Crime Analysis and Best Locations To Live in Chicago

Sree Lakshmi Addepalli
NYU Courant
New York, United States

Divya Juneja
NYU Courant
New York, United States

Sree Gowri Addepalli
NYU Courant
New York, United States

***Abstract*—**

This paper describes an approach to build a system that would suggest best places to live in Chicago. Suggestions of best locations are made to a user in the neighborhood of a given address by leveraging the datasets of crime, sex offenders, food inspection, and affordable rental housing. KNN has been used for this purpose. Another application is built for use by the police department to check if the arrest would be made for a crime based on its crime description in communities of Chicago. Crime, public health statistics, socioeconomic factors, and community areas datasets are used for this system and Logistic Regression is used for predicting the occurrence of arrest. Both machine learning algorithms have been implemented in Spark using MlLib library.

I. Introduction

In this paper, we try to analyze the pattern using crime-related data and predict future arrests and best locations to live in the communities of Chicago. The crime inference across locations would help prevent victimization and other forms of crime exposures. Chicago is selected as target of study because according to the FBI crime statistics for 2013, it has more homicides and non negligent manslaughter rates (15.2) per 100,000 residents than New York (4.0) and Los Angeles (6.5) and has experienced no decline in the past decade compared to the latter two urban areas [1]. One of our application is made for user searching for housing in Chicago. User enters the address which could be the place he/ she will be working at and the gender in our system and using the classification algorithm, i.e., K-Nearest Neighbor (KNN), we suggest the best locations to live in the neighborhood of given address. Various socioeconomic factors such as poverty, illiteracy and other factors such as life expectancy, food inspection, etc., influence crime occurrence in the area and our ML model statistically analyzes data redundancy and pattern. In case the user is female, sex offender dataset is also taken into consideration. Thus, based on crime occurrence probability in the area, these suggestions are made.

In our another application, the report of crime is entered by user and based on crime dataset, it is predicted if the arrest will happen for that crime. This arrest forecast is done by applying Logistic Regression over the given datasets. This application can be used by police department to analyze the arrest pattern and improve their police patrols. Machine learning models of both the applications are implemented using MlLib library of Spark.

Data collection, classification, pattern identification, prediction, and visualization are usually involved in machine-learning-based crime analysis. The rest of this paper is describes these in detail as follows: Section III of this paper provides a brief survey of previous work done on the topic of crime analysis. In sections IV and V, the data-analysis, application architecture, machine-learning methodology, and UI/ visualization methods used in this work are explained. The results are presented and compared in VI. Experiments and conclusions are presented in sections VII and VIII respectively.

II. Motivation

Being an international student, when we joined the New York University, our major concern was related to finding safe housing. Crime was major factor that influenced our choices. The idea of developing this application is highly influenced by this fact. And in urban areas like Chicago, crime is one of the biggest social problems. Reports of direct and indirect victimization and exposures to crime remain very high [22]. Hence, we try to analyze crime probability across communities which helps user to identify the best locations to live in the neighborhood of his given address. Motivation of arrest prediction is that it will contribute to effective police patrols. Depending on the location of known districts where crime is rampant or based on the empirical knowledge of police and arrest probability, these patrols have been undertaken. But this problem can be solved by the forecast of arrest occurrences, made by analyzing and modeling previous crime dataset.

III. Related Work

Problems regarding crime control have been researched a lot in past and different crime-prediction algorithms have been proposed for the same. The accuracy of prediction depends on the attributes selected and the dataset used as a reference. In [2], deep learning is used to predict crime occurrence from multimodal data. The data in this paper is collected from various online databases of crime statistics, demographic and meteorological data, and images in Chicago, Illinois. Crime rate inference at the neighborhood level is done in [3] using two types of urban data, i.e., Point-Of-Interest and taxi flow. These datasets show significantly improved performance in crime rate inference compared to using traditional features. The systematic comparison between various regression models is done in this paper. Bogomolov et. al. [11] use human behavioral data derived from mobile network and demographic sources, together with open crime data to predict crime

hotspots. Various classifiers are compared by them and random forests are found to have the best prediction performance. On the other hand, crime counts are predicted as a function of yard characteristics and surrounding tree canopy in [4]. Ordinary least squares (OLS), spatial error regression (SER), and Poisson regression are utilized as three statistical approaches.

Analysis of Vancouver crime data for the last 15 years is done in [5] using two machine learning predictive models, i.e., K-Nearest Neighbor and boosted decision tree. Crime type is chosen as the target to train the algorithm. Whereas in [6], the paper focuses on applying different classification algorithms on the real crime data and comparing the accuracy of their results in predicting the crime category attribute. Decision Tree and Naïve Bayesian are used to perform classification on the dataset. The dataset 'Crime and Communities' is acquired from UCI machine learning repository website [8]. Decision Tree is found to perform better than Naïve Bayesian.

A crime incidence-scanning algorithm was applied to train Artificial Neural Network (ANN) in [12] to predict the crime hotspots in Bangladesh. To analyze drug-related crime data in Taiwan and predict emerging hotspots, a data-driven machine-learning algorithm based on broken-window theory, spatial analysis, and visualization techniques was used in [13]. To model the dependency between the offense data and environmental factors such as the demographic characteristics and the spatial location in the state of New South Wales (NSW), Australia, a fully-probabilistic algorithm based on Bayesian approach was applied in [14]. In order to forecast crime trends in urban areas, an approach based on Auto-Regressive Integrated Moving Average model (ARIMA) was utilized in [15] to design a reliable predictive model. An approach to detect patterns of crime is defined in [7]. Pattern is observed to determine which ones may have been committed by the same individual/ individuals. A pattern detection algorithm called Series Finder is proposed in order to achieve this, that grows a pattern of discovered crimes from within a database, starting from a "seed" of a few crimes. Overall, many classic data mining techniques have been successful for crime analysis generally, such as association rule mining [17–20], classification [21], and clustering [16].

IV. Application Design

6 datasets and one shape file data are used as big data for our applications and they are stored in HDFS for cleaning and profiling using Spark. MlLib is used for running machine learning algorithms and models like KNN and logistic regression are used to analyze the influence of various factors on the crime rate near the given area by user. We analyze the crime rate merging various datasets and taking various factors into consideration and suggest best places to live in the neighborhood areas in Chicago. We also take input from user to predict arrest occurrence in other application for which we apply logistic regression. Input is taken and output is produced in these applications using command line. In one application, 12 inputs are taken from user to predict occurrence of crime and output is given categorically as 0 or 1. Correlation between various datasets is removed as part of feature selection. In other application, we have 2 inputs as address and gender of user and output is just the address of best locations to live in the neighborhood of address provided. Project diagrams are shown in Fig 1 and 2. Various analytics are done using these datasets and visualization of analytics is done by graphs using Tableau as shown in Fig 3.

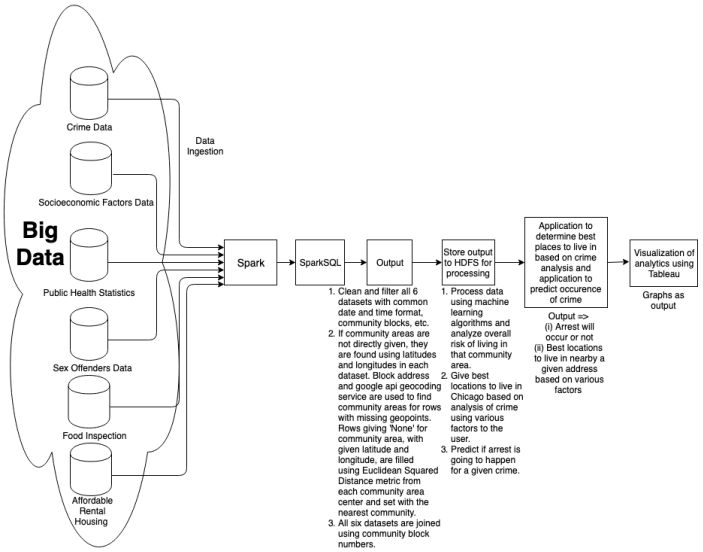

Fig 1: Project Design Diagram

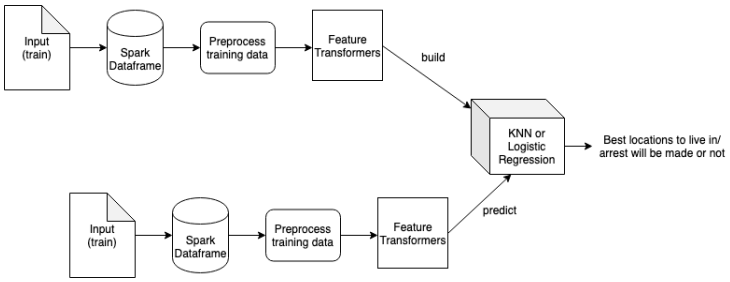

Fig 2: Project Architecture

5 analytics is done on the datasets. One of the analytics includes, how unemployment of people aged 16+ and per capita income, both extracted from socioeconomic factors dataset, are related to crime rate. Here, we get output as community area 25, which is Austin. As it can be validated from the website: https://www.niche.com/places-to-live/n/austin-chicago-il/, crime rate is high in Austin with high poverty and unemployment.

V. Datasets

Six datasets used are that of crime, socioeconomic factors, sex offenders, food inspection, affordable rental housing, and public health statistics. Shape file of community areas has also

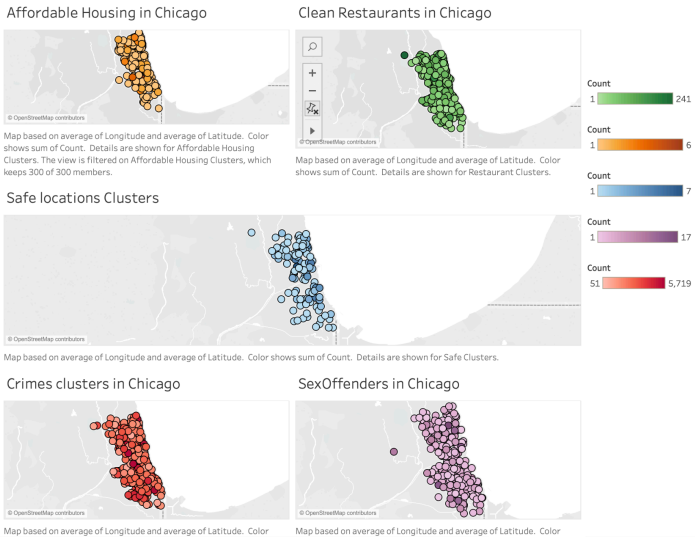

Fig 3: Example graph for visualization made using Tableau

been used to join the datasets. The data can be found on https://data.cityofchicago.org/. For pre-processing, if community areas are not directly given, they are found using latitudes and longitudes in each dataset. Block address and google api geocoding service are used to find community areas in case the rows are missing geopoints. Even then, if the rows give 'None' for community area, with given latitude and longitude, we use Euclidean Squared Distance metric from each community area center and set it with the nearest community. Datasets are extracted from 01/01/2010 till 12/31/2011. Date and time formats are changed, columns are dropped, missing values are filled using average value as part of profiling and cleaning of datasets.

First dataset, i.e., crimes dataset collected in Chicago, has detailed information about time, location (i.e., latitude and longitude), and types of crime. The term crime count refers to number of crime incidents in a region (i.e., community area) in a year. The community area is used as our geographical unit of study, since it is well-defined, historically recognized and stable over time. In total, there are 77 community areas in Chicago. We use this crime dataset using factors like number of arrests made and the total crime cases reported. This dataset is used in both the prediction of arrest application and application of suggesting places to live in. This dataset is 1.3 GB in size.

Crime dataset can be found on: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Second dataset, i.e., socioeconomic dataset, contains a selection of six socioeconomic indicators of public health significance and a "hardship index," by Chicago community area. Thus, this dataset describes unemployment, percentage of households below poverty, per capita income, etc. in the particular area. We use this dataset for crime prediction given youth population and illiteracy rate in the area. This dataset is used in prediction of arrest application and is 4 KB in size.

Variable Name	Variable type	Represents
Year	Integer	Case number which acts as unique ID of crime
Month	Integer	Date on which crime happened
Day	Integer	Community block where crime happened
Time	String	Time of the crime occurence
IUCR	String	Unique Crime case ID
Primary Type	String	Type of crime
Description	String	Description of crime
Location Description	String	Location where crime happened
Community Area	String	Community area where crime happened
Arrest	Boolean	If arrest if made for the crime
FBI Code	String	FBI code for the case
Latitude	Double	Latitude of the place where crime happened
Longitude	Double	Longitude of the place where crime happened

Table 1: Schema of final crime dataset

Socioeconomic factors dataset can be found on: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

Variable Name	Variable type	Represents
Community Area	String	Community area for which analysis is done
Community Area Name	String	Community area name for which analysis is done
Percentage of housing crowded	String	Percentage of crowded houses in given community
Percentage of households below poverty	String	Percentage of households below poverty in given community
Percent Aged 16+ Unemployed	String	Percentage pf people aged above 16 that are unemployed
Percent Aged 25+ without high school diploma	String	Percentage pf people aged above 25 that donot have high school degree
Percent aged under 18 or over 64	String	Percentage pf people aged under 18 or above 64
Per capita income	String	Per capita income in that community
Hardship index	String	Hardship index considering all the factors given to that community

Table 2: Schema of final socioeconomic factors dataset

Variable Name	Variable type	Represents
Block	String	Community area block for which analysis is done
Victim minor	String	If the victims are minor
Community Area	String	Community area for which analysis is done
Community Area Name	String	Community area name for which analysis is done
Latitude	Double	Latitude of the place where crime happened
Longitude	Double	Longitude of the place where crime happened

Table 3: Schema of final sex offenders dataset

Third dataset is Sex Offenders dataset. This dataset is given included only if user is a female. It contains indicators of age of victims by Chicago community area. This dataset is used in the application of suggesting places to live in and is 82 KB in size.

Sex Offenders dataset can be found on: https://data.cityofchicago.org/Public-Safety/Sex-Offenders/vc9r-bqvy

Fourth dataset is Public Health Statistics dataset. It contains a selection of public health indicators of age of victims by Chicago community area. This dataset is used in application for prediction of arrest and is 8 KB in size.

Public Health Statistics dataset can be found on: https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

Variable Name	Variable type	Represents
Community Area	String	Community area for which analysis is done
Community Area Name	String	Community area name for which analysis is done
Birth Rate	Double	Birth rate in the area
Low Birth Weight	Double	Birth weight low
Prenatal Care Beginning in First Trimester	Double	Prenatal Care
Preterm Births	Double	Preterm births happening
Teen Birth Rate	Double	Teen birth rate
Assault	Double	Assault in the area
Breast Cancer in Females	Double	Breast Cancer in Females
Cancer	Double	Cancer
Colorectal Cancer	Double	Colorectal Cancer
Diabetes related	Double	Diabetes
Firearm Related	Double	Firearm
Infant Mortality Rate	Double	Infant Mortality Rate
Lung Cancer	Double	Lung Cancer
Prostate Cancer in Males	Double	Prostate Cancer in Males
Stroke	Double	Stroke
Childhood Blood Lead Level Screening	Double	Childhood Blood Lead Level Screening
Childhood Lead Poisoning	Double	Childhood Lead Poisoning
Gonorrhea in Female	Double	Gonorrhea in Female
Gonorrhea in Males	Double	Gonorrhea in Males
Tuberculosis	Double	Tuberculosis

Table 4: Schema of Public Health Statistics dataset

Fifth dataset is Food Inspection dataset. It contains a selection of public health indicators of age of victims by Chicago community area. This dataset is used in the application of suggesting places to live in and is 218.5 MB in size.

Food Inspection dataset can be found on: https://catalog.data.gov/dataset/food-inspections-8cc79

Variable Name	Variable type	Represents
Risk	String	Overall risk factor of restaurant
Latitude	Double	Latitude of the place where restaurant is there
Longitude	Double	Longitude of the place where restaurant is there

Table 5: Schema of final Food Inspection dataset

Sixth dataset is Affordable Rental Housing dataset. It contains a selection of public health indicators of age of victims by Chicago community area. This dataset is used in the application of suggesting places to live in is 86 KB in size.

Affordable Rental Housing dataset can be found on: https://catalog.data.gov/dataset/affordable-rental-housing-developments-ef5c2

Variable Name	Variable type	Represents
Location	String	Location where affordable housing is available
Latitude	Double	Latitude of the place where affordable housing is available
Longitude	Double	Longitude of the place where affordable housing is available

Table 6: Schema of final Affordable Rental Housing dataset

Seventh is a shape file used for Community Areas. It provides the neighborhoods names, given a longitude and a latitude. It can return either None or the corresponding communities names. This dataset is basically used to join all other datasets.

Community areas shape file can be found on: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

VI. REMEDIATION

As response to the 12 input values given by user in first application, we predict if the arrest will happen for the given crime. And for the second application, in response to two inputs given by user, i.e., address and gender of user, we output the locations in neighborhood of given address. These are considered to be safe and evaluated with respect to low crime, sex offenders, better housing affordability and low risk food conditions. Our application is run by intervention by user. When the user gives input on command line, output is generated on the same terminal.

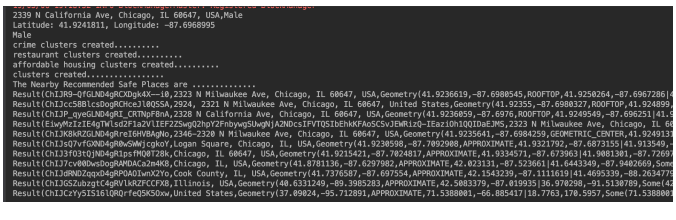

Fig 4: Command line UI for suggestions of safe places output for female

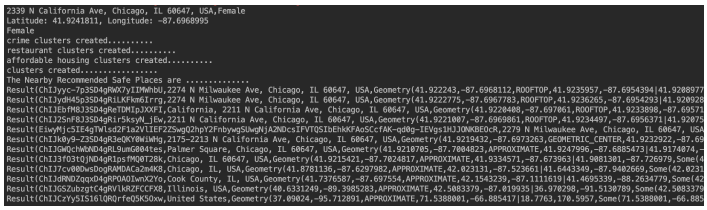

Fig 5: Command line UI for suggestions of safe places input

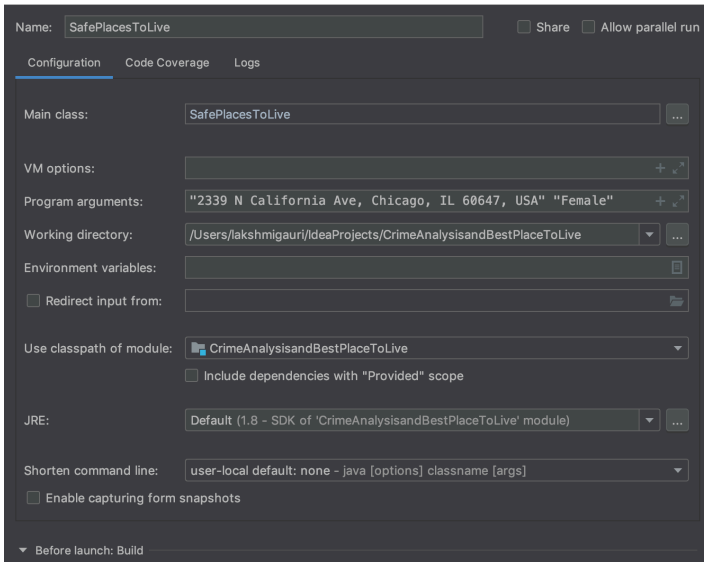

Fig 6: Command line UI for suggestions of safe places output for male

VII. EXPERIMENTS

(In this section, you can describe: Your experimental setup, problems with: data, performance, tools, platforms, etc. Discuss your experiments, describe what you learned. Discuss limitations of the application. Discuss what you would do to expand it given time - how would you improve it, etc.)

VIII. CONCLUSION

(One paragraph about the value, results, usefulness of your application.)

IX. FUTURE WORK

(Discuss possible future work for extending this project.)

ACKNOWLEDGMENT

We thank government of Chicago city who has provided us the datasets on their website. We are grateful to Matthew Cross whose geocode mapping is used by us (https://github.com/KoddiDev/geocoder/blob/master/README.md). We also thank Craig Booth whose code we have used for conversion of latitude and longitude to community area (https://github.com/craigmbooth/chicago_neighborhood_finder).

REFERENCES

1. C. Tribune, A tale of 3 cities: La and nyc outpace chicago in curbing violence, 2015. http://www.chicagotribune.com/news/ ct-violence-chicago-new-york-los-angeles-met-20150918-story.html
2. Hyeon-Woo Kang, Hang-Bong Kang. Prediction of crime occurrence from multimodal data using deep learning, Jan 2017. https://pdfs.semanticscholar.org/1275/f2fcdf911abf013301173d2d06ad35c2b739.pdf
3. Hongjian Wang, Huaxiu Yao, Daniel Kifer, Corina Graif, Zhenhui Li. Non-Stationary Model for Crime Rate Inference Using Modern Urban Data, Apr 2017. https://faculty.ist.psu.edu/jessieli/Publications/2017-TBD-GWR.pdf
4. Steven D. Levitt and Thomas J. Miles. Economic Contributions To The Understanding Of Crime, Aug 2006. https://www.researchgate.net/publication/321398204_Using_sentiment_analysis_to_predict_interday_Bitcoin_price_movements
5. Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime Analysis Through Machine Learning, 2018. https://www.researchgate.net/publication/330475412_Crime_Analysis_Through_Machine_Learning
6. R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahy, and N. Khanahmadliravi. An experimental study of classification algorithms for crime prediction. Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219-4225, Mar. 2013. http://www.indjst.org/index.php/indjst/article/view/31230
7. T. Wang, C. Rudin, D. Wagner, and R. Sevieri. Learning to detect patterns of crime. Springer, 2013. https://www.researchgate.net/publication/279841124_Learning_to_Detect_Patterns_of_Crime
8. UCI Machine Learning Repository (2012). Available from: http://archive.ics.uci.edu/ml/datasets.html
9. H. Adel, M. Salheen, and R. Mahmoud. Crime in relation to urban design. Case study: the greater Cairo region. Ain Shams Eng. J., vol. 7, no. 3, pp. 925-938, 2016. https://www.sciencedirect.com/science/article/pii/S2090447915001379
10. T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.
11. A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. Once upon a crime: towards crime prediction from demographics and mobile data. ACM. Sep. 2014. https://arxiv.org/abs/1409.2983
12. N. Mahmud, K. Ibn Zinnah, Y. Ar Rahman, and N. Ahmed. CRIMECAST: a crime prediction and strategy direction service. IEEE 19th Intl. Conf. on Comput. and Inform. Technol., Dhaka, Bangladesh,.Dec 2016. https://ieeexplore.ieee.org/document/7860234
13. Y. L. Lin, L. C. Yu, and T. Y. Chen. Using machine learning to assist crime prevention. IEEE 6th Intl. Congr. on Advanced Appl. Inform. (IIAIAAI), Hamamatsu, Japan. Jul 2017. https://ieeexplore.ieee.org/document/8113405
14. R. Marchant, S. Haan, G. Clancey, and S. Cripps. Applying machine learning to criminology: semi−parametric spatial−demographic Bayesian regression. Security Inform, vol. 7, no. 1. Dec. 2018. https://link.springer.com/article/10.1186/s13388-018-0030-x
15. E. Cesario, C. Catlett, and D. Talia. Forecasting crimes using autoregressive models. IEEE 14th Intl. Conf. on Dependable, Auton. and Secure Comput., Auckland, New Zealand. Aug 2016. https://ieeexplore.ieee.org/document/7588936
16. Nath, S.V. Crime pattern detection using data mining. In: Proceedings of Web Intelligence and Intelligent Agent Technology Workshops. Dec 2006 41–44 https://ieeexplore.ieee.org/document/4053200
17. Brown, D.E., Hagen, S. Data association methods with applications to law enforcement. Decision Support Systems 34(4) (2003) 369–378. Mar 2003. https://dl.acm.org/citation.cfm?id=640407
18. Lin, S., Brown, D.E. An outlier-based data association method for linking criminal incidents. In: Proceedings of the Third SIAM International Conference on Data Mining. 2003. https://epubs.siam.org/doi/abs/10.1137/1.9781611972733.39

19. Ng, V., Chan, S., Lau, D., Ying, C.M. Incremental mining for temporal association rules for crime pattern discoveries. In: Proceedings of the 18th Australasian Database Conference. Volume 63. 123–132. Jan 2007. https://www.researchgate.net/publication/221152620_Incremental_Mining_for_Temporal_Association_Rules_for_Crime_Pattern_Discoveries
20. Buczak, A.L., Gifford, C.M. Fuzzy association rule mining for community crime pattern discovery. In: ACM SIGKDD Workshop on Intelligence and Security Informatics. Jul 2010. https://dl.acm.org/citation.cfm?id=1938608
21. Wang, G., Chen, H., Atabakhsh, H. Automatically detecting deceptive criminal identities. Mar 2004. https://dl.acm.org/citation.cfm?id=971617.971618
22. J. Truman and L. Langton. Criminal victimization, 2014. Bureau of Justice Statistics (BJS), U.S. Department of Justice, Tech. Rep. NCJ 248973. Sep 2015. https://www.bjs.gov/content/pub/pdf/cv14.pdf