-

Kernel Density Estimation (KDE) vs. Hot-Spot Analysis - Detecting Criminal Hot Spots in the City of San Francisco

Maja Kalinic
University of Augsburg/Department for Applied Geoinformatics
Alter Postweg 118, 86159
Augsburg, Germany
maja.kalinic@geo.uni-augsburg.de

Jukka M. Krisp
University of Augsburg/Department for Applied Geoinformatics
Alter Postweg 118, 86159
Augsburg, Germany
jukka.krisp@geo.uni-augsburg.de

Abstract

Point pattern analysis is one of the most fundamental concepts in geography and spatial analysis. A range of methods can be applied to point pattern analysis, from basic summary to advanced spatial statistics. This paper considers two frequently used types of pattern analysis – kernel density estimation and Gi* statistics and their performance in detecting criminal hot spots in the city of San Francisco. Very often areas identified as hotspots from the Gi* results are not grouped into high density threshold on the kernel density estimation map. When analyzing and mapping spatial patterns (e.g. crime patterns), it is important to ensure that the identification of hotspots is as accurate and effective as possible.

Keywords: point pattern analysis, kernel density estimation, hot spot analysis, crime analysis.

1 Introduction

The analysis of point patterns appears in many different areas of research (Cressie, 2015). In general, a point process is a stochastic process in which we observe the locations of some events of interest within a bounded region (Bivand et al., 2008). Point pattern analysis involves the ability to compare and describe patterns and test whether there is a significant difference to a random spatial pattern (O'Sulivan and Unwin, 2003). As such, it represents one of the most fundamental concepts in geography and spatial analysis.

In its most basic form point pattern analysis attempts to analyse the occurrence of points in a particular space. This involve calculating summary statistics such as Count, Mean, Median, and Standard Deviation. However, these summary statistics are too basic and may hide more valuable information about the observed patterns. We investigate additional analysis methods for exploring point patterns, such as density analysis or statistical operations with different refinements and extensions. Therefore, in this paper we focus on two frequently used types of point pattern analysis – kernel density analysis (KDE) and Gi* hot spot analysis. We perform both types of analysis on the same dataset – criminal data of San Francisco, and discuss their performance based on data characteristics and the case study objective.

2 Related research

Kernel density estimation (Silverman, 1986) methods are often used in visualizing and analyzing spatial data, with the objective of understanding and potentially predicting event patterns (Smith et al. 2015). These methods have a wide variety of applications such as risk assessment and damage analysis (Ahola et al., 2007), emergency planning for the fire and rescue services (Krisp et al., 2005), road accidents (Anderson, 2009). KDE maps are in high demand in crime analysis (Chainey and Ratcliffe, 2005; Ratcliffe, 2010; Bruce and Smith, 2011; Rey et al., 2011; Mburu and Zipf, 2014; Levine, 2017).

KDE is specifically useful in detecting hot spots due to the series of estimations which are made over a grid placed on the entire point pattern. Each of these estimations show the intensity at a certain location and therefore detect the highs and lows of point pattern densities. The user’s only role is to specify appropriate bandwidth for the estimation which plays a decisive role. When the bandwidth is set too large, important information may be lost. In case of a small bandwidth local data information has a more significant impact on the result. To ease this process (Krisp et al., 2009) suggest a bandwidth slider tool which outputs the pre-processed KDE maps simultaneously with the specified bandwidth. This way the influence of the kernel bandwidth to the KDE can be demonstrated clearly and an appropriate bandwidth can be determined visually. Moreover, (Krisp and Špatenková, 2009) indicate on second important issue related with KDE – classification of the kernel density output raster. The aim of the classification is to approximate the original surface as closely as possible by preserving characteristic patterns of the phenomenon. According to (Gilmartin and Shelton, 1989), it is desirable to classify the density values into several categories for visualization. A suitable number of classes should not exceed seven due to limited ability of human eye to discriminate shades. Jenks (1967), proposed a method which uses a measure of classification error (sum of absolute deviations about class means) to keep similar data values in the same class. In this way the classification gives in the most accurate and objective overview of the original data. The method is commonly called natural breaks and has been implemented in variety of GIS software.

It is often desirable to examine patterns at a more local scale. (Getis and Ord, 1992) introduce a family of statistics G, that can be used as measures of spatial association in a number of circumstances. The local statistics, Gi and Gi*, enable us to detect pockets of spatial association that may not be evident when using global statistic. Here, the study area is subdivided into number of regions where each region is identified with a point (often called feature) associated with a value. The null hypothesis says that there is no association between a specific feature of one region and its neighbours (Getis and Ord, 1996). The Gi* statistic measures the degree of this association that results from the concentration of weighted points (or area represented by a weighted point) and all other weighted points included within a radius of distance from the original weighted point (including the point itself). In other words, based on a selected distance, the results for locations containing common neighbours are likely to be correlated (Getis and Ord, 1992; Anselin, 1995). If exists, this correlation will be exhibited by a spatial clustering of high or low values. When there is a prevalence of high values, the resulting Gi* will be positive, while low values will yield negative Gi*.

In practise, Gi* statistic found its application in variety of fields such as crime analysis, epidemiology, voting pattern analysis, economic geography, retail analysis, traffic incident analysis, and demographics. (Kuo et al., 2013) use Gi* statistic to detect hot spots in crash and crime data and suggest more effective organization in police patrolling system. This statistic can be carried out in assessment of spatial clustering of road accidents and hot spot spatial densities. The results can be effectively used for adopting better planning and management strategies and therefore improving traffic conditions as well as accident reduction (Prasannakumar et al., 2011). (Goodwin, Schoby and Council, 2014) use Gi* for examining crash data of teenage drivers to help competent authorities better understand, manage and control high accident locations. In order to identify important factors that influence the distribution of domestic fires in Helsinki (Špatenková and Stein, 2010) apply Gi statistics as a measure of second-order effects indicating dependence in relationships between incidents and their influences. Their study demonstrates how this statistic can provide a useful opportunity for fire brigades to improve planning their efforts.

All this yields to a question what is the difference between a KDE and a Gi* hot spot analysis? It is often the case that results of KDE and Gi* visually yield the same. However, kernel density function (Silverman, 1986) and the Getis-Ord Gi* (Getis and Ord, 1992) statistic are completely different analysis. While KDE aims to detect clusters of high values within the data, Gi* statistic not only detects, but deepens understanding of spatial clusters of the phenomena under study.

3 Case study – detecting criminal hot spots in the city of San Francisco

For performing our analyses, we use San Francisco criminal data. The dataset is free and open-source, available at San Francisco Open Data Portal[1]. It contains crime records stored as separate points which carry spatial component – longitude and latitude of where have they occurred. Table 1 shows dataset attributes (variables) and their detailed description.

Table 1: San Francisco criminal records and their corresponding attributes

Column name	Description
Dates	timestamp of the crime incident
Category	category of crime (e.g. homicide)
PdDistrict	Police Department District Names
X	Longitude
Y	Latitude

The goal of our analyses is to identify areas which have the highest and lowest crime rates. In other words, to detect the safest and the most dangerous zones of the city. In wider application, this answer would be of a great benefit for police departments in relocating their resources across the city for reducing the number of crimes occurring.

Variety of open-source and commercial software provide tools for calculating kernel density and hot spot analysis. Our approach consists of applying both methods kernel density analysis and hot spot analysis for calculating hot spot areas. Furthermore, we compare the outputs and suggest which method showed better results, given the data characteristics and study objectives.

3.1 Kernel Density analysis for detecting crime hot spots

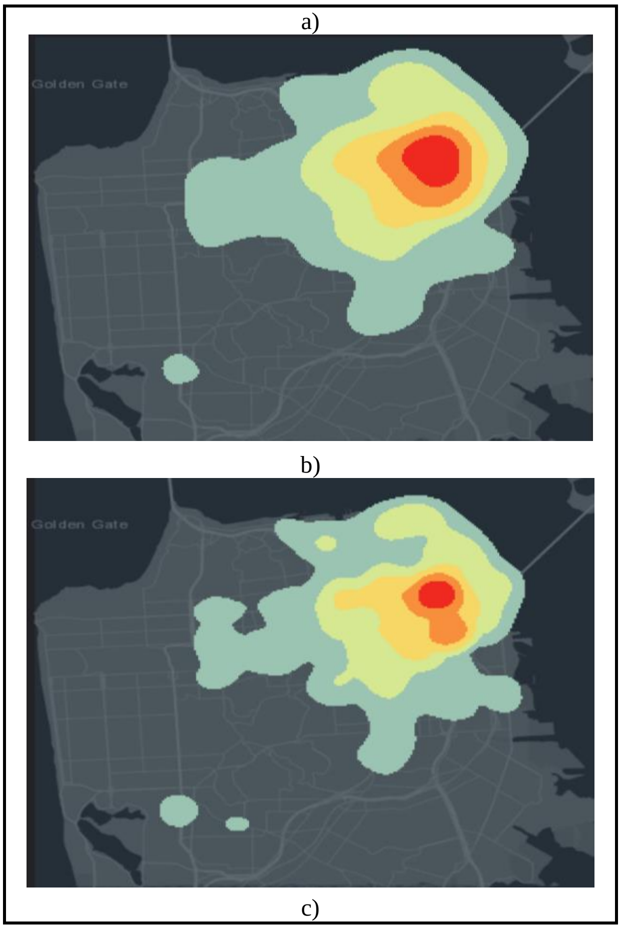
We use kernel density for point features tool to calculate the density of point features around each output raster cell. The algorithm behind the tool fits smoothly curved surface over each point. The surface value is highest at point location and diminishes as the distance from the point increases. It becomes zero at the search radius (bandwidth) distance from the point. If not set otherwise, the tool calculates the bandwidth specifically to the input dataset. The search radius units are based on the linear unit of the projection of the output spatial reference. The cell output size defines the output raster that will be created. This is the value in the environment if specifically set. If the environment is not set, then cell size is the shortest of the width or height of the

[1] https://datasf.org/opendata/

extent of point features in the output spatial reference, divided by 250 (Esri, 2017)[2].

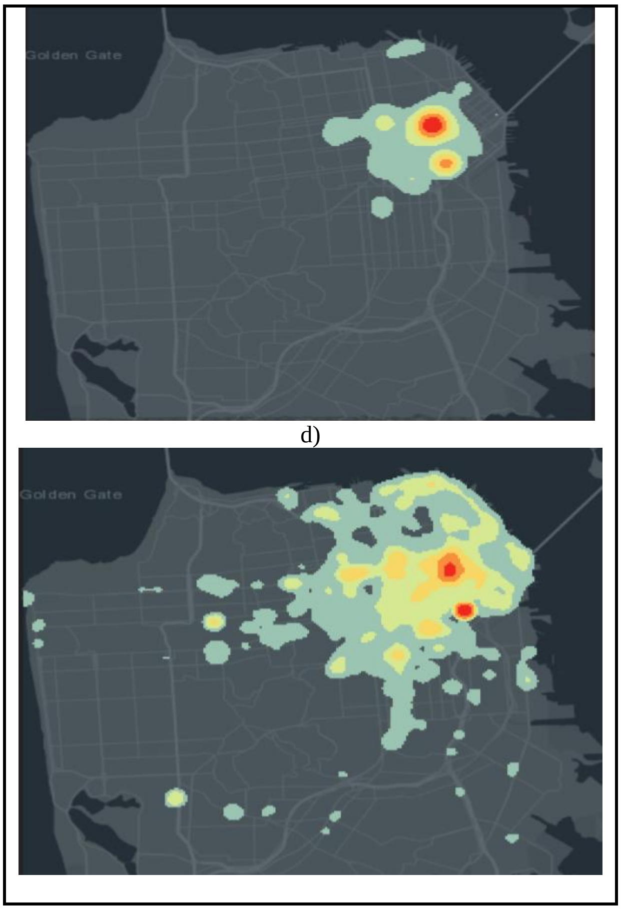
We choose several different bandwidths (1500m, 1000m, 750m and calculated value by default settings) to be able to visually compare their outputs. In addition, we leave the cell size to be calculated by tool's default (The cell size was calculated to 63m for all four examples). Classification method is natural breaks (Jenks), with six classes and green to red color scheme to point out towards highs (red) and lows (green) of point pattern densities. The output of each KDE has the same scale – 1:100 000. Figure 1 shows KDE outputs for the bandwidths of 1500m, 1000m, 750m, and default value, respectively.

Figure 1: Kernel density estimation outputs with search radius of a) 1500m, b) 1000m, c) 750m and d) calculated value by default settings, equal cell size of 63m, map scale of 1:100 000 and natural breaks classification method (where red color shows high and green color low density of points at a given location.


a)

b)

c)


d)

3.2 Hot Spot analysis for detecting crime hot spots

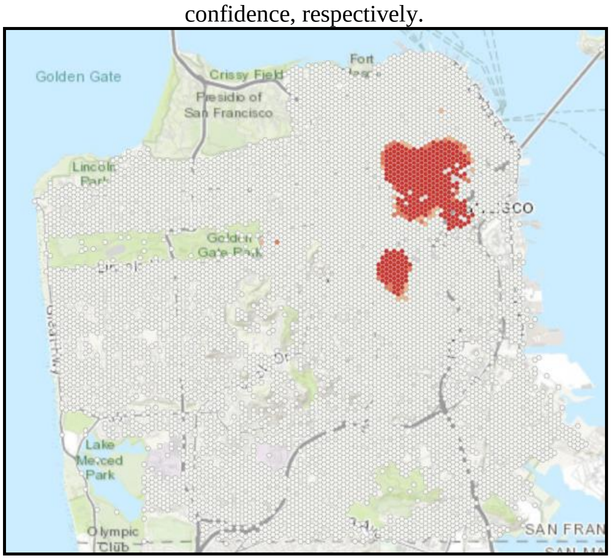
Performed kernel density analyses are able to tell us where clusters in our data exist. However, we are not confident to tell whether these clusters resulted randomly or there is some underlying spatial process behind. Therefore, with our next analyses we want both to detect the patterns and inspect how meaningful they are. Moreover, we want to base our answer on something more confident than human (visual) perception. With hot spot analysis we are able to detect clusters of high and low values in our data. And with the p and z value we are 99%, 95% or 90% confident to tell how statistically significant these clusters are. Hot spot analysis considers a feature (e.g. crime event) in the whole dataset. A feature has a value or, in case of crime events, features are aggregated and their count within the aggregation area represents the value. A feature has a neighborhood which is a group of features around it, including the feature itself. A feature with a high value is interesting but may not be a statistically significant hot spot. To be a statistically significant hotspot, a feature will have a high value and be surrounded by other features with high values as well. The local sum for a feature and its neighbors is compared proportionally to the sum of all features; when the local sum is very different from the expected local sum, and that difference is too large to be the result of random choice, a statistically significant z-score results. For statistically significant positive z-scores, the larger

[2] http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/kernel-density.htm

the z-score is, the more intense clustering of high values (hot spot). For statistically significant negative z-scores, the smaller the z-score is, the more intense the clustering of low values (cold spot).

Since our data include points, we perform aggregation by hexagon bins in order to get the crime counts in each bin as a feature value. The resulted output (Figure 5) shows clusters of statistically significant hot spots in the city of San Francisco. Dark and bright red hexagons are zones of intense clustering with 99 and 95 percent of confidence, respectively. White hexagons are zones of not significant clustering.

Figure 2: Dark and bright red hexagons showing areas with an intense clustering of high values with 99% and 95% confidence, respectively.



4 Disccusion

KDE is able to tell us where clusters in our dataset are, however, we could not tell whether these clusters are statistically significant or not. Moreover, we see that by changing the search radius the outputs came different. For our analysis, we do not consider the cell size nor the scale of the outputs. All these parameters are left out or set up by default calculations. The KDE outputs leave space for additional calculations and further discussions. Therefore, we agree that these analyses give useful outputs upon which we can make good assumptions. Nevertheless, we want to minimize the subjectivity from the analysis and make our decisions and conclusions more reliably and confidently.

In comparison to KDE, hot spot analysis is able to tell us both where the clusters in our dataset are and how significant are they. On our output (Figure 5), the dark red hexagon bins signify areas where there is intense clustering of high values with 99 percent confidence. These are areas where there are high numbers of crimes occurring, and as such require special attention from crime authorities. What we find interesting is that there are no statistically significant cold spots (areas of clustering of low crime counts).

5 Conclusion and Outlook

In case of KDE, it is difficult to give a general suggestion on parameter settings, as they are dependent on user requirements. Information represented by the resulting density surface depends on the choice of the kernel bandwidth and the output grid size. It is therefore necessary to experiment with these parameters to acquire map suitable for the user needs. The hotspots resulting from the KDE map are not statistically significant, and different cell sizes and search bands may obviously affect the results. In such circumstances, users should be vigilant with the area of treatment, the study area, and the study case.

On the other side, with hot spot analysis we are able to estimate density distribution of events at the local level, and identify statistically significant hot spots in our dataset. Considering our case study objective, we suggest that KDE should be used in conjunction with hot spot analysis to increase efficiently and efficacy in results interpretation.

Our further work should focus on enhancing our current methods and identifying new ones for understanding and visualizing how specific (spatial) phenomenon behaves.

References

Ahola, T., Virrantaus, K., Krisp, J.-M., and Hunter, G.-J. (2007) A spatiotemporal population model to support risk assessment and damage analysis for decision-making. *International Journal of Geographical Information Science*, 21(8):935 – 953.

Anderson, T.-K. (2009) Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention*, 41(3):359–364.

Anselin, L. (1995) Local Indicators of Spatial Association—LISA. In *Geographical Analysis* 27(2):93–115.

Bivand, R.-S., Pebesma, E.-J., Gómez-Rubio, V. (2008) Spatial Point Pattern Analysis. In: *Applied Spatial Data Analysis with R. Use R!*, pages 155-190. Springer, New York.

Bruce, C.-W. and Smith, S.-C. (2011) Spatial Statistics in Crime Analysis. In *International Association of Crime Analysts.* Overland Park, Kansas.

Chainey, S. and Ratcliffe, J. (2005) Identifying Crime Hotspots. In *GIS and Crime Mapping*, pages 145–182. Chichester, UK: John Wiley & Sons.

Cressie, N.A.-C. (2015) Spatial Point Patterns. In *Statistics for Spatial Data*, pages 575–723. New Jersey, USA: John Wiley & Sons.

Getis, A., and Ord, J.-K. (1992) The analysis of spatial association by use of distance statistics. In *Geographical Analysis* 24(3):189-206.

Getis, A., and Ord., J.-K. (1996) Local spatial statistics: an overview. In Longley, P. and Batty, M. (eds) *Spatial analysis: Modelling in a GIS environment*, pages 261-282. New York, USA: John Wiley & Sons.

Gilmartin, P., and Shelton, E. (1989) Choropleth maps on high resolution CRTs: the effects of number of classes and hue on communication. *Cartographica*, 26(2):40-52.

Goodwin, G.-C., Schoby, J. and Council, W. (2014) A Hot Spot Analysis of Teenage Crashes: An Assessment of Crashes in Houston, Texas. TX: Texas Southern University.

Jenks, G.-F. (1967) The Data Model Concept in Statistical Mapping. *International Yearbook of Cartography* (7):186-190.

Krisp J.-M., Virrantaus K., Jolma A. (2005) Using explorative spatial analysis methods in a GIS to improve fire and rescue services. In Oosterom P., Zlatanova S., Fendel E.M. (Eds.), *Geo-information for Disaster Management*, pages 1282-1296. Springer, Berlin, Heidelberg.

Krisp J.-M., Peters S., Murphy C.-E., and Fan, H. (2009) Visual Bandwidth Selection for kernel Density Maps *Photogrammetrie Fernerkundung Geoinformation*, (5): 441–450.

Krisp J.-M. and Špatenková, O. (2009) Kernel density estimations and their application in visualizing mission density for fire & rescue services. *Proceedings on Cartography and Geoinformatics for Early Warning and Emergency Management: Towards Better Solutions*. Prague, Czech Republic.

Kuo, P.-F., Lord, D. and Walden, T.-D. (2013) Using geographical information systems to organize police patrol routes effectively by grouping hotspots of crash and crime data. *Journal of Transport Geography*, (30): 138–148.

Levine N. (2017) CrimeStat: A Spatial Statistical Program for the Analysis of Crime Incidents. In: Shekhar S., Xiong H., Zhou X. (eds) *Encyclopedia of GIS*. Springer, Cham.

Mburu, L. and Zipf, A. (2014) A Spatial Approach to Surveying Crime-problematic Areas at the Street Level. In Proceedings of the Agile'2014 International Conference on Geographic Information Science. Castellón, Spain.

O'Sulivan D., and Unwin D.-J. (2003) Point Pattern analysis. In *Geographic Information Analysis*, pages 121-154. New Jersey, USA: John Wiley & Sons.

Prasannakumar, V., Vijith, H., Charutha, R, and Geetha, N. (2011) Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment. In *Procedia - Social and Behavioral Sciences,* (21):317–325.

Ratcliffe, J. (2010) Crime mapping: spatial and temporal challenges. In *Handbook of quantitative criminology*. 5-24. New York, NY: Springer.

Rey, S.-J., Mack, E.-A., and Koschinsky, J. (2011) Exploratory Space-time analysis of Burglary Patterns *Journal of Quantitative Criminology*, (28):509-531.

Silverman, B.-W. (1986) Density Estimation for Statistics and Data Analysis. London- New York, Chapman and Hall.

Smith, M.-J., Goodchild, M.-F., and Longley, P.-A. (2015) Geospatial analysis: a Comprehensive Guide to Principles, Techniques and Software Tools. The Winchelsea Press, Winchelsea, UK.

Špatenková, O. and Stein, A. (2010) Identifying factors of influence in the spatial distribution of domestic fires. In *International Journal of Geographical Information Science*. Taylor & Francis, 24(6):841–858.