ISSN: 1533 - 9211

DOI: 10.5281/zenodo.6876638

PREDICTING CRIME HOTSPOTS BY COMPARING MACHINE LEARNING ALGORITHMS

P. Suneel Kumar
Professor,Department of Electronics and Communication, Sridevi Women's Engineering College, Hyderabad, India. psunilkumar.ece@gmail.com

V. Preethi Chowdary
U.G Student,Department of Electronics and Communication, Sridevi Women's Engineering College, Hyderabad, India.

Ch. Vyshnavi
U.G Student,Department of Electronics and Communication, Sridevi Women's Engineering College, Hyderabad, India.

M. Chandana
U.G Student,Department of Electronics and Communication, Sridevi Women's Engineering College, Hyderabad, India.

ABSTRACT

Objective:

Now-a-days crime is one of the biggest and dominating problem in our society and its prevention is an important task. Daily there are huge numbers of crimes that are being committed frequently. Occurrence of these types of crimes requires keeping track of all the crimes and maintaining a database for same which may be used for future reference. The present problems faced are maintaining of proper dataset of crimes occurred and analyzing this data to help in predicting and solving the crimes in the future. The objective of this project is to analyze dataset which consist of number of crimes and predicting the types of crimes which may occur in future depending upon various circumstances. The thesis aim is to visualize and explore the methodological approach in finding the spatial patterning of crime through a geographical information system as a means for future guidance within spatial crime analysis. Two spatial analysis are performed, the Optimized Hot Spot-analysis tool and the kernel density estimation analysis tool. The average Nearest Neighbor-model was applied to the data for further statistical accuracy. For this supervised classification problem, Decision tree, Gaussian Tree, Gaussian Naïve Bayes, KNN, logistic Regression. This approach involves predicting crimes, classifying, Pattern detection and visualization with effective tools and technologies. Usage of past crime data trends helps us to correspond factors which might help understanding the future scope of crimes. In this work, various visualizing techniques and machine learning algorithms are acquired for predicting the crime distribution over a particular area. In the beginning step, the raw datasets were processed and visualized based on the requirement.

Introduction

Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide.

Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist.

There is a need of technology through which the case solving could be faster. The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the accuracy. The K-Nearest Neighbour (KNN) classification and other algorithm will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in the country.

This work helps the law enforcement agencies to predict and detect crimes in Chicago with improved accuracy and thus reduces the crime rate. There has been tremendous increase in machine learning algorithms that have made crime prediction feasible based on past data. The aim of this project is to perform analysis and prediction of crimes in states using machine learning models. It focuses on creating a model that can help to detect the number of crimes by its type in a particular state.

In this project various machine learning models like K-NN, boosted decision trees will be used to predict crimes. Area Wise geographical analysis can be done to understand the pattern of crimes. Various visualization techniques and plots are used which can help law enforcement agencies to detect and predict crimes with higher accuracy. This will indirectly help reduce the rates of crimes and can help to improve securities in such required areas. Crimes can be predicted as the criminals are active and operate in their comfort zones. Once successful they try to replicate the crime under similar circumstances.

Related Work

Crime prediction is done on Chicago data set in which various machine learning models are applied. Comparison of different model's algorithms like KNN, Naïve Bayes, SVM, decision tree is done this paper. It is seen that prediction varies depending upon the dataset and features that have been selected. The prediction accuracy found in is 78% for KNN, 64% for Gaussian NB, 31% for SVC. Auto regressive integrated Moving average models were used to make machine learning algorithms to forecast crime trends in urban areas. One of the major problems in crimes is detecting and analyzing the pattern of crimes. Understanding datasets is also an important concept in this case. We surely want to accurately predict so that we don't waste our

DOI: 10.5281/zenodo.6876638

resources due to false signals. Also proposed a method for classifying the crime rate as high, medium or low. None of them has classified the type of crime that can occur and its probability of occurring. Analysis and prediction of crime is an important activity that can be developed using various techniques and processes. Lot of research work is done by various researchers in this particular domain. The existing work is limited to use the datasets to identify locations of crime.

De Bruin et.al. introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. Manish Gupta et.al. highlights the existing systems used by Indian police as e-governance initiatives and also proposes an interactive query-based interface as crime analysis tool which can assist police in their activities. He proposed interface which is used to extract useful information from the vast crime database maintained by National Crime Record Bureau (NCRB) and find crime hot spots using crime data mining techniques such as clustering etc. The effectiveness of the proposed interface has been illustrated on Indian crime records. Nazlin Mohamad Ali et al. discuss on a development of Visual Interactive Malaysia Crime News Retrieval System (I-JEN) and describe the approach, user studies and planned, the system architecture and future plan. Their main objectives were to construct crime-based event; investigate the use of crime based event in improving the classification and clustering; develop an interactive crime news retrieval system; visualize crime news in an effective and interactive way; integrate them into a usable and robust system and evaluate the usability and system performance and the study will contribute to the better understanding of the crime data consumption in the Malaysian context as well as the developed system with the visualization features to address crime data and the eventual goal of combating the crimes. Suta pat Thiprung Sri examines the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. The purpose of his study is to examine the use of clustering technology to automate fraud filtering during an audit. He used cluster analysis to help auditors focus their efforts when evaluating group life insurance claims. A. Malathi et al. look at the use of missing value and clustering algorithm for a data mining approach to help predict the crimes patterns and fast up the process of solving crime. Malathi. An et. al. used a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyze the city crime data from Police Department. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years. Dr. S. Santhosh Baboo and Malathi. A research work focused on developing a crime analysis tool for Indian scenario using different data mining techniques that can help law enforcement department to efficiently handle crime investigation. The proposed tool enables agencies to easily and economically clean, characterize and analyze crime data to identify actionable patterns and trends. Kadhim B. Swadhi Al-Janavi presents a proposed framework for the crime and criminal data analysis and detection using Decision tree Algorithms for data classification and Simple K Means algorithm for data clustering.

The paper tends to help specialists in discovering patterns and trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and identifying possible suspects. Aravindan Mahendiran et al. apply myriad of tools on crime data sets to mine for information that is hidden from human perception. With the help of state-of-the-art visualization techniques, we present the patterns discovered through our algorithms in a neat and intuitive way that enables law enforcement departments to channelize their resources accordingly. Suta pat Thiprung Sri examine the possibility of using clustering technology for auditing. Automating fraud filtering can be of great value to continuous audits. The objective of their study is to examine the use of cluster analysis as an alternative and innovative anomaly detection technique in the wire transfer system. K. Zakir Hussain et al. tried try to capture years of human experience into computer models via data mining and by designing a simulation model.

PROPOSED WORD FOR PREDICTING CRIME HOTSPOTS

Here in the proposed system, we use the random forest algorithm in order to get good results and better accuracy when compared to the other existing algorithms. We use random forest for accuracy. Random forest is a most known and powerful supervised machine learning algorithm capable of performing both classification, regression tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over fitting to their training set. The data sets considered are rainfall, perception, production, temperature to construct random forest, a collection of decision trees by considering two-third of the records in the datasets. These decision trees are applied on the remaining records for accurate classification. Accuracy score for random forest algorithm is 86.6%.

PREDICTION MODEL

In this paper, random forest algorithm, KNN algorithm, SVM algorithm and LSTM algorithm are used for prediction of crime hotspots. First, historical crime data alone are used as input to calibrate the models. Comparison would identify the most effective model. Second, built environment data such as road network density and poi are added to the predictive model as covariates, to see if prediction accuracy can be further improved.

A. KNN

KNN, also known as k-nearest neighbor, takes the feature vector of the instance as the input, calculates the distance between the training set and the new data feature value, and then selects the nearest K classification. If k = 1, the nearest neighbor class is the data to be tested. KNN's classification decision rule is majority voting or weighted voting based on distance. The majority of k neighboring training instances of the input instance determines the category of the input instance. K-Nearest Neighbour is one of the simplest Machine Learning Algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new caste/data and available cases and put the new case into the category that is mostly similar to the available categories. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

B. RANDOM FOREST

Random forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of the dataset. The random forest is a set of tree classifiers {h (x, βk), k = 1. . .}, in which the meta classifier h (x, βk) is an uncut regression tree constructed by CART algorithm; x is the input vector; βk is an independent random vector with the same distribution, and the output of the forest is obtained by voting. The randomness of random forest is reflected in two aspects: one is to randomly select the training sample set by using bagging algorithm; the other is to randomly select the split attribute set. Assuming that the training sample has M attributes in total, we specify an attribute number F ≤ M, in each internal node, randomly select F attributes from M attributes as the split attribute set, and take the best split mode of the f attributes Split the nodes. The multi decision tree is made up of random forest, and the final classification result is determined by the vote of tree classifier.

C. SVM

SVM stand for Support Vector Machine Algorithm. SVM is one of the most popular supervised learning algorithms, which is used for classification as well as regression problems. However, primarily, it is used for classification problems in Machine Learning. SVM, based on statistical learning theory, is a data mining method that can deal with many problems such as regression (time series analysis) and pattern recognition (classification problem, discriminant analysis) very successfully. The mechanism of SVM is to find a superior classification hyperplane that meets the classification requirements, so that the hyperplane can ensure the classification accuracy and can maximize the blank area on both sides of the hyperplane. SVM chooses the extreme points/vector that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as support vector machine. In theory, SVM can realize the optimal classification of linear separable data.

A. NB

NB stands for naïve bayes. Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes Theorem and used for solving classifications problems. It is mainly used in text classification that includes a high-dimensional training dataset. In the field of probability and statistics, Bayesian theory predicts the occurrence probability of an event based on the knowledge of the evidence of an event. In the field of machine learning, the naïve Bayes (NB) classifier is a classification method based on Bayesian theory and assuming that each feature is independent of each other. In abstract, NB classifier is based on conditional probability, to solve the probability that a given entity belongs to a certain class. Naïve Bayes Classifier is one of the simple and most effective classification algorithms which help in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

B. CNN

CNN stands for Convolutional neural network is a class of artificial neural network, most commonly applied to analyze visual imagery. CNN uses one-dimensional convolution for sequence prediction, which is the convolution sum of discrete sequences. To convolve the sequence, CNN first finds a sequence with a window size of kernel size, and perform convolution with the original sequence to obtain a new sequence expression. The convolutional network also includes a pooling operation, which is to filter the features extracted by the convolution to get the most useful characteristics. CNNs are relatively little pre-processing compared to other image classification algorithms. This means that the network learns to develop the filters (or kernels) through automated learning, whereas in traditional in traditional algorithms these filters are hand engineered.

C. LSTM

LSTM stands for Long short-term memory. LSTM is an artificial neural network used in the fields of artificial intelligence and deep learning. LSTM is a kind of deep neural network based on RNN. The core of LSTM is to add a special unit (memory module) to learn the current information and to extract the related information and rules between the data, so as to transfer the information. LSTM is more suitable for deep neural network calculation because of memory module to slow down information loss. Unlike standard feedforward neural networks, LSTM has feedback connections. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals an =d the three gates regulate the flow of information into and out of the cell.

EXPERIMENTAL AREA AND DATA VISUALIZATION ANALYSIS

A. EXPERIMENTAL AREA

The area XT selected in this paper is a town in a coastal megacity in Southeast China. The population density of this community is relatively large, with a total area of about 6.5 square kilometers, a total population of about 400000, and a household registration population of only 50000, suggesting that the overwhelming majority of the population domestic migrants or non-local population. The town consists of several large-scale city villages. The complex composition of built environment and population makes it a high crime area.

B. SELECTION OF CRIME TYPES

The crime of property in public places mainly refers to the crime that takes occupying the property ownership of others as the main purpose in public places. It mainly includes theft, robbery, snatching and other types of embezzlement crimes that completely obtain property against the will of others. It is of great practical significance to choose the public property crime in this town for the prediction of crime hotspots. Accurate crime prediction can help guide the deployment of the local police resources, changing from passive policing to active prevention

DOI: 10.5281/zenodo.6876638

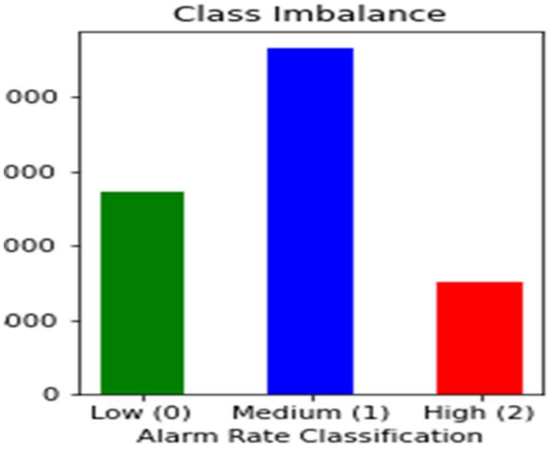
and control, thus improving local public security.

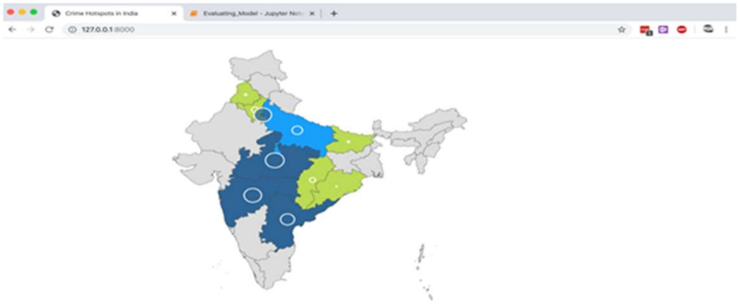
C. DATA VISUALIZATION ANALYSIS

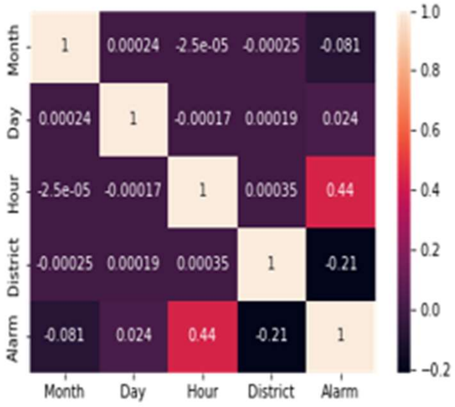
The historical crime data used in this paper comes from the police receiving data from 2015 to 2018 in the P-GIS database of the Public Security Bureau of the experimental district. The text coordinate information recorded in the database is extracted, and the case point data within the street range of the study area is extracted after it is located on the map of the study area. In order to meet the needs of practical police work, the spatial scale of crime hot spot prediction experiments should be as small as possible. According to the calculation formula of gridding processing study area of Griffith et al., the study area is divided into 150m ∗ 150m grids according to the investigation of actual police work and the data distribution of case points. Compared with grids with smaller spatial scales, grids divided by 150m will make case points more concentrated in certain grids and reduce the contingency of hotspot grids. Such a division will also reflect the mechanism and distribution of cases better and improve the prediction accuracy and preciseness of the crime hot spots. According to the investigation of the actual police work, 150m is the largest patrol area that a single police officer can cover in a time unit, which can better use the prediction results in crime prevention and control.

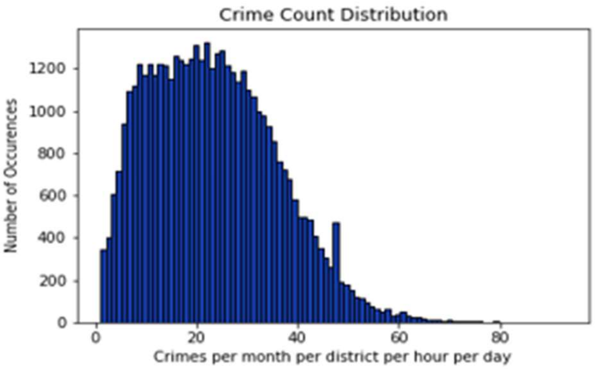
RESULTS AND DISCUSSION

This thesis primarily grew out of an interest for physical variations within space and its relation and impact on crime-trends and clustering. However, after initial analysis and research it became evident that such studies reached beyond the extent of both time and availability of data. That being said, it provided a framework for analyzing how and where crime cluster as well as how it changes on a temporal scale. It is clear from this thesis that the outdoor-crimes analyzed in this thesis have their highest peaks at and around large gathering-points, which is not all too surprising. If one was to move forward with this thesis it would be of great interest to change scope and identify those locations which show a relatively high intensity while geographically limited. If those were isolated one could move forward with the initial idea for this master’s thesis and measure possible correlation to physical factors or structural presence/absence. With limited number of locations and fewer physical and structural factors to measure it would increase the chance for gaining reliable results from a thesis. This master’s thesis preliminary focus was on the whole of Malmö and outdoor crimes for 2007. Something which was likely too hard to measure successfully. One issue to overcome remain however. If one work under the assumption that the production and spatiality of crime works under the routine activity theory or rational choice theory one must acknowledge that crime is inherently based on opportunistic behavior. In there the conflict lies. Weisburd, et., al., (2016:56) argue that the relationship between crime and opportunity is that of a non-linear relationship. If Weisburd's claims hold true, can one measure such correlation? It is possible that the limitation lies within the software used.









Using different machine learning algorithms, and using the past crime data, the occurrence of future crimes can be predicted. The future crimes can be predicted using all algorithms but the accurate one is preferred. So here we compare all the machine learning algorithms that are being used and choose the better one so that we get the accurate results and crime hotspots can be predicted properly so that proper safety measures can be taken accordingly. Here on using the past data the future crime hotspots are detected and shown in the map that in which areas which types of crimes are going to be taken place.

DOI: 10.5281/zenodo.6876638

CONCLUSION

With the help of machine learning technology, it has become very easy to find out relation and patterns among various data. The work in this project mainly revolves around predicting the type of crime which may occur if we know the location of where it has occurred. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation. The model predicts the type of crime with accuracy of 0.789. Data visualization helps in analysis of data set. The graphs include bar, pie, line and scatter graphs each having its own characteristics. We generated many graphs and found interesting statistics that helped in understanding Past crimes datasets can help in capturing the factors that can help in keeping society safe. The tool we have developed provides a framework for visualizing the crime networks and analyzing them by various machine learning algorithms using the Google Maps. The project helps the crime analysts to analyze these crime networks by means of various interactive visualizations. The interactive and visual feature applications will be helpful in reporting and discovering the crime patterns. Many classification models can be considered and compared in the Analysis. It is evident that law enforcing agencies can take a great advantage of using machine learning algorithms to fight against the crimes and saving humanity. For better results, we need to update data as early as possible by using current trends such as web and Apps.

REFERENCES

[1]Agnew, John, 1987, Space and Place, The SAGE Handbook of Geographical Knowledge, Agnew, John, Livingstone, N. David (eds.), Sage Publications, Thousand Oaks.

[2]Agnew, John, (1987), Place and politics: The geographical mediation of state and society, Boston MA, Allen and Unwin.

[3]Allan, George, (2003) A critique of using grounded theory as a research method, Portsmouth University, UK.

[4]Andresen, A. Martin, Farrell, Graham, (2015) The Criminal Act – The Role and Influence of Routine Activity Theory, Palgrave Macmillan, London.

[5]Blakemore, Michael, Longhorn, Roger, (2004) Ethics and GIS: The Practitioner's Dilemma – a position paper on GIS ethics, AGI 2004 Workshop on "GIS Ethics", London.

[6]Bohman, Helena, Gerell, Manne, Lundsten, Jonas, Tykesson, Mona, (2013) Stadens Bränder Del 2

[7] Fördjupning, Malmö, Publikationer I Urbana Studier.

[8]Ceccato, Vania, (2012) The Urban Fabric of Crime and Fear, Springer, London.

[9]Chainey, Spencer, Ratcliffe, Jerry, (2005) GIS and Crime Mapping, John Wiley & Sons, Ltd, West Sussex.

[10]Chainey, Spencer, (2013) Examining the influence of cell size and bandwidth size on kernel density estimation crime hot spot maps for predicting spatial patterns of crime, Bulletin of the Geographical Society of Liege, Vol:60, pp:7-19.

[11]Cloke, Paul, Cook, Ian, Crang, Philip, Goodwin, Mark, Painter, Joe, Philo, Chris, (2004) Practicing human geography, Sage Publications, Thousand Oaks.

DOI: 10.5281/zenodo.6876638

[12]Cohen, E. Lawrence, Felson, Marcus, (1979) Social Change and Crime Rate Trends: A Routine Activity Approach, American Sociological Revive, 44:4, pp:588-608.

[13]Cozens, P., 2008, Crime Prevention Through Environmental Design, Environmental Criminology and Crime Analysis, Wortley, R., Mazerolle, L. (eds.), pp: 153-177, Devon, UK.

[14]Dagens Nyheter, Rekordfå brott klaras upp efter Polisenos om organization, published 2016-08-30. [http://www.dn.se/nyheter/sverige/rekordfa-brott-klaras-upp-efter-polisens-omorganisation/, accessed 2017-05-29].

[15]Denzin, Norman, Lincoln, Yvonna, 2000, Handbook of Qualitative Research, Sage publications, Thousand Oaks.

[16]Eck, E. John, Weisburd, David, (1995) Crime and Place, Criminal Justice Press, Washington.

[17]Esri (2017), http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-
analysis-getis-ord-gi-spatial-stati.htm, accessed 2017-08-16].

[18]Farrell, Graham, Phillips, Coretta, Pease, Ken, Like Taking Candy: Why does Repeat Victimization Occur? The British Journal of Criminology, 1995, 35:3, pp:384-399, Oxford University Press.

[19]Glaser, G. Barney, Strauss, L. Anselm, (1967) The Discovery of Grounded Theory – Strategies for Qualitative Research, Transaction Publishers, London.

[20]International Association of Crime Analysts (IACA), Identifying High Crime Areas, White Paper, Overland Park.

[21]Ratcliffe, Jerry (2010), Crime Mapping: Spatial and Temporal Challenges, Piquero, R, Alex, Weisburd, David (eds.), Handbook of Quantitative Criminology, Springer, London.

[22]Knigge, Ladona, Cope, Meghan, (2004) Grounded visualization: integrating the analysis of qualitative and quantitative data through grounded theory and visualization, Environment and Planning A, 2006:38, pp:2021-2037.

[23]Massey, Dorian (1994), Space, place and gender, Polity Press, Minneapolis.

[24]V. Prasannakumar, H. Vijith, R. Charutha, N. Geetha, (2011) Saptios-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment, Procedia – Social and Behavioral Sciences, 2011, Vol: 21, pp:317-325.

[25]Schuurman, Nadine, 2000, Trouble in the heartland: GIS and its critics in the 1990s. Progress in human geography, 2000, 24:4, pp:569-590.

[26]Sherman, W. Lawrence, Gartin, R. Patrick, Buerger, E. Michael, (1989) Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place, Criminology, 27:1, pp:27-55.

[27]Steinberg, Lakshmi Sheila, Steinberg, J Steven, (2015), GIS Research Methods - Incorporating Spatial Perspectives, Esri Press, Redlands, California.

[28]Teichman, Doron, (2005) The Market for Criminal Justice: Federalism, Crime Control, and Jurisdictional Competition, Michigan Law Review, 2006, 103:7, pp:1831-1876.

[29]Thomas, Gary, James, David, (2006) Re-inventing grounded theory: some questions about theory, ground and discovery, British Educational Research Journal, 2006, 32:6, pp:767-795.

[30]Vilalta, J. Carlos, (2013) How Exactly Does Place Matter in Crime Analysis? Place, Space, and Spatial Heterogeneity, Journal of Criminal Justice Education, 2013, 24:3, pp:290-315.
[31]Weisburd, David, Lorraine, Green, Mazerolle, (2000), Crime and Disorder in Drug Hot Spots: Implications for Theory and Practice in Policing, Police Quarterly, 2000, 3:3, pp: 331-349

.