GDPattern Cluster 7

H-GDPattern Cluster 8

N-GDPattern Cluster 1

N-6DPattern Cluster 2

N-GDPattern Cluster3

N-GDPattern Cluster4

N-GDPattern Cluster5

N-6DPattern Cluster 6

N-GDPattern Cluster 7

N-GDPattern Cluster8

Fig. 9. Footprint maps of GDPatterns' clusters. Areas inside blue circle are where most colleges located in the studied city and the green circle indicate the centre park of the city. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

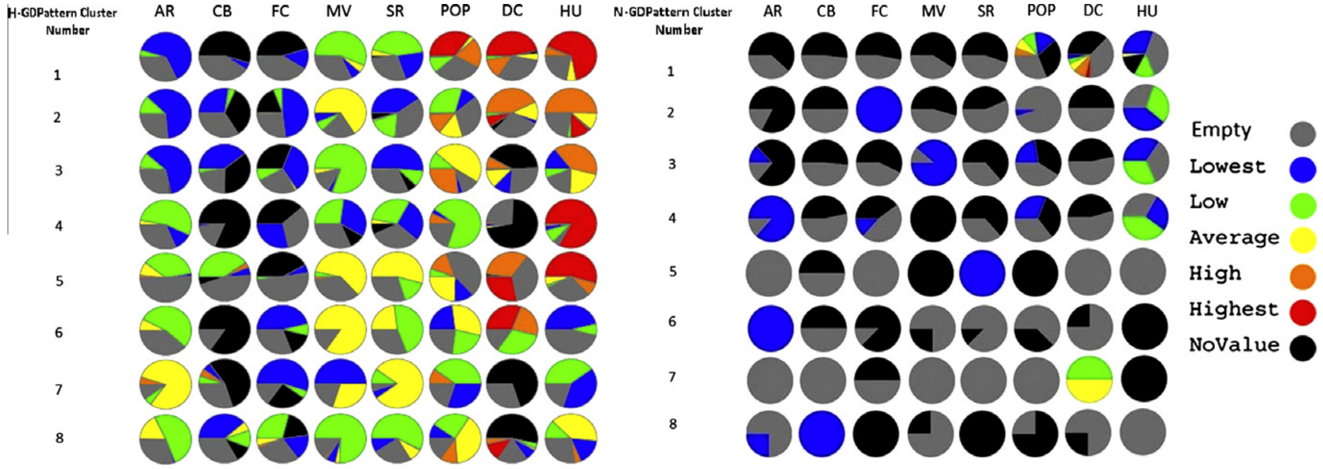


Fig. 10. Pie-charts of GDPatterns' clusters. The values of each related variable are shown in different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

characterized by high residential density (HU), as well as low motor-vehicle larcenies (MV) and street robberies (SR) rates (Cluster 4). The other has low residential density and average MV and SR rates (Cluster 4). The locations of the two categories are shown in H-GDPattern Cluster 4 and H-GDPattern Cluster 7 of Fig. 9, respectively.

The information revealed by our approach has been verified by domain scientists. For example:

- Offenders are known to focus on neighborhoods with large proportions of college students living in off-campus residences (the blue circles in Fig. 9 show areas where most colleges located), (Fig. 10, H-GDPattern Cluster 1 in which the value of DC is high).
- Where college students are less significantly represented, offenders take a different approach, and the FC rates become a more important indicator of RB offenders (Fig. 10, H-GDPattern Cluster 8 in which the value of FC is relatively high). This also explains why high RB is associated with low FC in most areas of the city.
- The footprint map of N-GDPattern Cluster 7 (green circle in Fig. 9) covers mostly non-residential areas like parks, because these areas have similar conditions and no RB incidents.

The case study and the comparison experiments have shown the potential of using crime related variables in hotspot mapping. Our method helps maintain the mapping precision during the hotspots representation process and also provides a comprehensive way for further analysis.

5. Conclusion

In this paper, we present a spatial data mining framework to study the spatial distribution of crimes through their related variables. To the best of our knowledge, it is the first attempt to use related variables in crime hot spot mapping. Spatial data mining is often said to ''let the data speak for themselves''. But the data cannot tell stories unless appropriate questions are formulated and asked, and appropriate methods are needed to solicit the answers from the data. In the framework we address an iterative and inductive learning process to study the spatial properties of crime. Experiment results show that our HOT model outperforms HSA in precisely identifying crime hotspots. Additionally, by using a similarity measure method, we demonstrate the characteristics of target crime's related variables using GDPattern clusters and footprint maps, which help explaining the varying of crime over space and deliver the knowledge in a quantitative, as well as comprehensive and systematic manner.

Acknowledgement

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

References

Agrawal, R., Srikant, R., (1994). Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB (Vol. 1215, pp. 487–499).

Bailey, T., & Gatrell, A. (1995). *Interactive spatial data analysis*. Longman Scientific & Technical Essex.

Bates, S. (1987). Spatial and temporal analysis of crime. Research Bulletin, April.

Boba, R. (2005). *Crime analysis and crime mapping*. Sage Publications, Inc..

Brantingham, J., & Brantingham, L. (1984). *Patterns in crime*. New York: NCJ.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal, 21*(1), 4–28.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cohen, L., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 588–608.

Cook, W., Ormerod, P., & Cooper, E. (2004). Scaling behaviour in the number of criminal acts committed by individuals. *Journal of Statistical Mechanics: Theory and Experiment, 2004*, P07003.

Cornish, D., & Clarke, R. (1986). *The reasoning criminal: Rational choice perspectives on offending*. New York: Springer-Verlag.

Deane, G., Beck, E., & Tolnay, S. (1998). Incorporating space into social histories: How spatial processes operate and how we observe them. *International Review of Social History, 43*(S6), 57–80.

Ding, W., Stepinski, T., & Salazar, J. (2009). Discovery of geospatial discriminating patterns from remote sensing datasets. In: Proceedings of SIAM international conference on data mining.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 43–52). ACM.

Eck, J., Chainey, S., Cameron, J., Leitner, M., & Wilson, R. (2005). *Mapping crime: Understanding hot spots*. National Institute of Justice.

ESRI (2011). Arcgis desktop: Release 10.

Ester, M., Kriegel, H., & Sander, J. (1997). Spatial data mining: A database approach. In *Advances in spatial databases* (pp. 47–66). Springer.

Getis, A., & Ord, J. (2010). The analysis of spatial association by use of distance statistics. *Perspectives on Spatial Data Analysis*, 127–145.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., & Hsu, M. (2000). Freespan: Frequent pattern-projected sequential pattern mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 355–359). ACM.

Harries, K. (1999). Mapping crime: Principle and practice. US Dept. of Justice, Office of Justice Programs, National Institute of Justice, Crime Mapping Research Center.

Herrera, F., Carmona, C. J., González, P., & del Jesus, M. J. (2011). An overview on subgroup discovery: Foundations and applications. *Knowledge and information systems, 29*(3), 495–525.

Hirschfield, A. (2001). *Mapping and analysing crime data: Lessons from research and practice*. CRC.

Janeja, V. P., & Palanisamy, R. (2012). Multi-domain anomaly detection in spatial datasets. *Knowledge and Information Systems*, 1–40.

Jenks, G. (1967). The data model concept in statistical mapping. *International Yearbook of Cartography, 7*, 186–190.

Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. In *Advances in spatial databases* (pp. 47–66). Springer.

Lin, D. (1998). An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning, San Francisco (Vol. 1,, pp. 296–304).

Ludwig, J., Duncan, G., & Hirschfield, P. (2001). Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment. *The Quarterly Journal of Economics, 116*(2), 655–679.

Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., et al. (2010). A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics, 16*(2), 205–220.

Malerba, D., Esposito, F., Lisi, F., & Appice, A. (2002). Mining spatial association rules in census data. *Research in Official Statistics, 5*(1), 19–44.

Mennis, J. (2006). Socioeconomic-vegetation relationships in urban, residential land: The case of denver, colorado. *Photogrammetric Engineering and Remote Sensing, 72*(8), 933.

Mennis, J., & Liu, J. (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS, 9*(1), 5–17.

Miller, H., & Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC.

Mu, Y., Ding, W., Morabito, M., & Tao, D. (2011). Empirical discriminative tensor analysis for crime forecasting. Knowledge Science. *Engineering and Management*, 293–304.

Nguyen, H., & Nguyen, S. (1998). Discretization methods in data mining. *Rough Sets in Knowledge Discovery, 1*, 451–482.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. *Database Theory ICDTT, 99*, 398–416.

Qian, F., He, Q., Chiew, K., & He, J. (2012). Spatial co-location pattern discovery without thresholds. *Knowledge and Information Systems*, 1–27.

Ratcliffe, J., & Taniguchi, T. (2008). Is crime higher around drug-gang street corners? Two spatial approaches to the relationship between gang set spaces and local crime levels. *Crime Patterns and Analysis, 1*(1), 17–39.

Rossiter, D. (2004). Technical note: Statistical methods for accuracy assessment of classified thematic maps. *Enschede (NL): International Institute for Geo-information Science & Earth Observation (ITC), 25*(92), 107. <http://www.itc.nl/personal/rossiter/teach/R/R_ac.pdf>.

Sah, R. (1991). Social osmosis and patterns of crime: A dynamic economic analysis. *Journal of political Economy, 99*(6).

Sampson, R., Raudenbush, S., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277*(5328), 918–924.

Short, M., Bertozzi, A., & Brantingham, P. (2010). Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression. *SIAM Journal on Applied Dynamical Systems, 9*, 462.

Skogan, W. (1992). *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ. of California Pr..

Tita, G., & Greenbaum, R. (2009). Crime, neighborhoods, and units of analysis: Putting space in its place. *Putting Crime in its Place*, 145–170.

Van Patten, I., McKeldin-Coner, J., & Cox, D. (2009). A microspatial analysis of robbery: Prospective hot spotting in a small city. *Crime Mapping: A Journal of Research and Practice, 1*(1), 7–32.

Wand, M., & Jones, M. (1995). *Kernel smoothing* (Vol. 60). Chapman & Hall/CRC.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician, 63*(2), 179–184.

Williamson, D. McLafferty, S., McGuire, P., Ross, T., Mollenkopf, J., Goldsmith, V., et al. (2001). 9 tools in the spatial analysis of crime.*Mapping and Analysing Crime Data: Lessons from Research and Practice*, 187, CRC.

Yu, K., Ding, W., Simovici, D. A., & Wu, X. (2012). Mining emerging patterns by streaming feature selection. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*(pp. 60-68). ACM.