

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/335361962>

Crime Analysis in Chicago City

Conference Paper · June 2019

DOI: 10.1109/IACS.2019.8809142

CITATIONS

8

READS

6,238

3 authors, including:



Ayidh Alqahtani

University of Maryland, Baltimore County

3 PUBLICATIONS 10 CITATIONS

SEE PROFILE

CRIME ANALYSIS IN CHICAGO CITY

Ayidh alqahtani
Information Systems department
University of Maryland Baltimore
County
Baltimore ,United States
ayidh1@umbc.edu

Ajwani Garima
Information Systems department
University of Maryland Baltimore
County
Baltimore ,United States
garimal@umbc.edu

Ahmad Alaaaid
CIS Department
Jordan University of Science and
technology
Irbid, Jordan
aiaiad@just.edu.jo

Abstract—Security has always been one of the most significant concerns. Government and security agencies are working hard to prevent crimes and protect their people. However, challenge of dealing with large amount of data has become a major issue for all organizations. Therefore, a crime information system that is able to process large amount of data in a short period of time is needed for investigators to know crimes hotspots, crime patterns and to predict future ones. This paper provides design of Crime Data Information System. Data preprocessing is done in Crime Database and two approaches for crime analysis is performed. These two approaches are compared, and results are confirmed with ground truth.

Keywords—component, data mining, crime, preprocessing, clustering, spatial clustering, Data analytics, analysis.

I. INTRODUCTION

There is a rapid increase in crime in almost every country. There is a strong need to identify crime patterns and analyze different areas of crime. Security agencies all over countries are working hard to reduce these crimes, however, size of crime information is increasing rapidly, and it becomes difficult to manage such a huge amount of data and to keep record of crimes that are geographically widespread and at different time period. Thus, it is very necessary to have a crime information system which can process large amount of data in short period of time. Data mining using clustering, classification, association mining is one of the most effective ways to explore, analyze, detect patterns and predict future crimes in huge amount of data. Traditional approaches did not provide any kind of central crime database to dig and find out relations. They used paper based methods to keep the record of crime which made extremely difficult for security agencies to find out patterns and better use of their resources. With the increasing use of the computerized systems to track crimes, computer data analysts have started helping the security officers to speed up of solving crimes. However, data analysis was done in a superficial way. Data was not being collected in scientific manner, so by and large data was sporadic. They did not have any formal process or technique of data collection, cleaning and extracting knowledge out of it. Therefore, data mining has been known as a powerful tool that aids investigators to explore and work on large amount of data. Also, various technologies like Satellite Images, Cellular Phones, Sensor networks, and GPS devices helps in collecting data related to space and time. This helps in Spatial and spatial-temporal mining of data. It helps identifying non-trivial information which traditional systems are unable to discover. Motivation: As we know there is a rapid increase in crime, if it is not controlled, it can have adverse effect or can create hostile environment for

everyone. It must be a top priority to save the nation from criminals. Police and other security departments have number of resources to help the citizens of country. However, placing right resource at right time and right place is the key factor to overcome this problem. News and records indicate that there have been a large number of crimes in Chicago city. It is called as “murder capital” or “crime capital” of the U.S in 2012. Chicago had more murders in 2012 than any other city in the country. There were 500 murders in Chicago last year; the FBI said which were highest in entire nation. [1] Also, there has been a report by New York Times which states that Rate of Killings Rises 38 Percent in Chicago in 2012.[2]Therefore, there is a requirement to make more informed decisions and analyze crimes occurring in different region of Chicago city. By the help of data mining we can detect different patterns and therefore benefit the residents of Chicago city by employing the correct resources at correct time and place. This will help in utilizing maximum efficiency of security departments and help prevent crime in the city.

This research paper is about analyzing historical data related to different crimes occurring in different region of Chicago city. In this we retrieved data from Chicago Data Portal and designed a crime data information system. Next step was data preprocessing; processed raw data into uniform format. In our dataset, type of crime, time and location of crimes were important attributes. More specifically, we will be using two approaches. First one is to do clustering using K-means algorithm and the second one is to do Spatial Mining to find out hotspots of crime. For clustering using K-means we will be using WEKA tool and Euclidean distance as metric. Through this we will build different clusters and identify places of crime. Clustering helps us to find similar kinds of crime in the given geography of interest. Such clusters are also useful in identifying a crime pattern. The densely populated group of crime is used to visually locate the ‘hot-spots’ of crime. This will help in the deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources. Hot-spots detection will be done by using SatScan software. Data is collected from Chicago Data Portal and will be analyzing 2 years of crime data of Chicago from 2010-2012.

II. RELATED WORK:

“Crime Pattern Detection Using Data Mining” by Shyam Varan Nath [3] proposed idea of crime detection and said clustering algorithm using data mining can help to detect the crime patterns and speed up the process of solving crime by using K-means clustering algorithm. This paper applied clustering to real crime data from a sheriff’s office and

validated results. Author of this paper identified significant attributes; using expert based semi-supervised learning method and developed the scheme for weighting the significant attributes. This allows placing different weights on different attributes dynamically based on the crime types being clustered. This also allows weighing the categorical attributes. Limitation in this paper was that there was no prediction of the crime hot-spots that will help the police utilizing their resource for any given window of time.

“Detecting and Mapping Crime Hot Spots Based on Improved Attribute Oriented Induce Clustering” [4] discusses about detecting high-crime-density areas. It says most useful method for detecting crime hotspots is the spatial clustering. Crime Data includes many various crime events such as event time, event class, event spatial information and event object. These data have many attributes at different levels. So the attribute oriented induce method is chosen to deal with these data. This paper presents an improved attribute oriented induce method and algorithm related to crime hot spots detecting and a simple mapping method is depicted. Limitation and future work for this paper is further design and optimization of taxonomy. Taxonomy of attribute is the key related to the accuracy of the results. The mapping method in this paper is relatively simple. Improving the mapping method is another important work later.

“Detecting and Mapping Crime Hot Spots Based on Improved Attribute Oriented Induce Clustering” [4] discusses about detecting high-crime-density areas. It says most useful method for detecting crime hotspots is the spatial clustering. Crime Data includes many various crime events such as event time, event class, event spatial information and event object. These data have many attributes at different levels. So the attribute oriented induce method is chosen to deal with these data. This paper presents an improved attribute oriented induce method and algorithm related to crime hot spots detecting and a simple mapping method is depicted. Limitation and future work for this paper is further design and optimization of taxonomy. Taxonomy of attribute is the key related to the accuracy of the results. The mapping method in this paper is relatively simple. Improving the mapping method is another important work later.

“Crime hotspot mapping using the crime related factors—a spatial data mining approach”

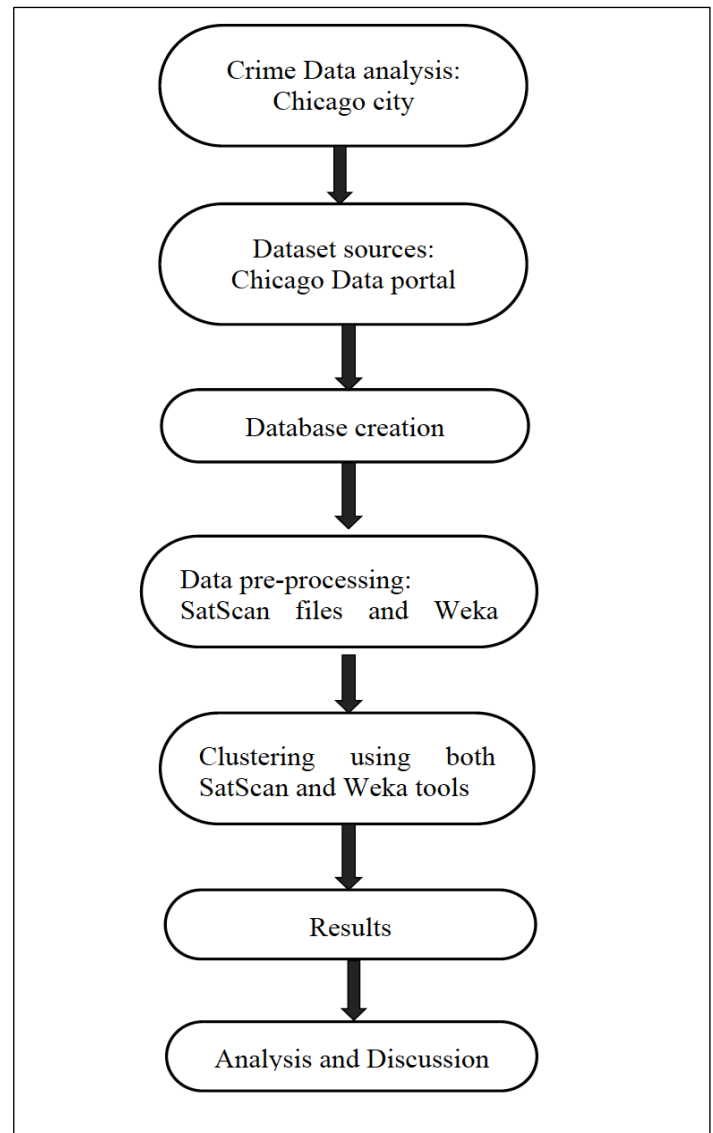
[5] also discusses about the use of Hotspot Mapping. Author says that spatial distribution of crime is considered to be related with a variety of socio-economic and crime opportunity factors but none of the existing techniques focus on it. In this study, a new crime hotspot mapping tool—*Hotspot Optimization Tool* (HOT) is introduced which is an application of spatial data mining. Experiments are done using a real-world dataset from a northeastern city in the United States and the pros and cons of utilizing related factors in hotspot mapping are discussed.

III. METHODOLOGY:

In this paper, the methodology part includes all knowledge discovery and data mining stages from the beginning of data gathering, database creation and design, data processing, transformation, and data mining, to the analysis and evaluation stages. Our research paper followed two approaches in our methodology:

- SAT SCAN to find the hot spots and do the clustering for the spatial data.
- Clustering technique by using WEKA tool by applying K-means algorithm for Euclidean distance measure.

The following diagram shows the workflow of our methodology:

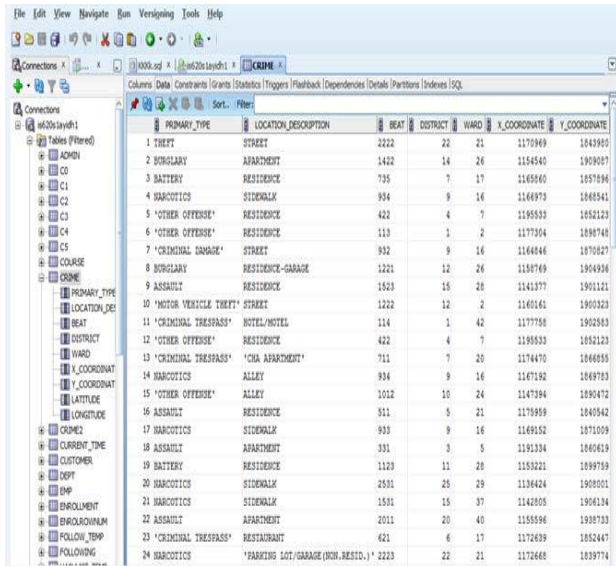


A. Data Collection Process:

This was the first and foremost step of our project. It was to collect crime data from Chicago Data Portal. It was from 2010- 2012 when there was a high crime rate in Chicago and it was named as ‘crime capital’ as reported by FBI. Also, population was taken from government website and census site. Data collected here was used as input in further process.

B. Database Creation & Design:

In this step, we have created crime database and have imported crime data into this database. SQL Developer was used for this database creation and design. First step here was to create crime table. The crime table is shown in below after successful creation:



The screenshot shows the SQL Developer interface with a table named 'CRIME'. The table has the following columns: PRIMARY_TYPE, LOCATION_DESCRIPTION, BEAT, DISTRICT, WARD, X_COORDINATE, and Y_COORDINATE. The table contains 24 rows of data, including crimes like THEFT, BATTERY, AGGRAVATED BATTERY, and CRIMINAL DAMAGE.

C. Data Preprocessing:

As mentioned earlier we have used two approaches using SatScan and WEKA tool. Before experimenting and using these tools we need to clean our data, since, our data was in raw format. It was necessary to transform it into uniform data, removing missing values, normalizing data, removing errors, perform some aggregation etc. Data Preprocessing was done for both tools SatScan & WEKA files.

1) Preprocessing for SatScan files:

Spatial and space-time Clustering method in SAT SCAN requires three input files viz. case file, population file, and coordinate file. However, these files do not exist in our dataset directly, so with the use of our crime information database which we have created using Oracle SQL Developer, we created these 3 input files. For this lot of aggregation functions were performed by using SQL queries. One of the important preprocessing done here was Normalizing the coordinate file since location id in coordinate file were wards and SatScan doesn't accept that location id has many latitudes, longitude. Thus. it was required to normalize coordinate file. Finally, by using SQL queries and performing aggregate functions we had our case,

population and coordinate file ready for input.

2) Preprocessing for WEKA file.

For using WEKA tool, we needed only 1 file as input to perform clustering using K-means algorithm. Primary attributes for this file was Primary type of crime, location description, date and time of crime, districts, wards, latitude and longitude. Extension for this file was .arff. Data Preprocessing done here and it was removal of missing values and transforming data into uniform format.

Thus. summary of DATASETS and attributes used for clustering:

SatScan files:

Case file	Location(id) ward	Number of cases	Date (2010-2012)
Population file	Location(id) ward	Date (2010-2012)	Population
Coordinate file	Location(id) ward	Longitude	Latitude

Weka file:

Date	Primary Type	Location description	District	Ward	Longitude	latitude
------	--------------	----------------------	----------	------	-----------	----------

D. Clustering:

Cluster (of crime) refers to a geographical group of crime, i.e. number of crimes in a given geographical region. Such clusters can be visually represented using a geo-spatial plot of the crime overlayed on the map to make best use of police resources. The densely populated group of crime is used to visually locate the ‘hot-spots’ of crime. Currently, it is said that the most useful method for detecting crime hot spots is the spatial clustering. Also, K-means clustering is one of the typical methods and the most widely used data mining clustering technique. Therefore in our paper, we have used both approaches: Spatial Clustering using SatScan and K-means clustering using WEKA tool.

1) SatScan Spatial Clustering:

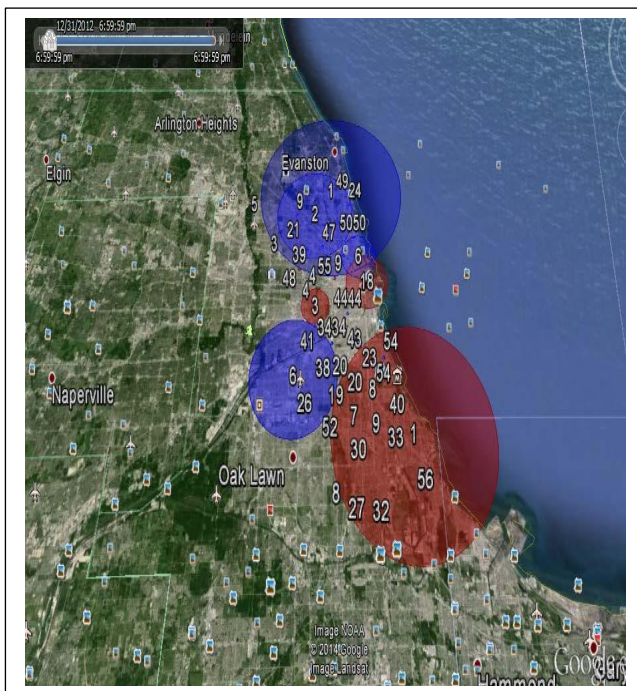
SatScan is a free software that analyze spatial, temporal and space-time data using the spatial, temporal or space-time scan statistics. It is designated to perform different jobs such as detect spatial or space-time crime clusters and to see if they are statistically important or not. It can be used to see whether a crime is randomly distributed over space, over time or over space and time.

So, our goal is to detect spatial crime clusters by using Poisson model. For discrete spatial analyses, we need 3 files. Data containing the spatial coordinates of a set of locations (coordinates file). For each location, the data must further contain information about the number of cases (crimes) at that location (case file) and population size for

each location (population file). Scan statistics will detect and evaluate clusters in purely spatial setting. This is done by scanning a window across space and noting the number of observed and expected observations inside the window at each location. In this the scanning window is a circle or an ellipse (in space). Three files are ready for clustering and three steps to perform this type of clustering are:

1. The Input Tab is used to specify the names of the 3 input data files.
2. In analysis tab, we select purely spatial analysis and Poisson model in the probability model. In Poisson model cases are included as part of the population count. It is possible to scan for areas with high rates only (clusters), for areas with low rates, or simultaneously for areas with either high or low rates.
3. In output tab, we define where SatScan should save our results. By specifying results file information about the clusters detected, summary information about the data, computing time and the analysis parameters chosen are shown completely. Also there is KML File that will show the detected clusters in Google Earth and other geographical software. [6]

After running SAT SCAN and opening the results in google earth, we found out that SAT SCAN cluster the data (for both high and low rate) and hotspots as can be seen below.



2) Clustering using K- means:

In this approach, we performed clustering by using WEKA tool using K-means algorithm for Euclidean distance measure. Moreover, clustering is data mining technique where cluster or group data depends on similarities and the distance between points. We also measured quality of

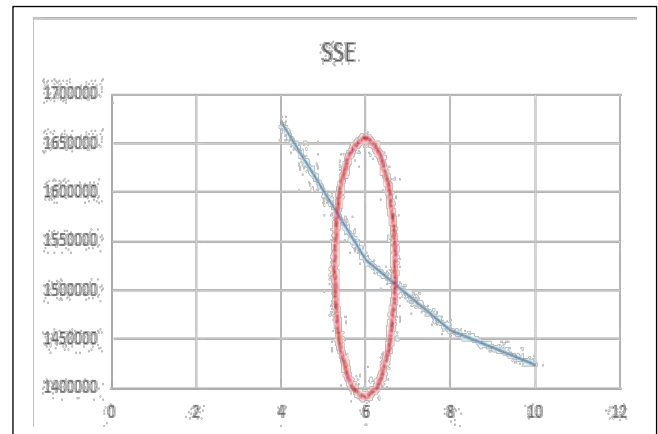
cluster using Sum of Squared errors (SSE).

E. EXPERIMENT&RESULTS:

This part of the document contains all the results retrieved from the two approaches.

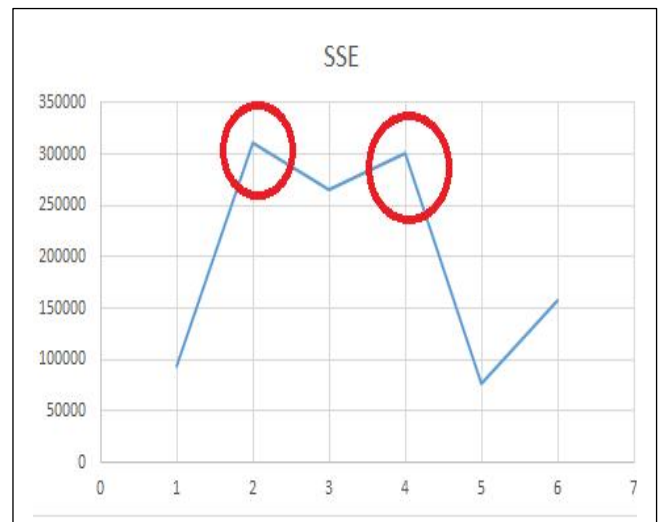
1) K- means Clustering results:

First step in using K-means clustering was calculating sum of squared errors for different number of clusters on our data set from 4 to 10 to find the ideal k cluster as can be shown in the below graph.



From the diagram, it shows that the k=6 is the ideal K for cluster.

After identifying the ideal cluster for the data set, the next step is to test each cluster of the results when k=6 and calculate the SSE for each cluster and plot the results on the graph as can be seen in the below figure.



From the diagram, it shows that there are two clusters which have high SSE which gives a clue that these two clusters might have unusual patterns or are disorganized. Further investigation is needed for these two clusters. We tested both clusters whether they have outliers or not by applying interquartile range for anomalies detection by using WEKA. We found out that cluster number 4 has outliers as shown in

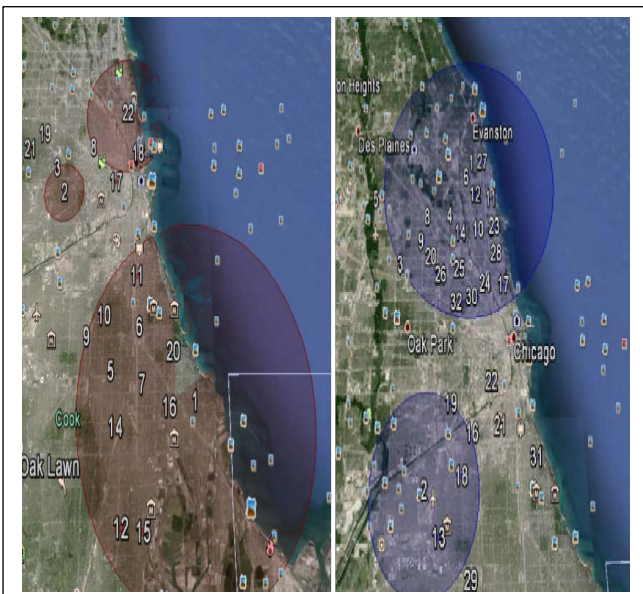
the following table:

Date	District	Ward	Latitude	Longitude	Outlier
12/29/2012	4	7	41.75157	-87.5704	no
12/29/2012	16	38	41.96057	-87.769	no
12/29/2012	4	7	41.75001	-87.5527	no
12/29/2012	3	6	41.76367	-87.617	no
12/29/2012	24	49	42.02264	-87.6727	yes
12/23/2012	2	3	41.81457	-87.6228	no
12/23/2012	24	40	41.99644	-87.6701	yes

Above table illustrate that there are outliers in the cluster, so we need to do more investigation why these two points are outliers and far from the centroid of the cluster. Therefore, we searched about these two wards in our data set and calculated the number of crimes. We found out that ward number 49 and 40 have the lowest number of crimes comparing by the other points or wards in the cluster. However, the number of crimes for ward 49th= 306 crimes, and ward 40th = 339 crimes. Also, the number of crimes for ward 7th = 7811 which is not outlier point since it is a ward with high risk and it is showing high number of crimes showing similarity with other points. Thus, these two wards are outlier and out of the range of clusters because they do not have the similarities that other points have in the low risk cluster.

1) *Spatial Clustering Results:*

When analysis is done using SatScan, we selected two types of scanning: Scanning for clusters with high rate and scanning for clusters with low rates. Scanning with high rate shows clusters having large number of crimes and also shows number of wards under high rates of crime. Scanning with high (shown in red) and low rate clusters (shown in blue) are shown below:



Wards which are under high crime rates are ward number 6,7,24,28 etc. Number of cases are 2,64,954 in these wards and relative risk is 1.93. Scanning for clusters with high rates gives sorted results from highest likelihood to lowest likelihood. SatScan provides complete information about results file. Here, is a sample of both high and low rate scanning file:

Purely Spatial analysis
scanning for clusters with high rates
using the Discrete Poisson model.

SUMMARY OF DATA

Study period.....	2010/1/1 to 2012/12/31
Number of locations.....	50
Total population.....	2695598
Total number of cases.....	686553
Annual cases / 100000.....	8487.7

CLUSTERS DETECTED

```

1. Location IDs included.: 7, 8, 5, 10, 6, 20, 21, 17, 9, 4, 34, 3, 16, 1!
Overlap with clusters.: 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 20
Coordinates / radius.: (41.746397 N, 87.561151 W) / 10.76 km
Gini Cluster.....: No
Population.....: 633315
Number of cases.....: 264954
Expected cases.....: 168942.44
Annual cases / 100000.: 13311.3
Observed / expected...: 1.57
Relative risk.....: 1.93
Log likelihood ratio.: 32729.970090

```

Purely Spatial analysis
scanning for clusters with low rates
using the Discrete Poisson model.

SUMMARY OF DATA

Study period.....	2010/1/1 to 2012/12/31
Number of locations.....	50
Total population.....	2695598
Total number of cases.....	686553
Annual cases / 100000.....	8487.7

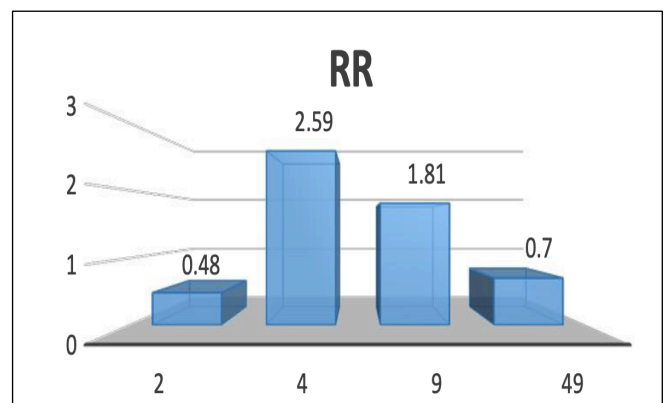
CLUSTERS DETECTED

```

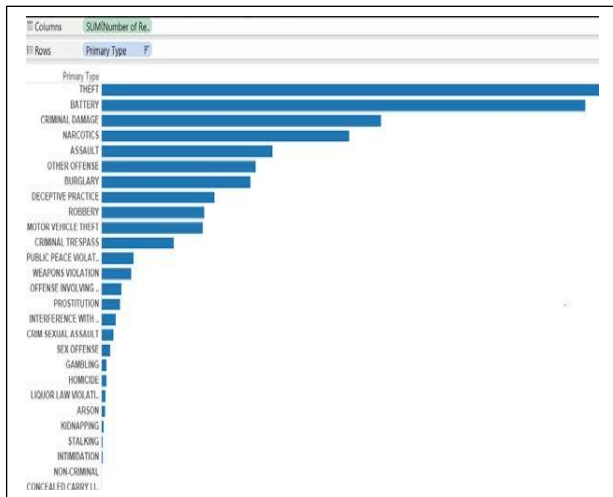
1. Location IDs included.: 50, 40, 49, 48, 39, 47, 33, 46, 45, 44, 35, 38, 30, 31, 32, 1, 43,
                           26, 36
Overlap with clusters.: 3, 4, 6, 8, 9, 10, 11, 12, 14, 17, 20, 23, 24, 25, 26, 27, 28, 30,
Coordinates / radius.: (42.000481 N, 87.694800 W) / 11.47 km
Gini Cluster.....: No
Population.....: 1062793
Number of cases.....: 164719

```

We identified various clusters showing high crime rate and low crime rate wards and no. of crimes in that ward.

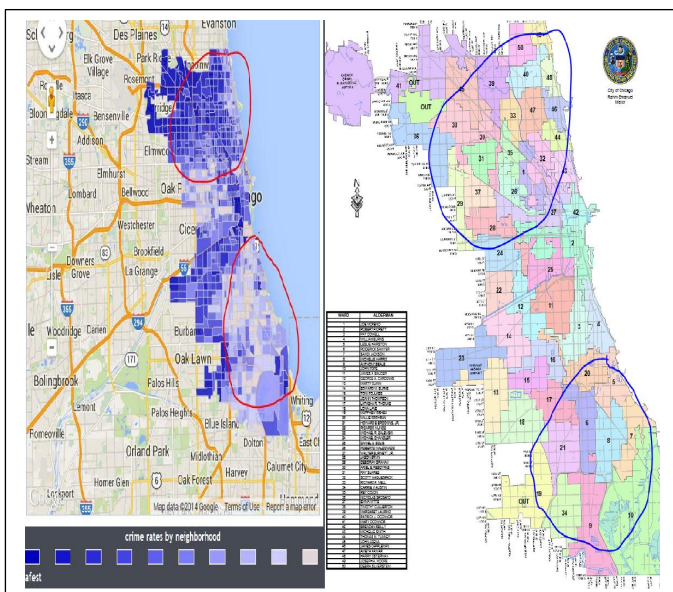


Experiments also shows us that that theft was the most highly occurred way of crime as can be seen in the below.

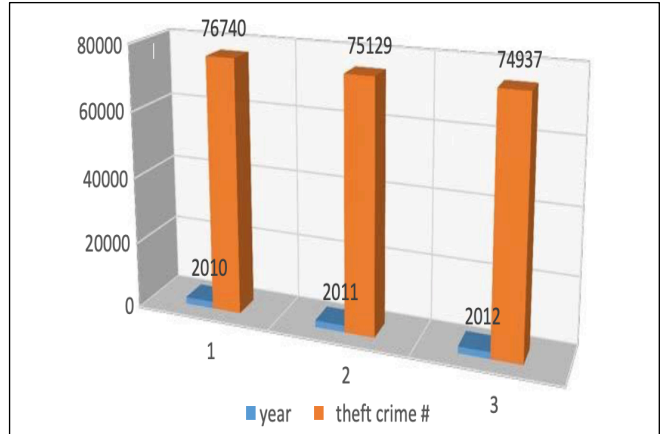


F. ANALYSIS:

We compared results of 2 approaches Spatial Mining using SatScan and K-means clustering using WEKA. By comparing both approaches we found out that the wards with high rate of crimes in Sat Scan are same like the wards with high number of crimes in K-means clustering. Moreover, we wanted to confirm our results with some ground reality. So, we searched for some safest neighborhoods in Chicago city. Also, tried to find out areas with high crime rate and how it has changed over the years. Ward # 20 has high rate of crime in both Spatial and K-means clustering. We looked for some ground truth to cross check our 2 approaches and found that Ward 20 is bad neighborhood known as Englewood, and different reasons were seen for high crime rates in this place. One important factor which we came up was education; the statistics show that the percentage of people who do not have high school diploma degree is around 29.4%. [7] This high percentage of illiteracy contributes to increase of the number of crimes in this place. Similarly, ward #49 is a low rate crime ward in K-means clustering, and if we look for the same ward in the Spatial clustering, it also gives us the same result. We crossed checked these results by ground truth and found out this ward is a good neighborhood and is one of the safest wards. Another very interesting ground reality found was direct comparison of safest wards of Chicago City labelled from safest to least safe [8] which proves our SatScan and WEKA results.



Now, we tried to find out about the most common and highly occurred way of crime i.e. Theft. Analysis shows some interesting facts about this. There were few important changes after 2010 where numbers of theft cases were decreased. We tried to identify reason behind this sudden change.



Ground reality behind this was found from one of the news shown below:



After finding all these results and ground truths about wards and places in city, we performed some in-depth analysis for the reason behind this high number of crimes in few locations and low number in other. In this part, we found some facts and common properties between safe neighborhoods and places which are under high risk. Few of them are: People living in these high crime rate wards generally have lower-middle income. Most of the children in almost all these places are living below the federal poverty line. There were some neighbors who had higher rate of childhood poverty as high as 91% of U.S. neighborhoods.

Rents here are lower in price and neighborhoods are primarily filled with college students. Also, in all these places there is a greater number of Singles (never married) people. These neighborhoods identify their ethnicity or ancestry as Sub-Saharan African and African and Mexican. One of such area under high risk is S Indiana Ave / E 60th

St which is unique for having just 4.3% of adults having earned a bachelor's degree. This is a lower rate of college graduates than Neighborhood Scout found in 96.5% of America's neighborhoods. Other such area was S Racine Ave / W Marquette Rd which has more single mother. Often high concentrations of single mother homes can be a strong indicator of family and social issues such as poverty, high rates of school dropouts, crime, and other societal problems. On the other hand, facts about safest wards were just opposite of facts for high crime rates. These places are more expensive of the neighborhoods in Illinois. The average rental cost in this neighborhood is very high. These places have some of the lowest rates of children living in poverty. Analysis reveals that most adults here are well-educated and with the greatest number of residents employed. These places have high percentage of married people. Also, this type of neighborhood is classified as quiet and sophisticated. Some of the safest neighborhoods are N Caldwell Ave / N Lehigh Ave, W Devon Ave / N Central Ave, S Western Ave / W 95 St. [9]

IV. CONCLUSION & RECOMMENDATIONS:

We looked at the use of data mining for identifying crime patterns using the clustering techniques. This paper presented two approaches for detecting the crime hotspots. We have mined historical crime data sets with typical K-means algorithm and Spatial Clustering algorithm and setup visualization over Google earth to know area of crime and investigate crime data. We analyzed our results with some ground truths and founded some interesting facts surrounding high and low areas of crime rate. This project should help Police to track crime incidents within Chicago city in real-time, retrieve historical crime incidents for given wards and also make best use of resources in high crime area. We would also like to recommend more deployment of security people and patrolling in wards of high crime like ward # 5, 6, 7, 8, 9, 10, 24, 28.

V. INSIGHTS GAINED & FUTURE WORK:

We have really gained a lot of knowledge about the use of Data mining and clustering which can be applied in the field of criminology and many others. Also, data mining is sensitive to quality of input data that may be inaccurate, have missing information, be data entry error prone etc. Thus, preprocessing is the primary and most important step in data mining. Secondly, after preprocessing and using various tools, it becomes quite easy to get results. However, the important part lies in analyzing those results accurately. WEKA and specially SatScan for Spatial data mining were one of the effective tools for clustering.

In future, we would like to do more Spatial Temporal mining, so that we can utilize deployment of police at places of crime for any given window of time. In addition, Crime Data Information System, which will be able to retrieve and process crime data in various forms and perform association analysis and prediction on the data sets. Furthermore, mining applications for audio files and image data are still in their beginning stage. These types of information can be used to identify vehicles, persons and unique characteristics of criminals.

ACKNOWLEDGEMENT:

We are grateful to our Professor Dr. Vandana Janeja who took out her valuable time for reviewing and providing feedback for our research. We would also like to thank Chicago Data Portal for making crime datasets available.

REFERENCES

- [1] Dylan Stableford (2013 September 19). Chicago now murder capital of U.S., FBI says. *Yahoo News*. Retrieved from
- [2] Monica Davey (2012 JUNE, 25) Rate of Killings Rises 38 Percent in Chicago in 2012.
- [3] Shyam Varan Nath, "Crime Pattern Detection Using Data Mining," Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on , vol., no., pp.41,44, Dec. 2006
- [4] Xiang Zhang; Zhiang Hu; Rong Li; Zheng Zheng, "Detecting and mapping crime hot spots based on improved attribute oriented induce clustering," *Geoinformatics, 2010 18th International Conference on* , vol., no., pp.1,5, 18-20 June 2010 doi: 10.1109/GEOINFORMATICS.2010.5568075
- [5] Dawei Wang, Wei Ding, Henry Lo, Tomasz Stepinski, Josue Salazar, Melissa Morabito "Crime hotspot mapping using the crime related factors—a spatial data mining approach" *Applied Intelligence, 2013*, vol.,no., pp.772-781
- [6] SaTScanTM User Guide for version 9.3, By Martin Kulldorff March, 2014 (<http://www.satscan.org/>)
- [7] Crime Reports in Englewood (2014 October 27) Retrieved from <http://crime.chicagotribune.com/chicago/community/englewood#note-1>
- [8] Enterprise-grade data for every neighborhood and city in the U.S. Retrieved from <http://www.neighborhoodscout.com/il/chicago/crime/#content-data>
- [9] Enterprise-grade data for every neighborhood and city in the U.S. Retrieved from <http://www.neighborhoodscout.com/il/chicago/devon-central/>
- [10] Chicago crime rates by community area (2014 November 26) Retrieved from <http://crime.chicagotribune.com/chicago/community>
- [11] Aldermanic Wards for the City of Chicago Map. Retrieved from http://www.cityofchicago.org/content/dam/city/depts/doit/general/GIS/Chicago_Maps/Citywide_Maps/Wards.pdf