Spatial and Temporal Analysis of Crime for the Discovery of Hot Spots in Road Networks: the case of Boston

Sofia M. Nikolakaki

Abstract

Abstract The formation of criminal hot spots is a result of high concentration in criminal activity in a specific area. Restricting their discovery to human reasoning reduces the possibility of confining this phenomenon. A major challenge in hot spot analysis research is the lack of withstanding definitions and techniques thus making the specific task application-oriented.

This work studies the dynamic behavior of criminality in the example of Boston with the goal to discover crime hot spots within road network neighborhoods. In order to do so we follow a two-step methodology. The first step involves temporal criminal analysis and dynamic crime mapping using Geographic Information System (GIS) techniques. We extract and integrate spatial, temporal and criminal occurrence information to construct the visual product of the spatio-temporal criminality dynamics on the road network of Boston. We utilize graduated symbols and color gradient dots to illustrate data points. The second step quantifies the "dangerousness" of an area surrounding a location to allow hot spot mining. For this purpose we follow a statistical approach using crime counts and the Location Quotient (LQC) crime measure as well as a clustering analysis method with the use of the well-known agglomerative hierarchical clustering algorithm. Both procedures successfully discover high-crime density places, with the former confirming the results with numerical values and the latter with community visualizations.

Keywords: Crime Hot Spots, Boston Road Network, Spatio-temporal Clustering

I. INTRODUCTION

A. The Role of Crime Mapping in the Study of Crime Dynamics

Criminal activity in all its forms has indisputably always been a matter of significant essence.The spatial and temporal levels of detail in the criminal activity analysis render crime mapping imperative towards comprehending this ever lasting phenomenon. Crime mapping is the act of assigning points to a map to represent aggregated criminal incidents reported at specific locations. Developing different visual map displays of the distribution of criminality over time makes the identification of crime trends, patterns and hot spots more natural and less complex for the crime analysts. The term of crime mapping is closely related to the Geographic Information System (GIS) term. The latter is defined as a computer system that combines geographic coordinates with layers of information to produce a representation of the earth's surface integrated with the quantitative data related to specific positions. Related work suggests that it is the most dominant tool in crime analysis since the strong majority of crime reports contain location-related attributes which link crime information to a map[1],[2]. That said, the first part of this report focuses on the spatio-temporal analysis of crime occurrences and on crime mapping of incident reports. For this purpose we created our own Geographic Information System (GIS) methodology. This section provides answers to questions derived from spatio-temporal analysis, such as "What trends does criminality follow over time in Boston?" and "How has criminality spatio-temporary evolved in Boston?".

B. The Challenge Of Defining Hot Spots

The second part of this work addresses the discovery and computation of crime hot spots. Even though mining such high-crime density places is critical for community policy there is no straightforward approach towards this direction due to the lack of withstanding definitions of this notion. The common understanding is that a crime hot spot is an area that displays frequent occurrences of criminal activity. The research community suggests that hot spots should be defined in terms of the study case where they are encountered in. To this we add that they are also determined based on the techniques and methods followed towards their identification.

Criminal analysis is a field that combines knowledge from disciplines such as sociology, geography, law enforcement agencies and statistics. Its interdisciplinary background led

to visualization being the most convenient method for a crime interpretation. Problems encountered with this method are ambiguity within the results (overlapping, points stacked on top of one another) which make the task of analysis arduous and subjective to the current reader. Statistical crime analysis attempts to resolve these issues by introducing measurements such as crime counts and the LQC measure. Using this approach however, the results become variable dependent and less naturally interpretable. In this work we explore both approaches to evaluate their efficiency in our setting. This section provides answers to questions derived from hot spot analysis, such as "Which are the hot spots in the Boston area and how are they distributed?" and "Do hot spots change over time?".

II. BACKGROUND

Attempts for a sophisticated hot spot definition have been made without significant resonance. The methods we followed for hot spot areas discovery determined our interpretation of hot spots. More specifically, for our statistical crime analysis we define hot spots to be circular areas with a radius of k km that display a significantly high and dense criminality behavior compared to other areas over the years 2012 to 2015, based on their LQC index. The value of k is recommended to be defined by the user, based on the desired resolution. In our setting we consider k to be equal to 1 because within this radius locations exhibited common criminal activity patterns. For our visual analysis we define a hot spot to be a community produced by the clustering technique that displays a significantly high and dense criminality behavior compared to other areas over the years 2012 to 2015. The produced number of hot spots is user dependent due to the subjective notion of "dangerousness" of an area. Also, we are interested in hot spots persistent in time and therefore we examine a 4-year time window. We could reduce this number to a smaller period of time but the criminal data did not contain certain crime incident reports in 2012 and 2015.

Spatio-temporal crime mapping has become easier with the development and improvement of Geographic Information Systems (GIS), software systems that visually represent information on maps. We consider GIS to be any software which satisfies the aforementioned definition. In general however, data analysts associate this term with powerful, closed source and expensive software packages running as desktop applications with rich mature feature sets.

GIS utilize certain basic visualization techniques to complete their task. Two of these

are graduated symbols and color gradient dots. The graduated symbols method is used to depict the number of crime incidents on map locations with repeated occurrences of crime activity. This technique represents the concentration of a crime between other geographical locations by changing the size of the dot. The biggest the dot, the more criminal acts were performed in the corresponding place. The color gradient dots method utilizes color palettes to illustrate differences in crime concentration among different locations. Similarly to the graduated symbols technique different colors show different crime intensity.

Statistical crime analysis is less widespread due to the interdisciplinary background of crime study, which leads to the prevalence of visual tools. In this work we evaluated two crime analysis measurements, crime counts and the Location Quotients Criminal (LQC) measure. We also attempted a visualization analysis by introducing the well-known agglomerative hierarchical clustering algorithm. Note that the visualization technique also considers statistics to produce an outcome. The crime counts measurement simply aggregates crime incidents in a specific area. LQC[3] is an index of criminality activity in an area compared to its surroundings. It is therefore an indicator of the criminal activity at a regional area relative to the overall studied region that contains the specific area of interest, i.e. it computes crime concentration. Let C_i, i=1,2,...,K denote the number of crimes within the area i. Then, in our application LQC equals to:

$$LQC = \frac{S_i}{S} \quad S_i = \frac{\sum_{i=1}^{K} C_i}{K} \quad \text{and} \quad S = \frac{\sum_{i=1}^{T} C_i}{T} \tag{1}$$

where K shows intersections in a region and T denotes all the intersections in Boston.

The Fowlkes-Mallows index is an external evaluation method that was proposed to determine the similarity between two hierarchical clusterings. It is interpreted as the geometric mean of precision and recall. For its definition and detailed description we refer the reader to[4].

In this report we consider the agglomerative hierarchical clustering method to be basic and therefore we omit a detailed description. We refer the interested reader to[5]. We will focus on its use for spatial clustering which gathers locations in a geographic space that display excess of an event, which in our case is the number of criminal incidents. In our spatial clustering analysis we also consider these locations to be close.

III. RESEARCH METHODOLOGY

A. Spatio-temporal Crime Mapping

The application-oriented direction of this work requires the integration of two data sets for the road network construction. The first is the publicly available **Crime Incident Reports** data set, provided by the City of Boston. It comprises 268057 crime incident reports in the Boston area that were evaluated and confirmed as crimes by the police. The report time indexing begins on 07/09/2012 and ends on 08/10/2015. The data set attributes that were used in our implementation were the reported date and time, the street and the location coordinates (latitude, longitude) of the incident. The geographical data required for crime mapping were offered by Open Street Maps (OSM). In particular, we extracted the latest OSM data of the city of Boston.

Our spatio-temporal crime mapping approach involved the following GIS methodology of two layers. The first and basic layer is the construction of the Boston road network in a convenient to be mapped manner. Let the undirected graph $G = (V, E, W)$ be a road network, where vertices V show intersections, edges E denote roads and weights W be crime counts on intersections. To create this network we used information from the OSM file that represented roads. After performing filtering and data cleansing to the data set we created the final road network of Boston which contained 35924 road intersections and 18685 roads.

The second layer introduced crime information and time parameter to the Boston road network. One major challenge during this process was the assignment of crime incidents to locations. A naive approach would correspond each incident to its corresponding location on the map. However, determining the resolution of crime mapping is not purely a function of crime and location, as the scale of analysis should also be taken into account. In our case we considered point precision analysis, instead of area precision, to offer deeper insight into the crime information illustrated on the map. Another challenge was introducing the parameter of time. This work considers snapshots of the network, each illustrating the aggregated number of crime incidents on the map during successive years, months and hours. The final step of the spatio-temporal crime mapping was to provide a visualization of the data structured road network snapshots on a map. Note that all visualizations where created with the use of Tableau. When aiming for both, quantification and visualization, criminality dot maps are superior to other forms of mapping. For this purpose we used graduated symbols and color gradient dots. These methods were selected due to their efficiency in city-scale

analysis and comparison. In particular, on a single map we used the former method to show the distribution of crime counts given a specific temporal parameter and the latter to denote another comparable temporal parameter.

B. Discovering Hot Spots

This stage required computing the crime counts and the LQC statistical measurement, as well as performing the agglomerative hierarchical clustering algorithm. We created our own implementation for the computation of crime counts and LQC and used the igraph package of the R project to perform clustering.

The main challenge during this stage was to create an appropriate object prior to clustering and to map the produced outcome to the physical road network for interpretation. Recall that we consider a weighted road network, where the weights are crime counts on road intersections (vertices). To perform clustering we aimed for an equivalent graph with the weights being on the edges. For this purpose, we transformed intersections into edges and roads into vertices, thus preserving information of locality and connectivity.

The output of this algorithm contained clusters with their respective members. To support straightforward visual representation of communities, we mapped the produced data to the natural Boston map. Since the intersections are the denoted elements of the map, each road within a cluster was dissociated to its individual intersections so that road communities were transformed to intersection communities. Then, we designated locations within the same cluster to have the same color and used the graduated symbols method to show crime incident numbers.

IV. DISCUSSION OF THE RESULTS

A. Dynamics of Criminal Activity on Road Network

During the discussion of the results bare in mind that year 2012 comprises incidents reported after July and year 2015 contains incidents reported until June, therefore data before and after these months are missing.

Spatio-termporal analysis of crime answers to the questions presented in I. Figure 1 shows the total crime incidents of each year per month. We observe that criminal activities tend to remain higher during warm months, especially during July and August, when the weather is warm and tourists arrive in the city, while it considerably decreases during the winter, when the cold weather limits outdoor activities. The temporal analysis of crime incidents over days

showed very small fluctuations between subsequent days. A slight increase was observed on Friday and Saturday when there are more people on the streets. A 4-hour window temporal analysis suggested that criminality becomes intense during the hours 12:00-16:00 and 08:00-12:00. Given that we would expect late night hours to be more dangerous, we consider that the time of the crime incident corresponds to the report time stamp and not to when it occurred. This is not stated however in the dataset.

At this point we include geographical information into our analysis and the results are illustrated in Figure 2. All the map snapshots compare the crime distribution of February and August. These months were selected because their respective temporal analysis showed dissimilar crime numbers and it would be interesting to explore if they also presented dissimilar space distribution. The gray color represents February and the size of the dot denotes August. As the color becomes darker or the dot becomes bigger, the number of criminal acts increases. Furthermore, the left and right maps represent the overview of Boston in years 2013 and 2014 respectively so that we can also compare how crime concentration changed over subsequent years. We omit years 2012 and 2015 due to lack of space. Note that in general, criminality in Boston is gathered towards the center rather than in the suburbs. Also, from 2013 to 2014 criminality seems to increase and expand around the most high-crime density areas. This phenomenon becomes clear with the additional numerical results presented in Table I, which contains the intersections with the biggest crime counts in 2013 and the crime counts of the same locations in 2014. We notice a slight increase in the numbers of crime incidents.

Location	Crimes Feb 2013	Crimes Aug 2013	Crimes Feb 2014	Crimes Aug 2014
(42.3524,-71.0646)	25	109	25	67
(42.3290,-71.0861)	52	38	52	48
(42.3489,-71.0818)	13	32	25	30
(42.2848,-71.0912)	25	25	26	27
(42.3329,-71.0923)	21	14	26	19

TABLE I: Locations with a high number of crime incidents.

B. Discovering Hot Spots

We computed the LQC measurement based on the hot spot definition for statistical crime analysis provided in I B. The center locations of the five areas with the highest overall LQC index are shown in II. The same table displays the total number of crimes that occurred solely in the center locations to evaluate whether crime aggregation is a strong indicator of crime concentration. Observe that lines 2,3 and 4 have similar crime counts but considerably different LQC. Thus, we conclude that crime counts cannot summarize crime information of the surrounding area. The presented hot spots display the same high LQC trends between the years. We present the total LQC because it is a stronger indicator of the overall differences between crime concentrations in these areas as it aggregates the amount of difference encountered in each year.

Rank	Location	Crime Count	Crime LQC
1	(42.35238,-71.064550)	1631	18.13
2	(42.329010,-71.08607)	1103	16.87
3	(42.348890,-71.08178)	1017	12.58
4	(42.325867,-71.06340)	1019	11.31
5	(42.284813,-71.09116)	986	10.96

TABLE II: Top ranked locations based on LQC and the respective crime count measurement.

The results of the agglomerative hierarchical clustering analysis are illustrated in Figure 3. This representation includes the total aggregated crime incidents and spatial information from 2012 to 2015. Dots with the same color depict intersections within the same cluster, whereas bigger dot sizes show more criminal activity. Both maps correspond to the same snapshot, with the right being a magnified version of the left. The visual representation of the clustering outcome allows validating its performance. Observe the existence of dense communities with well defined community boundaries and proximity regarding crime behavior within a community.

To extend the spatial clustering to a spatio-temporal analysis we used the Fowlkes-Mallows index. Community differences between 2012 and 2013 yielded an index value equal to 0.6381647. The same value equals 0.6729145 and 0.5449903 from the differences between years 2013 to 2014 and 2014 to 2015 respectively. A higher index is proportional to

the number of true positives so the clusters of 2013 and those of 2014 display the biggest similarity.

We identified the top 5 hot spots with the use of LQC. Observing Figure 3 also reveals hot spots defined based on the definition of visual analysis. For spatial-clustering we observed the maps between different years. The bold communities that are conspicuous over time are considered the hot spots and are denoted with a black circle in Figure 3. We cannot directly compare the results of LQC and clustering directly because they consider different regions shapes and there are no comparable results in the literature. It is worth mentioning though, that the areas identified by LQC were also discovered as communities by agglomerative, but the spatio-temporal clustering. Also, selecting one of the two approaches depends on whether numerical precision or area identification on the map is more important for an application.

V. CONCLUDING REMARKS

The interdisciplinary interest in crime analysis stresses out the importance of comprehending this ever lasting phenomenon and mining hot spots that significantly contribute to it. Defining the properties of these areas alone is challenging as they become application dependent, and therefore the possibility of comparison and evaluation of proposed methods is limited. Visualization approaches with GIS tools are dominant in this area, but also expensive due to proprietary GIS. To address this problem we provided the complete description of a GIS implementation. Our crime mapping methodology and final outcomes are comparable to results produced by professional tools. In addition, we address the problem of defining and discovering hot spots with a statistical crime analysis and a visualization approach. All of the proposed methods were tested for the case of Boston.

In the first section of our work we presented a temporal and a spatio-temporal crime analysis. Temporal analysis showed that crime incidents are intense during the warm months, towards the end of the weekend, and with some uncertainty, during the hours 12:00-16:00. Introducing the additional parameter of geographical information led to the conclusion that over time the hot spots in Boston remain and that criminality tends to increase in these locations while spreading to the surrounding areas. In the second section of this report we defined and discovered hot spots. Results showed that both the LQC and the agglomerative clustering approaches succeeded in finding high-crime density areas. We concluded that the

selection between the two approaches depends on whether an application requires numerical precision or a map overview of hot spot areas. At the end we used the Fowlkes-Mallows index to extend the spatial clustering interpretation to a spatio-temporal one and observed that communities generally remained over the years, with the most similar ones being in 2013 and 2014.

A possible future direction of this work could evaluate the performance of other statistical crime analysis and clustering methods in order to compare the presented findings. For this purpose other internal and external clustering evaluation indexes should also be tested to generalize our observations.

VI. APPENDIX

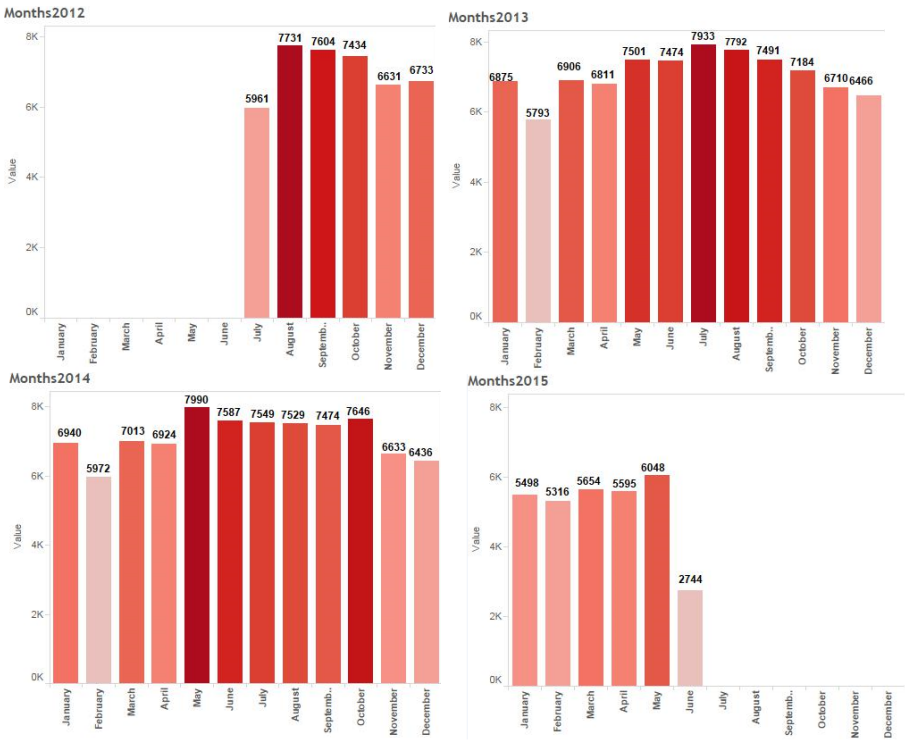

FIG. 1: Total crime incidents of each year per month

---

[1] J. Eck, S. Chainey, J. Cameron, and R. Wilson (2005).

[2] K. A. Harries, Tech. Rep. (1999).

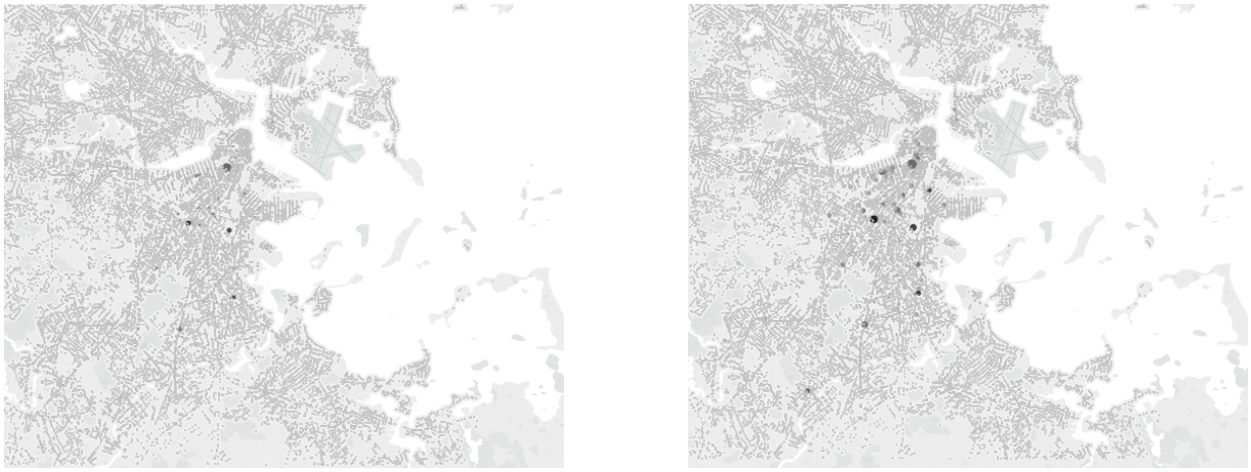

FIG. 2: Comparison of crime distribution between Feb and Aug in 2013(left) and 2014(right). Gray color of dot shows Feb and size of dot shows Aug.

FIG. 3: Discovering hot spots with agglomerative clustering. Same colored dots are in the same cluster. The size of the dot shows number of crimes. The right map is a magnified version of the left. The black circles denote hot spots

[3] P. L. Brantingham, P. J. Brantingham, et al., Crime mapping and crime prevention **8**, 263 (1998).

[4] E. B. Fowlkes and C. L. Mallows, Journal of the American statistical association **78**, 553 (1983).

[5] W. H. Day and H. Edelsbrunner, Journal of classification **1**, 7 (1984).