INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 70-79
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

CRIME ANALYSIS IN CHICAGO CITY

Ankita Thakur[1], Mr. Anurag Anand Duvey[2]

[1]Student, Computer Science and Artificial Intelligence, Poornima Institute of Engineering & Technology Jaipur, Rajasthan, India.

[2]Assistant Professor, Deptt. of Artificial Intelligence and Data Science, Poornima Institute of Engineering & Technology Jaipur, Rajasthan

ABSTRACT

Security has consistently been one of the most important issues. To stop crimes and keep its citizens safe, government and security forces are putting forth a lot of effort. However, handling vast amounts of data has become a significant problem for all enterprises. Therefore, in order for investigators to identify crime hotspots, patterns, and future trends, they require a criminal information system that can analyze massive amounts of data quickly. The design of the Crime Data Information System is presented in this paper. Two methods for crime analysis are carried out, together with data pretreatment in the Crime Database. These two methods are contrasted, and ground truth is used to validate the results.

Keywords- Component, Data Mining, Crime, Preprocessing, Clustering, Spatial Clustering, Data Analytics, Analysis.

1. INTRODUCTION

Crime is on the rise in practically every nation. It is imperative to evaluate various crime locations and spot trends in criminal activity. The scale of crime data is growing quickly, making it challenging to manage such a vast amount of data and to maintain records of crimes that are geographically dispersed and occur at different times, despite the efforts of security organizations around the world to reduce these crimes. Therefore, having a crime information system that can process a lot of data quickly is essential.

Exploring, analyzing, identifying patterns, and forecasting future crimes in vast amounts of data is made possible by data mining techniques such as clustering, classification, and association mining. Conventional methods lacked a central criminal database to go into and uncover connections. They kept crime records using paper-based systems, which made it very difficult for security authorities to identify trends and make better use of their resources. Computer data analysts have begun assisting security personnel in expediting the investigation of crimes as a result of the growing usage of computerized systems to track crimes. Nevertheless, the data analysis was done superficially. Data was generally irregular since it was not being gathered in a scientific way. There was no defined procedure or method in place for gathering, cleansing, and deriving knowledge from data. Therefore, data mining has gained a reputation as a potent technique that helps researchers work with and examine vast amounts of data. Data about time and space can also be gathered with the use of a variety of technologies, including GPS units, satellite images, cellular phones, and sensor networks.

This facilitates data mining in both spatial and spatial-temporal domains. It assists in finding non-trivial information that is not detectable by conventional systems. Motivation: As we all know, crime is on the rise, and if it is not stopped, it may have negative consequences or make the atmosphere unfriendly for everyone. Protecting the country from criminals must be the first concern. Numerous resources are available to the police and other security agencies to assist the nation's residents.

The secret to solving this issue, though, is to place the appropriate resource at the appropriate time and location. Records and news reports show that Chicago has had a high number of crimes. In 2012, it was referred to as the United States' "murder capital" or "crime capital." Chicago had more murders in 2012 than any other city in the country. According to the FBI, Chicago had 500 homicides last year, the most in the whole country. [1] Additionally, a New York Times story claims that Chicago's murder rate increased by 38% in 2012.[2] As a result, it is necessary to make better selections and examine crimes that take place in various parts of Chicago.

We can identify various trends with the aid of data mining, which will help the people of Chicago by using the appropriate resources at the appropriate time and location. This will assist prevent crime in the city and make the most of the security departments' efficiency. The analysis of historical data pertaining to various crimes that took place in various parts of Chicago is the focus of this research work. In order to create a crime data information system, we extracted data from the Chicago Data Portal. Data preparation, which transformed raw data into a standard format, came next. Crime type, crime time, and crime location were significant characteristics in our dataset.

To be more precise, we will employ two methods. The first is clustering using the K-means algorithm, and the second is spatial mining to identify criminal hotspots. We will use the WEKA tool and Euclidean distance as a metric for clustering using K-means.

By doing this, we will create several clusters and pinpoint criminal hotspots. Finding comparable types of crimes in the specified area of interest is made easier by clustering. To visually identify the "hot-spots" of crime, the highly populated group of crimes is used. In order to make the best use of police resources, this will assist in placing police officers at the most likely crime scenes within any given window of time. Software called SatScan will be used to find hotspots. Chicago's crime statistics for the years 2010–2012 will be analyzed using data gathered from the Chicago Data Portal.

2. RELATED WORK

The concept of crime detection was first forth in "Crime Pattern Detection Using Data Mining" by Shyam Varan Nath [3], who claimed that the K-means clustering algorithm, which uses data mining, can assist in identifying criminal patterns and expediting the crime-solving process. This study corroborated findings by using clustering to actual crime data from a sheriff's office. Using an expert-based semi-supervised learning approach, the author of this research determined which traits were important and created a weighting scheme for them. This makes it possible to dynamically assign varying weights to various qualities according to the sorts of crimes being clustered. This enables the category attributes to be weighed as well. The lack of a forecast for crime hotspots that would assist the police in making use of their resources during any particular time frame was a limitation of this study.

The topic of identifying high-crime-density locations is covered in "Detecting and Mapping Crime Hot Spots Based on Improved Attribute Oriented Induce Clustering" [4]. It claims that geographical clustering is the most effective technique for identifying crime hotspots. Numerous different crime events, including event time, event class, event spatial information, and event object, are included in crime data. Numerous properties at various levels are present in these data. In order to handle these data, the attribute-oriented induce approach is selected. An enhanced attribute-oriented induction approach and algorithm for identifying crime hot locations are presented in this study, along with a straightforward mapping technique. This paper's limitations and future work include additional taxonomy design and optimization. The key to the correctness of the outcomes is the attribute taxonomy. In this study, the mapping procedure is straightforward. Later, there will be more significant work to improve the mapping approach.

High-crime-density region detection is covered in "Detecting and Mapping Crime Hot Spots Based on Improved Attribute Oriented Induce Clustering" [4]. It claims that the most effective technique for identifying crime hotspots is spatial clustering. Event time, event class, event spatial information, and event object are only a few of the crimes that are included in crime data. Numerous features are present in these data at various levels. To deal with these data, the attribute-oriented induce approach is thereby selected. This work illustrates a basic mapping approach and proposes an enhanced attribute-oriented induce method and algorithm related to criminal hot zone detection. Further taxonomy design and optimization is a limitation and area for future research for this publication.

The key to the results' correctness is the attribute taxonomy. In this study, the mapping procedure is straightforward. Later, there will be more significant work to improve the mapping approach.

The application of hotspot mapping is also covered in "Crime hotspot mapping using the crime related factors—a spatial data mining approach" [5]. According to the author, the spatial distribution of crime is thought to be connected to several socioeconomic and crime opportunity elements, but none of the current methods concentrate on this. The Hotspot Optimization Tool (HOT), a new criminal hotspot mapping tool that uses spatial data mining, is presented in this paper. The benefits and drawbacks of employing related factors in hotspot mapping are examined, and experiments are conducted using a real-world dataset from a northeastern American city.

3. METHODOLOGY

The methodology section of this paper covers every step of the knowledge discovery and data mining process, from the initial stages of data collection, database design and creation, data processing, transformation, and data mining to the stages of analysis and assessment. Two methods were used in our study paper's methodology:

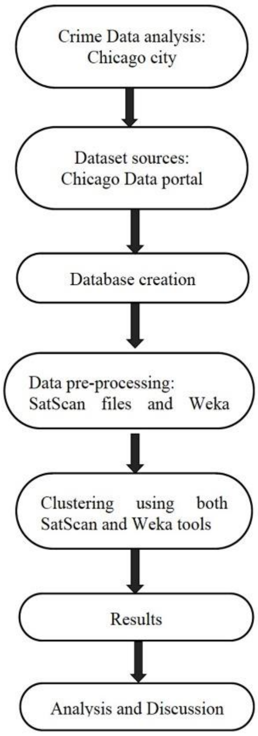
- SAT SCAN to cluster the geographic data and identify hot spots.
- Clustering technique employing the K-means algorithm for Euclidean distance measure with the WEKA tool.

The workflow of our methodology is depicted in the following diagram:

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 70-79

e-ISSN : 2583-1062
Impact Factor : 7.001



Dig-1: Methodology Workflow

A. Data Collection Process:

This was our project's first and most important phase. Its purpose was to gather information about crime from the Chicago Data Portal. Chicago was dubbed the "crime capital" by the FBI due to its high crime rate between 2010 and 2012. Additionally, the population was obtained from census and government websites. The information gathered here served as input for subsequent procedures.

B. Database Creation & Design:

This step involved creating a crime database and importing criminal data into it. This database was created and designed using SQL Developer as shown in Fig-1. Making a crime table was the first stage in this process. Once it has been successfully created, the crime table appears below:

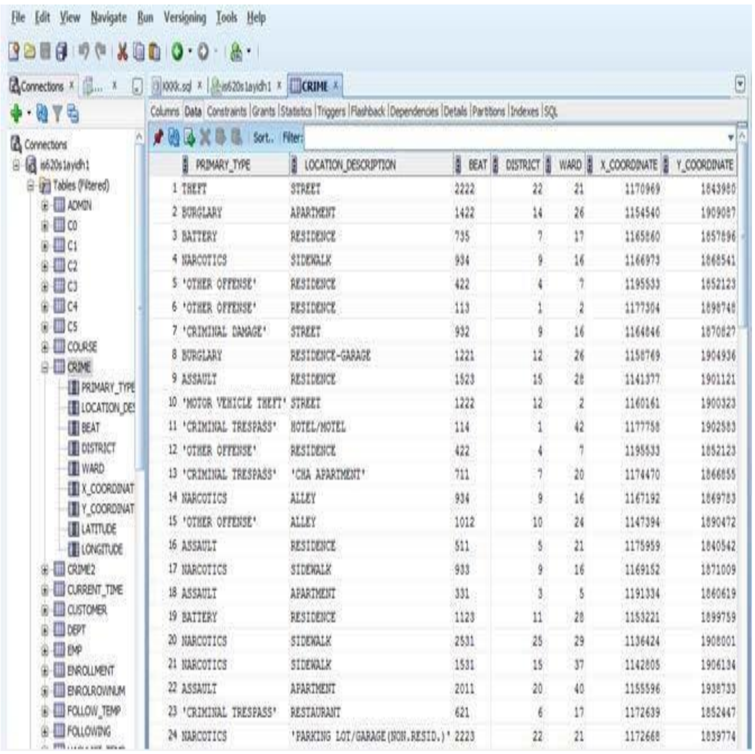


Fig-2

C. Data Preprocessing:

As previously stated, we have employed two strategies utilizing the SatScan and WEKA tools. Since our data was in raw format, we had to clean it before we could experiment and use these technologies. It has to be transformed into uniform data, with missing values removed, data normalized, errors removed, some aggregating done, etc. Both SatScan and WEKA tool files underwent data preprocessing.

1) **Preprocessing** for SatScan files:

Space-time and spatial Three input files—the case file, the population file, and the coordinate file—are needed for the SAT SCAN clustering method. But since these files aren't directly in our dataset, we made these three input files using our criminal information database, which we made with Oracle SQL Developer. SQL queries were used to carry out a number of aggregating tasks for this lot. Normalizing the coordinate file was a crucial preparation step in this case because the location IDs in the coordinate file were skewed, and SatScan does not accept location IDs with multiple latitudes and longitudes. As a result, normalizing the coordinate file was necessary. Lastly, by utilizing SQL we got our case, population, and coordinate files prepared for input by running queries and carrying out aggregate operations.

2) **Preprocessing for WEKA file:**

We only needed one file as input to use the K-means algorithm for clustering when utilizing the WEKA program. This file's primary properties included the following: district, ward, latitude, longitude, location description, date and time of the crime, and primary type of crime. This file has the extension.arff. Here, data preprocessing was carried out, which involved converting the data into a consistent format and eliminating missing values.

Consequently, a synopsis of the datasets and clustering attributes is as shown in Table-1 & Table-2:

SatScan files:

Table-1

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:

Table-2

Date	Primary Type	Location description	District	Ward	Longitude	latitude

D. Clustering:

A geographical group of crimes, or the quantity of crimes in a specific geographic area, is referred to as a cluster (of crime). To maximize police resources, such clusters can be graphically depicted by superimposing a geospatial plot of the crime on the map. The "hot-spots" of crime are visually located using the densely populated group of crimes. Spatial clustering is now thought to be the most effective technique for identifying crime hot regions. Additionally, the most popular and common data mining clustering methodology is K-means clustering. Therefore, we employed both methods in our paper: K-means clustering using the WEKA tool and spatial clustering using SatScan.

1) **SatScan Spatial Clustering:**

In order to evaluate geographical, temporal, or space-time data, SatScan is a free program that uses scan statistics. Among its many tasks is the detection of spatial or space-time crime clusters and the determination of their statistical significance. It is useful for determining whether a crime is dispersed randomly over time, space, or both. Therefore, our objective is to use the Poisson model to identify spatial crime clusters. Three files are required for discrete spatial analyses.

Coordinates files are data sets that contain the spatial coordinates of a group of locations. Additionally, the data must include the population size for each site (population file) and the number of cases (crimes) at each (case file). In a strictly spatial context, scan statistics will identify and assess clusters. Counting the number of observed and anticipated observations within a window at each point is accomplished by scanning the window across space. This scanning window is either an ellipse (in space) or a circle. Three files are prepared for clustering, and you can carry out this kind of clustering in three steps:

1. The names of the three input data files are specified using the Input Tab.
2. We choose the Poisson model in the probability model and only geographic analysis under the analysis tab. The population count in the Poisson model includes cases. places with low rates, places with high rates (clusters), or areas

with both high and low rates at the same time can all be scanned.

3. We specify where SatScan should save our data in the output tab. Information about the clusters found, data summary, computation time, and selected analysis parameters are all displayed in full by specifying the results file. In Google Earth and other geographic applications, the identified clusters can also be seen in a KML file. [6]

We discovered that SAT SCAN clusters the data (for both high and low rate) and hotspots, as shown in below Fig-2, after running the program and seeing the results on Google Earth.

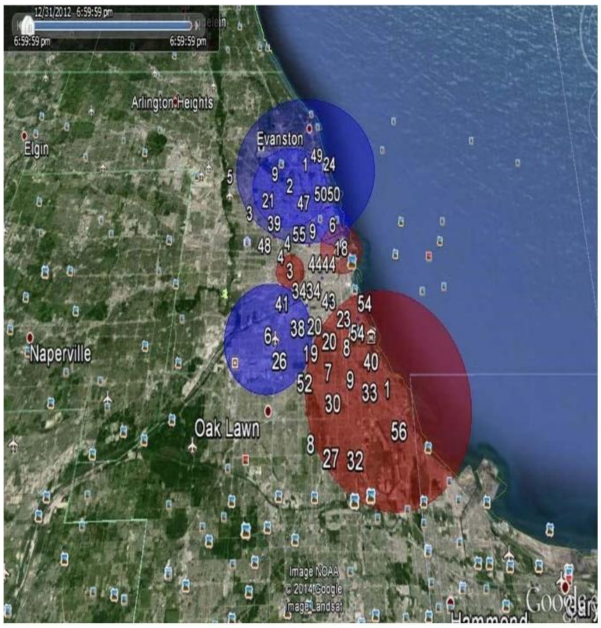


Fig-3

2) Clustering using K- means:

Using the K-means algorithm for the Euclidean distance measure, so that we clustered using the WEKA tool in this method. Additionally, clustering is a data mining approach in which the distance between the points and similarities determines whether to group or cluster data. We also used Sum of Squared Errors (SSE) to measure cluster quality.

4. EXPERIMENT&RESULTS

All of the findings from the two methods are included in this section of the publication.

K- means Clustering results:

In order to determine the optimal k cluster, as seen in the graph below as in Fig-3, the first step in applying K-means clustering was to compute the sum of squared errors for varying numbers of clusters on our data set, ranging from 4 to 10.

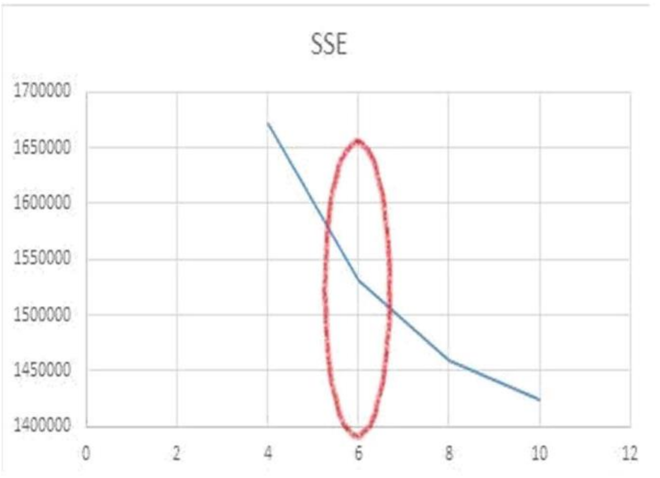


Fig-4

The graphic indicates that the optimal K for a cluster is k=6. Following the selection of the optimal cluster for the data set, each cluster's results are tested when k=6. The SSE for each cluster is then determined, and the results are shown on the graph as shown in the Fig-4.

Fig-5

The diagram indicates that there are two clusters with high SSE, which suggests that these two clusters may exhibit odd patterns or be disorderly. Both of these clusters require more research. We tested both clusters whether they have outliers or not by applying interquartile for anomalies detection by using WEKA.

We found out that cluster number 4 has outliers as shown in the following Table-3:

Table-3

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

The cluster contains outliers, as the above table shows, so further research is necessary to determine why these two sites are outliers and distant from the cluster center. In order to determine the number of offenses, we looked up information on these two wards in our data set. We discovered that, when compared to the other locations or wards in the cluster, wards 49 and 40 have the fewest offenses. Nonetheless, there are 306 crimes in ward 49 and 339 crimes in ward 40. Additionally, ward 7th has 7811 crimes, which is not an outlier point because it is a high-risk ward with a high crime rate that is comparable to other areas. Due to their lack of parallels with other locations in the low-risk cluster, these two wards are therefore outliers and outside of the cluster range.

1) Spatial Clustering Results:

We chose two scanning strategies while utilizing SatScan for analysis: searching for high-rate clusters and searching for low-rate clusters. High-rate scanning reveals both the number of wards with high crime rates and clusters with high crime rates. The following displays (Fig-5) scanning with low-rate clusters (blue) and high-rate clusters (red):

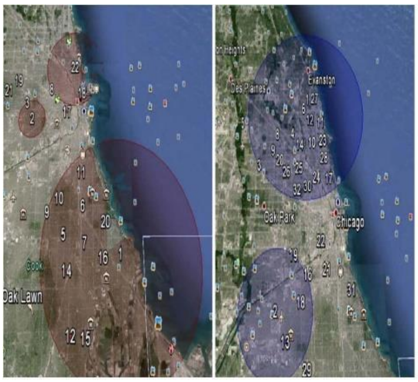


Fig-6

Wards number 6, 7, 24, 28, and so on have high crime rates. In these wards, there are 2,64,954 instances, and the relative risk is 1.93. Results are sorted from highest likely to lowest likelihood while looking for clusters with high frequencies. SatScan offers comprehensive details about the findings file. Here is an example of a scanning file with both a high and low rate in Fig-6:

```
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............: 663315
  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
```

Fig-7

We found several clusters that displayed wards with high and low crime rates as well as the number of crimes committed there as shown in Fig-7.

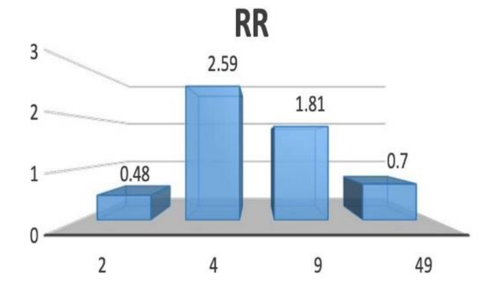


Fig-8

As may be seen below in Fig-8, experiments also indicate that theft was the most common form of criminal activity.

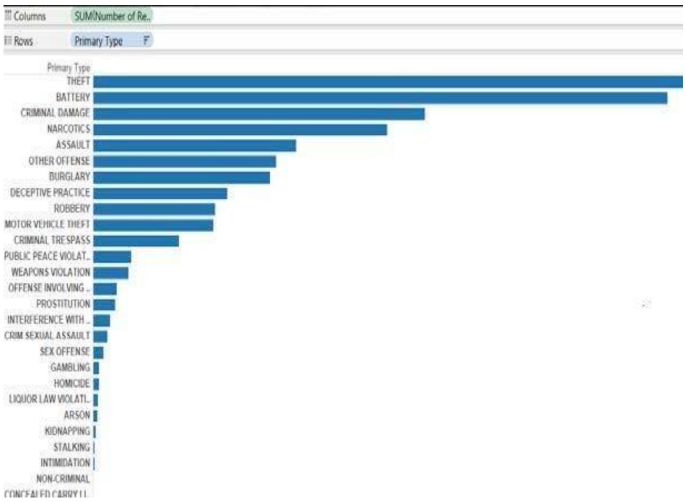


Fig-9

We then attempted to learn more about the most prevalent and frequently occurring type of crime in Fig-10, which is theft. An analysis of this reveals some fascinating insights. After 2010, there were not many significant changes, such as a decline in theft instances. We looked for the cause of this abrupt shift.

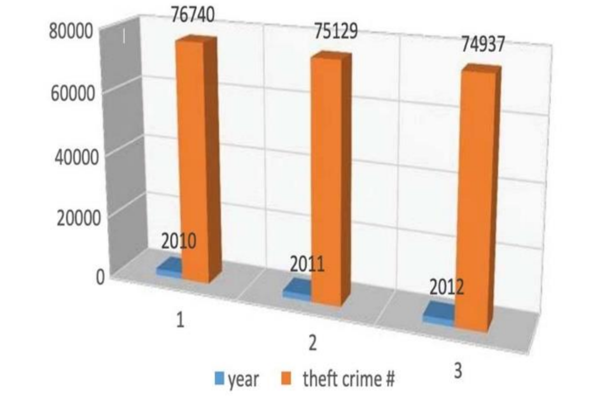


Fig-10

5. ANALYSIS

We contrasted the outcomes of two methods. SatScan for spatial mining and WEKA for K-means clustering. By contrasting the two methods, we discovered that the wards with the highest crime rates in Sat Scan and K-means clustering are identical. Additionally, we wanted some ground truth to support our findings & we looked for some of Chicago's safest neighborhoods. Additionally, an attempt was made to identify high-crime locations and track changes over time. The crime rate in Ward # 20 is high in both spatial and K-means clustering. In order to compare our two methods, we sought some empirical evidence. We discovered that Ward 20 is a dangerous neighborhood called Englewood, and that there are a number of reasons why the crime rate is so high there as shown in Fig-9. Education was a significant element that we discussed; according to the data, 29.4% of persons do not possess a high school diploma. [7] The high rate of illiteracy in this area is one factor contributing to the rise in crime. In a similar vein, ward #49 has a low crime rate according to K-means clustering, and the same outcome is obtained if we search for the same ward in spatial clustering. We discovered that this ward is among the safest and a good neighborhood. The direct comparison of Chicago City's safest wards, ranked from safest to least safe [8], is another fascinating underlying fact that supports our SatScan and WEKA findings.

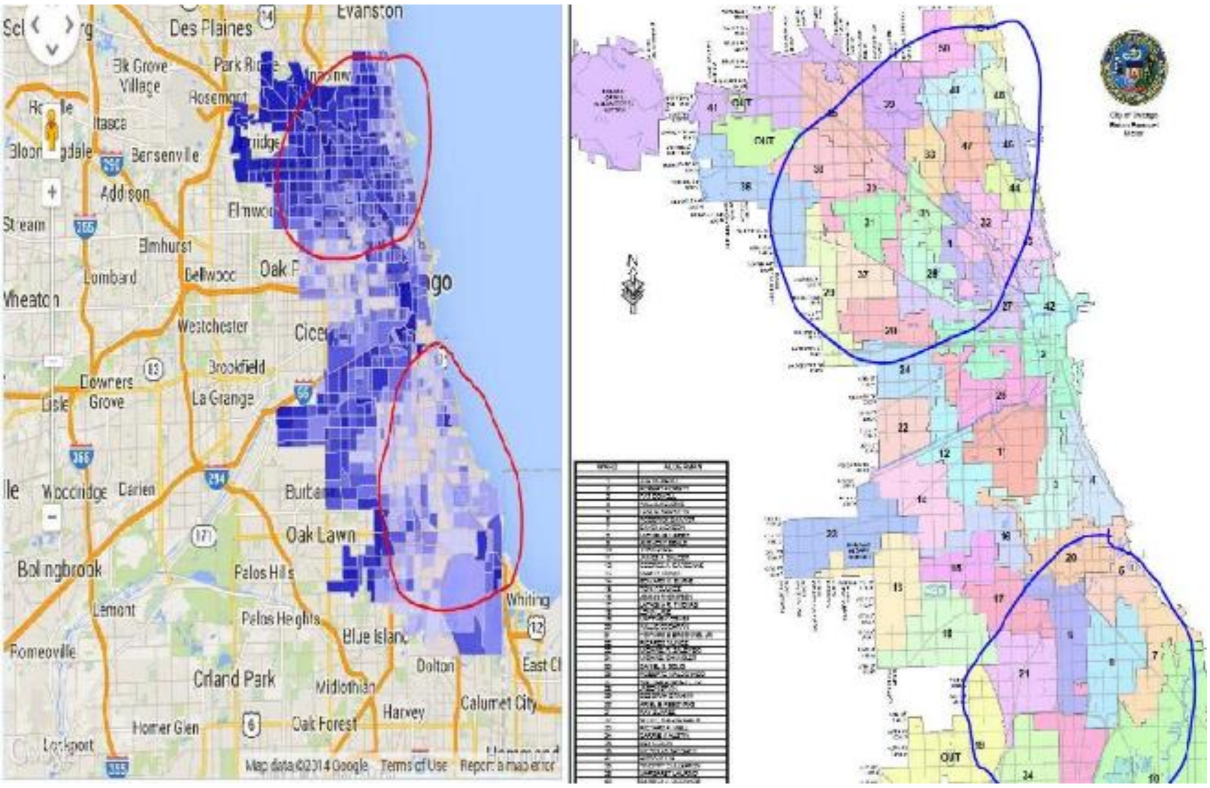


Fig-11

Ground reality behind this was found from one of the news shown below Fig-11:

Fig-12

After obtaining all of these findings and firsthand knowledge of the city's wards and locations, we conducted a thorough investigation to determine the cause of the high crime rate in some areas and the low crime rate in others. We discovered some information and similarities between high-risk areas and safe neighborhoods in this section. Among them are: Most residents in these wards with high crime rates are lower-middle class. In practically all of these locations, the majority of children live below the federal poverty level. Ninety-one percent of U.S. neighborhoods had a higher rate of childhood poverty than others. College students predominate in the neighborhoods, and rents are lower here. Additionally, there are more singles (those who have never married) in each of these locations. Sub-Saharan African and African and Mexican are the ethnicities or ancestries that these neighborhoods identify with.

Fig-9

One such high-risk neighborhood is S. Indiana Ave./E. 60th St., which is distinct in that just 4.3% of the adults there have a bachelor's degree. Compared to Neighborhood Scout, which reported 96.5% of American neighborhoods had college grads, this is a lower percentage. Another neighborhood with a higher number of single mothers was S Racine Ave/W Marquette Rd. High concentrations of single-mother households are frequently a good predictor of social and family difficulties like poverty, high school dropout rates, criminal activity, and other societal challenges.

However, the data regarding the safest wards was the exact reverse of the data regarding high crime rates. Compared to other neighborhoods in Illinois, these areas are more costly. In this neighborhood, the average rental price is extremely expensive. The percentage of children living in poverty is among the lowest in these areas. According to analysis, the majority of adults in this area have a good education, and the majority of citizens work. The proportion of married people in these areas is high. This kind of area is also categorized as classy and calm. N Caldwell Ave/N Lehigh Ave, W Devon Ave/N Central Ave, and S Western Ave/W 95 St. are a few of the safest neighborhoods.[9]

6. CONCLUSION & RECOMMENDATIONS

We examined the application of clustering algorithms in data mining to detect trends in criminal activity. Two methods for identifying criminal hotspots were provided in this research. To identify crime hotspots and look into crime data, we have mined historical crime data sets using the standard K-means and spatial clustering algorithms and built up visualization via Google Earth. We used basic ground realities to assess our results and discovered some intriguing statistics about high and low crime rates. This project should assist police in tracking crime episodes in real time throughout the city of Chicago, retrieving past crime incidents for specific wards, and optimizing resources in high-crime areas. Additionally, we would like to suggest increased security personnel deployment and patrols in high-crime areas, such as wards #5, #6, #7, #8, #9, #24, and #28.

7. INSIGHTS GAINED & FUTURE WORK

We now know a great deal more about the application of data mining and clustering, which has applications in criminology and many other fields. Additionally, the quality of the input data—which could be erroneous, lacking information, prone to data entry errors, etc.—affects data mining. Preprocessing is therefore the first and most crucial stage in the data mining process. Second, obtaining findings is rather simple after preprocessing and utilizing a variety of tools. The crucial aspect, though, is correctly evaluating such findings. Among the efficient techniques for clustering were WEKA and, in particular, SatScan for spatial data mining. In order to deploy police at crime scenes for any given window of time, we would like to undertake more spatial temporal mining in the future. Moreover, the Crime Data Information System will be able to handle and retrieve crime data in a variety of formats, as well as do prediction and association analysis on the data sets. Additionally, mining applications for image and audio data are still in their infancy. Vehicles, people, and distinctive traits of criminals can all be identified using this kind of information.

ACKNOWLEDGEMENT

We appreciate that our professor, Dr. Vandana Janeja, gave up her precious time to review and comment on our study. Additionally, we are grateful to the Chicago Data Portal for providing crime datasets.

8. 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/Cit ywide_Maps/Wards.pdf