



MONTCLAIR STATE
UNIVERSITY

Montclair State University
**Montclair State University Digital
Commons**

Theses, Dissertations and Culminating Projects

5-2019

Data Mining and Predictive Policing

Chanté L. Stewart-Wallace

Follow this and additional works at: <https://digitalcommons.montclair.edu/etd>



Part of the [Computer Sciences Commons](#)

Recommended Citation

Stewart-Wallace, Chanté L., "Data Mining and Predictive Policing" (2019). *Theses, Dissertations and Culminating Projects*. 310.

<https://digitalcommons.montclair.edu/etd/310>

This Thesis is brought to you for free and open access by Montclair State University Digital Commons. It has been accepted for inclusion in Theses, Dissertations and Culminating Projects by an authorized administrator of Montclair State University Digital Commons. For more information, please contact digitalcommons@montclair.edu.

Abstract

This paper focuses on the operation and utilization of predictive policing software that generates spatial and temporal hotspots. There is a literature review that evaluates previous work surrounding the topics branched from predictive policing. It dissects two different crime datasets for San Francisco, California and Chicago, Illinois. Provided, is an in depth comparison between the datasets using both statistical analysis and graphing tools. Then, it shows the application of the Apriori algorithm to re-enforce the formation of possible hotspots pointed out in a actual predictive policing software. To further the analysis, targeted demographics of the study were evaluated to create a snapshot of the factors that have attributed to the safety of the neighborhoods. The results of this study can be used to create solutions for long term crime reduction by adding green spaces and community planning in areas with high crime rates and heavy environmental neglect.

Thesis Signature Page

MONTCLAIR STATE UNIVERSITY

Data Mining and Predictive Policing

By

Chanté L. Stewart-Wallace

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of


Master of Sciences

May 2019


College of Science and Mathematics

Thesis Committee:


Department of Computer Science


Dr. Katherine Herbert

Thesis Sponsor


Dr. Bharath Kumar Samanthula

Committee Member


Dr. John Jenq

Committee Member

DATA MINING AND PREDICTIVE POLICING

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

By

CHANTÉ L. STEWART-WALLACE

Montclair State University

Montclair, NJ

2019

Table of Contents

Abstract	1
Thesis Signature Page	2
DATA MINING AND PREDICTIVE POLICING	3
Table of Contents	4
1 Introduction	6
1.1 Motivation	7
1.2 Research Goal	9
2 Literature Review	9
2.1 Assumption Issues and Algorithms	10
2.2 Active Technologies	12
2.3 Ethics and Human Impact	16
3 Datasets	18
4.2 San Francisco Crime Dataset	19
4.3 Chicago Crime Dataset	20
5 Methodology	22
5.1 Data Processing	23
5.1.1 Data Reduction	23
5.1.2 Data Cleaning	24
5.1.3 Data Integration	24
5.1.4 Data Transformation	24
5.2 Data Analysis	25
5.3 Model Construction	29
5.3.1 Apriori Algorithm	30
6 Results	30

6.1 Hotspots	30
7.1 Demographics	34
8 Adaptation Proposal	36
8.1 Application Implementation	37
9 Conclusion	41
References	41

1 Introduction

There is always a pressing demand to process faster, solve sooner, and to respond immediately to the daily strifes life may present. It is seen everywhere and everyday from personalized coupons printed at the end of a consumers transaction, to the bidding of which advertisements will pop up on a users browser at the next click. Big data problems arise from and are driven by the desire to predict, giving users something they want before knowing they want it. Where ever there is socio-economic development, there is criminal activity that is diminishing the overall advancement and security of its region. With that in mind, our governments and law enforcement organizations are perturbed with the demand to change how criminal activity is approached. Over the last 10 years, law enforcement professionals began exploring diverse technologies that offer advanced assistance in crime analysis. In the attempts to no longer be blinded by the trends of transgressions, these technologies aim to study the behavioral patterns that associate with certain crimes, as well as recognize signals that can lead up to similar situations.

Some crimes are random and hard to track. It is apparent that crimes like arson and burglary are on the decline while more premeditated or systematic crimes such as gang rape, murder, and sexual abuse are growing. It would be unrealistic to state that one can predict every victim of every crime but it is feasible to make a speculation from collected data because certain regions have concentrations of particular crimes. With this knowledge, patrols can be effectively dispersed to catch or prevent crimes before they have a chance to mature.

Machine learning agents that have been fostered through Artificial Intelligence and Data Mining have been working with fixed datasets and an array of procedures to find the similarities in data. Predictive analysis is based off of data collections from previous reports that create a probability of what is expected. That makes these machine learning agents a desirable tool in the assessment of event anticipation. Strategic patrolling is already a practice within law enforcement agencies to better maximize the use of resources. With the assistance of these diverse machines, the data from former police reports can be analyzed to produce hotspots that are made apparent from time, type, and location of prior incidents. This method boils down to classification which is useful in many forms of analysis.

In this research, the data types that are used to produce regional maps as well as assess the impact the software has on the productivity of law enforcement agencies will be explored. In addition, the study will briefly look through the lens of those that are aware of active “predictive policing” in their areas. It is important when assessing subjects concerning data mining, that the software is both mutually welcomed and considered ethical while maintaining effectiveness.

1.1 Motivation

Due to the recent and historical tension between authority and citizens, the interaction between the protector and those in need of protection has been blurred. Law Enforcement need solutions with minimal damage, while citizens desire protection without running the risk of being classified as a threat. There is an appeal in using Machine Learning Agents because it is designed to draw its conclusions from concrete data. An officer with the same intentions to

identify a trend may find it difficult to not consider all the data. The predictive policing softwares uses time, type and location, while an officer might also keep those data types in mind, they also might profile individuals for characteristics.

Machine learning agents that have been implemented for policing crime analysis are designed to not have an opinion when processing information. While this is crucial for the integrity of the product, information collected and supplied to the machine seem to show a pattern of bias, which has resulted in the opposition of the product by citizens. The problem arises in the embodied data that can be considered an attack on privacy as well as ethics. There is doubt that the data given to the machine can be removed of its bias. Understanding law, crime, and ethics seems to never be a black and white situation and should not be treated as such. A learning agent might not be as beneficial to the overall resolution that law enforcement is in need of finding.

The intention of this research is to find clarity in the purpose and effectiveness of predictive policing. Using analytical tools and finding understanding in the data as well as the algorithms that organize the data types into hotspot maps. A stimulant of a predictive policing program will be built and tested, the maps and graphs that result should be consistent in depicting trends of sequential crimes. Furthermore, the paper will help identify specific data and how it will hold this operation to ethical standards while still carrying out its purpose.

1.2 Research Goal

Being that there is always crime going on all over the world, it is best that the government and law enforcement agencies have the ability to deal with situations both effectively and ethically.

Not only would that give the people patrolling an upper hand on the regions they are securing, it will also create an understanding behind why these crimes happen in the first place.

With my research I hope to find the answers to the following questions:

- How does an agency use data that has been analyzed to implement the features of predictive policing?
- Is a machine learning agent designed to produce crime patterns from previous incidents truly beneficial when the information it is analyzing has a questionable bias?
- What are the restraints that predictive policing face and are the limitations a drawback from the intent of the product?
- Does predictive policing effectively disrupt crime in the long term ?

2 Literature Review

Growing knowledge in crime patterns as well as finding the causes fostering criminal activities has been a primary focal point for law enforcement agencies. There is an increasing belief that with better comprehension of offenses, information can be obtained to find patterns in criminal behavior to give law enforcement foresight of illegalities to come. This concept, honed the

strategy of predictive policing, a development that is designed to push law enforcement from the common practice of only responding to a crime after it has happened. The plan for the implementation of this is to get disrupt or avoid the crime before it has a chance to come to fruition.

This review is composed of 19 research articles and papers between the years 2012-2018 that analyze the makeup of predictive policing and how it directly affects the society it is utilized in. A majority of the collected readings are based on data in the United States although there are a few mentionings of Canada and the United Kingdom. The analysis taken from these articles will concentrate on specific software, strategies behind them, implementation and utilization, as well as social impact, both positive and controversial. The review has been divided into subsections that will aid in the readers intake of important information. The 3 sections are: algorithms backing prediction, review of actively used technologies, ethical boundaries and the impact police in adverse situation have on data.

2.1 Assumption Issues and Algorithms

This section of the literature review uses two papers that focus on the issues and assumptions surrounding the algorithms that bring life to the machine learning agents. Predictive policing companies have made many claims about their products, but it remains a fact that there is very few formal evaluations published and accessible to the public [1]. There has been no arguments supported enough to claim that the new data-driven policing agents lack possible results. There

has been a consistent mention that the data allowed to be used has conflicts that can breach the accountability of law enforcement in the stages of process and decision making.

One of the assumptions in defense of using algorithms is that data from the can accurately reflect what is bound to happen in the future of the real world. Depending on the context, there can be a degree of continuity found from historic crime patterns, but this is only true for crimes such as burglary while crime like kidnapping would have a harder time falling in this assumption[1]. Inherently, there are many things that can affect continuity which can change the degree of accuracy the a machine learning agent can provide. Policy changes, social views, and the manner in which cultures evolve can all change current reality to what it was historically. With that in mind, the assumption that the past will model the present loses its validity.

Another assumption commonly associated with the use of algorithms is that data analytics can not discriminate without just reason. Whereas the objective for predictive policing is to discriminate against locations and individuals based on the data designed to identify crime probable differences. “If predictive policing identifies a correlation between feature X and probability of offering, in what circumstances is it unjust to treat a person with feature X differently?” [1]. It is without doubt that this is a primary question and conflict with the implementation of predictive policing software. These algorithms supply law enforcement with times, locations, and characteristics that set claim that this will be the next lead for criminal activity. How does one decipher a boundary for utilizing such a tool, without labeling a class of people unjustly?

For those task forces already implementing predictive policing algorithm, there has been question of how much change policing will undergo. Traditional policing embodies involvement with the neighborhoods that the officers patrol, which creates a trust and also gives officials leverage to be able to intercept crime in youth. While data-driven policing creates a complete shift in police work. It's less about the people and more about the numbers. "Police officers are driving through areas predicted to have a high crime chance to scan whether the front doors have bad locks instead of stopping to talk to families behind those doors"[1].

In addition, law enforcement has made it a point to utilize the software to aid in the deployment of police. "Police response in a hotspot policing approach tend to be pre-packaged, cookie-cutter reactions rather than tailored, researched strategic plans for solving or eliminating the the problem over the long haul" [1]. Some departments prefer to use the data to focus strictly on the mobilization of police patrols and respectively decline its input when it comes to understanding why crime happens. Reason being, "some officers have knowledge not captured by the data (as where they know the data they themselves enter into the system are flawed or incomplete) and may thus be less inclined to trust the forecast"[1].

2.2 Active Technologies

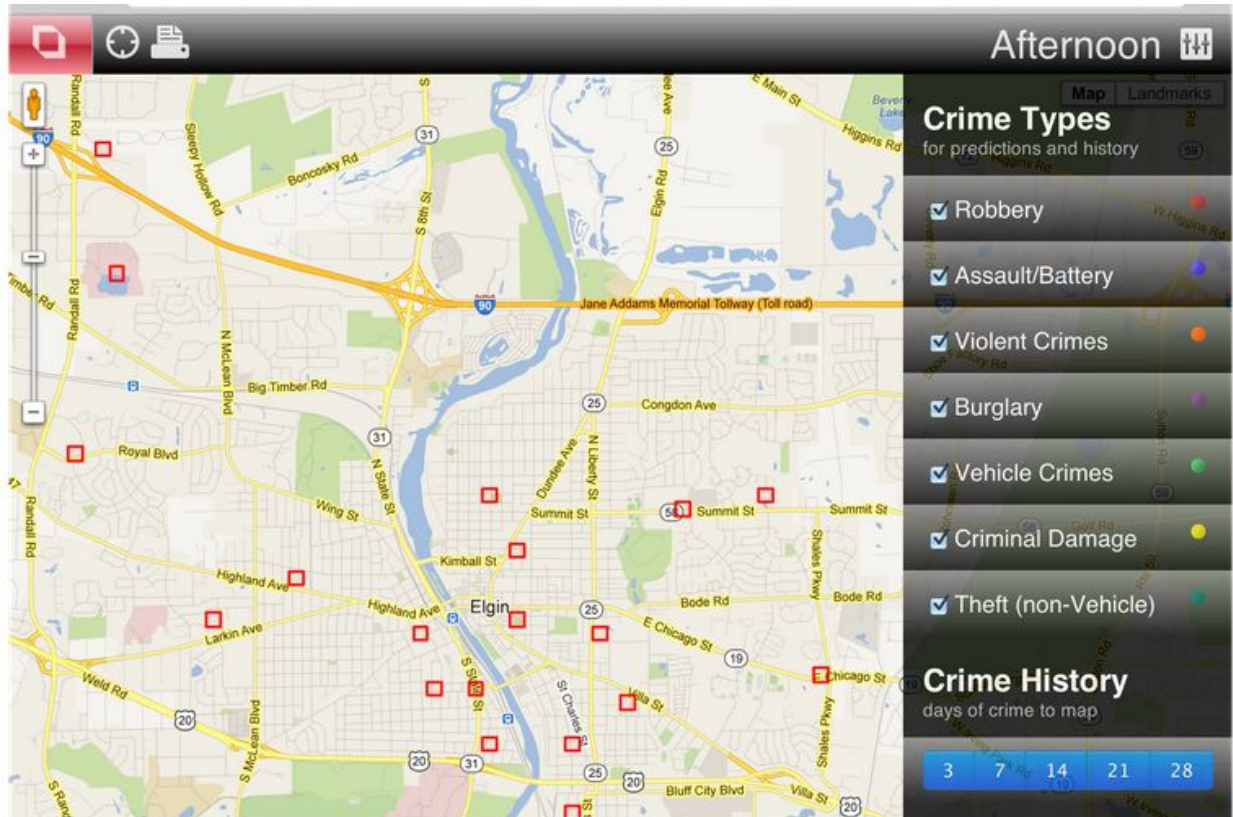
Upon considering the effects, both positive and negative of predictive policing, this literature review zones in on active technologies that are being utilized in is some of the dense and

crime-ridden communities. It has been established that the approaches of predictive policing fall in the following methods:

1. Predicting places and times of crime.
2. Predicting offender and pinpoint individuals that a probable to commit crimes.
3. Predicting the identity of perpetrators.
4. Predicting those who may become victims to crimes.

Although there are so many perspectives of how to execute, most predictive policing software will find itself in the first two categories [2]. Particularly in the United States and Europe, geospatial crime prediction will be seen which leans towards the first category. It is a method that has been prominent since the 1960's and the research surrounding it shows that that crime location is not a random occurrence and can be transfigured into strategic analysis and planning for the distribution of resources. The second category has collected some traction but has not yet been openly adopted. "To calculate the likeliness that a given person will commit a crime or is prone to behavior that puts others at risk" sounds like a dream come true to many law enforcement perspectives, but such programs have been put "under high scrutiny by privacy and human rights advocates" [2].

Figure 1: Screensnap of Predictive policing software highlighting prediction hotspots



PredPol is one of the well talked about and easily identifiable technologies catered to predictive policing. UCLA conducting a study on its effectiveness and found that “after 4 to 8 months, the study revealed that the areas assigned by the algorithm and patrolled by the officers, had reduction of 7.4%, while the analyst without the predictive model predicted 2 crimes a week” [3]. If PredPol maintains these types of numbers, it has been calculated that the utilization of the software could result in the LAPD saving around nine million dollars a year on average. What is really interesting about how PredPol, is that the software branches from an algorithm used in seismology. Just as an earthquake is expected to have an aftershock, serial crimes are expected to repeat in waves within a short amount of time and in close proximity[2].

HunchLab which was developed by Azavea has major similarity to PredPol including appearance and overall goal. What makes it different is that has integrated approaches like Risk Terrain Modeling to further develop the credibility of the results. The idea behind it is to divide map layers by different representations. For example, one layer can be the influence and the other can be intensity of a crime. Once those are established, the layers are combined and used to produce a map that can show values that display the probability of every crime in the area under analysis. Since repeat theory is focused on endogenous factor like behavior pattern, adding Risk Terrain Modeling inserts exogenous factors that considers things like landmarks and spacing [2]. An example of something that can be useful to detect on HunchLab would be prostitution. We know for prostitution to be successful, it would have to take place in areas that allow drivers to reduce speed near bars and party spots. These factors can be constructed into different layer that specify bars, nightclubs or banks [2].

Chicago's Heat List seems to be one of the more invasive versions of predictive policings. The police department analyses the networks or previously arrested individuals to calculate the chances of someone in their network being involved in major crimes. It focuses on the relevance a social network can have but concludes no ideas of what crime might be committed. In addition to creating a list of likely people, the software compiles a list of influencers. If someone appears on this list, it means that the person can have some sort of effect on a individual found on the heat list. Once these people are established, certain ones are sent notifications from the police department with a warning that they may end up facing charges if the continue engaging in criminal activities [2]. One can only imagine how it feels to receive a letter of possible charges

before even committing said crime. The community is asking for transparency in how the police department is coming to these scores but they have remained stern in declining to release the details surrounding the algorithm. “The most significant characteristic for computing an S.S.L. risk score is the age of a potential victim or offender. For every decade of age, the risk score declined by about 40 points. Practically speaking, this variable limits the list to young people: No one older than 30 falls within the highest-risk category with a score at or above 480” [4].

Another big concern people have towards the algorithm is that the numbers are not matching the long term logic of systematic crime. The scores were showing that “victims of assault and battery or shooting were much more likely to be involved in future shootings. Arrest for domestic violence, weapons or drugs were much less predictive. Gang affiliation, which applied to 16.3 percent of people on the list, had barely any impact on the risk score” [4]. The numbers were showing the opposite of what has been proved to be predictable over time. The algorithm has been updated many times but many are still uncomfortable with its operation.

2.3 Ethics and Human Impact

The literature review identifies three articles that mainly focus on the ethically line that predictive policing has been playing with. In addition these reading along with all of the other papers resourced for this research has mentioned the extensive concern there is a bias in the data [5-7]. This is believed to a start to creating further tension between the officials utilizing the software and the citizens in the areas that have been identified in hotspots. While highlighting these points, numbers do support that the implementation of the softwares can be increasing effectiveness of law enforcement without costing departments additional money.

Law enforcement agencies that have been actively testing these programs have emphasized that the intention is to push policing to a more proactive process versus the current approach which tends to be reactive. At the same time, the claim is that predictive policing is not designed to be a substitute for real police knowledge and experiences [5]. These statements are reassuring and seem of goodwill. It is apparent from agency feedback that the majority of law enforcement that work in the field favor the conceptual promises that machine learning agents can present.

It is essential for the sake of societal coherence that there is a way to apply and enforce law. It would be unrealistic to say that policing is a 'one size fits all' for a country or even a state [6]. It is no coincidence that cities with some of the highest crime rates such as Los Angeles, Chicago, and New York City have been the first to implement and test predictive policing software. In the eyes of law enforcement, predictive policing is a tool which can save lives, giving these programs a moral responsibility to be put to use [6].

With that in mind, the articles also consider the lens of the opposition, which includes but is not limited to researchers and citizens living in hotspot labeled areas. Predictive policing should not alter how policing is done, it should change efficiency. With that intention in mind, those opposing the implementation argue that, "there can be a placebo-like effect. The simple fact that data exists and officers have access to it means that they are more likely to change their behavior and the way they police" [5]. A change in method results in a change of mindset. Being that this technology is so new, there are no policies and procedures that adequately build trust within the

communities of how it is being used. There is worry that predictive policing could put neighborhoods on continued armed patrol while also reinforcing a temperament towards people due to the bias data that is being shared with the machine.

In addition, if these programs are considered to be putting people under surveillance, it would be a violation of privacy, process and civil liberty [5]. These hotspots point to locations of overpopulation and poverty, which not be coincidence, is dense with numbers of minorities. Predictive policing does not point to areas well spaced out and majority caucasian. Caucasians commit crimes yet on the radar of the predictive policing, there is little to no interference our prevention in crimes committed by them [6]. The data is skewed to a point where it models the historically political and racial climate in America. For predictive policing to be better received by citizen, there must be policy in place and complete transparency which so far has not been the case across the board [5].

3 Datasets

In this study, there are 2 different datasets pulled from the open data platforms of two cities in the United States. The cities used are the following: San Francisco in California and Chicago in Illinois. To establishing the data models for this study, focus was put on the Chicago dataset. After the construction of the data models were set, the same methods were applied to remaining datasets. This was to find any trends that expand past the bounds of State to State as well as evaluate the impact of demographics. This section is used to give a brief of the findings of each dataset.

4.2 San Francisco Crime Dataset

This dataset reflects the actual crime reports in San Francisco, California. Included in the set is criminal offenses and incidents during the calendar year of 2018. This data is pulled from the Open Data Online portal which is shared and maintained by the San Francisco Police Department. The dataset originated with 26 attributes and 155183 instances before it was put through the Data Processing specified in Section 5.

Table 1. San Francisco key attribute table

Attribute	Data Type	Number of Distinct Values	Value
Crime_Date	Date	Unlimited	mm/dd/yyyy
Crime_Time	Time	Unlimited	hh:mm
Crime_Day	Nominal	7 Categories	Monday Tuesday Wednesday etc.
Crime_Type	Nominal	20 Categories	Burglary Larceny Theft Robbery Assault Motor Vehicle Theft Other etc.
Crime_District	Nominal	42 Names	(See Figure 2)



Figure 2: Map of San Francisco Neighborhoods

4.3 Chicago Crime Dataset

This dataset reflects the actual crime reports in Chicago, Illinois . Included in the set is criminal offenses and incidents during the calendar year of 2018. This data is pulled from the City of Chicago Data Portal which is shared and maintained by the Chicago Police Department. The dataset originated with 22 attributes and 266297 instances before it was put through the Data Processing specified in Section 5.

Table 2. Chicago key attribute table

Attribute	Data Type	Number of Distinct Values	Value
Crime_Date	Date	Unlimited	mm/dd/yyyy
Crime_Time	Time	Unlimited	hh:mm
Crime_Day	Nominal	7 Categories	Monday Tuesday Wednesday etc.
Crime_Type	Nominal	17 Categories	Theft Deceptive Practice Robbery Battery Burglary Crim Sexual Assault Other etc.
Community Area	Nominal	78 Areas	(See Figure 3)

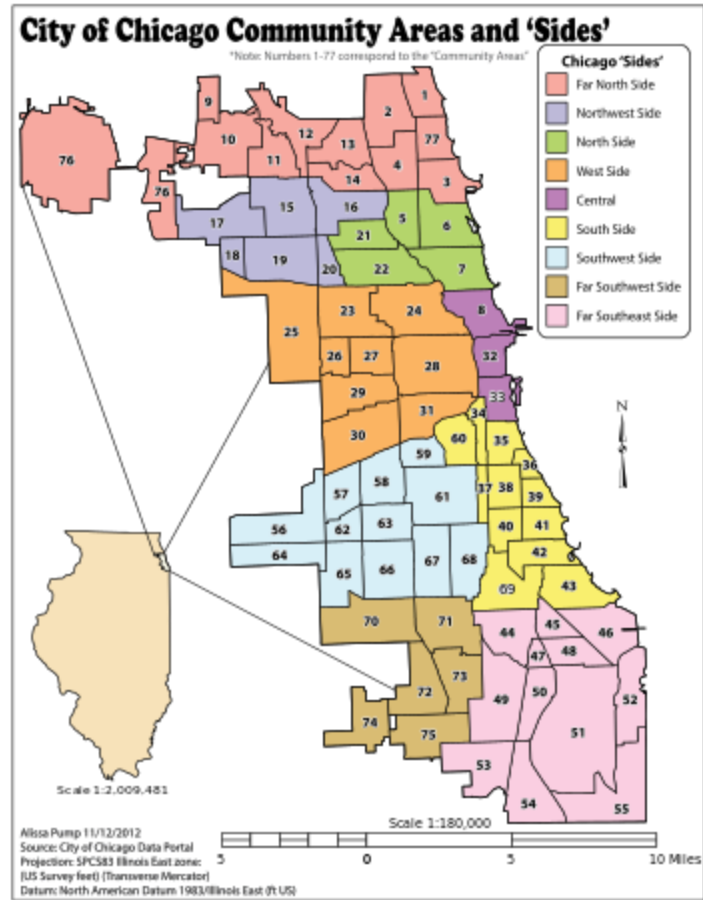


Figure 3: Map of Chicago Wards

5 Methodology

I firmly believe that finding the deep relationships between crime and the environmental neglect could help impact how predictive policing software is utilized. Currently it runs with the intent to interject before crime has the chance to happen. I find that this idea has just scratched the surface and has not been expanded to a fraction of its potential. There is a need for crime reduction while maintaining transparency and trust with general public. I attempt to extract thought-provoking

patterns found in the crime variables to understand which community planning or geographic additions could benefit areas and create a diffusion in high risk areas.

In this section, it is explained how the datasets were arranged. Then, there is an analysis of the data, followed by the data-mining models used to attain this papers motivation.

5.1 Data Processing

While working with the data, it was put through the following data processes:

5.1.1 Data Reduction

After taking a look at the data, it was apparent that data reduction should be applied to current datasets. Instead of taking the many versatile and repetitive attributes found in each set, what was utilized universally in each dataset in this study was cut down to four, which was Crime Type, Date, Time, and District or Neighborhood. All other data that was not beneficial and aiding in the goal of the study was removed from the datasets.

After that was applied, a data reduction was applied to the overall instances, When looking at the datasets, it was noticed that traffic tickets and car accidents were included. The attribute crime type was used to filter out and remove the listings that were not crime related since they served no purpose for the goal of the study. Once that was applied we were left with 225,554 instances for Chicago, Illinois and 76,048 instances for San Francisco, California.

5.1.2 Data Cleaning

It was discovered that there were blank and missing values scattered throughout the datasets. However, the attributes under question had no effect on the key attributes used for the study. As a result, the datasets did not have to go through a data cleaning stage. The attributes that are used are cleaned and have no inconsistencies that were noted.

5.1.3 Data Integration

The first step of data integration applied to the dataset in the study, was making adjustments to the names of the attributes. It is in the best interest of the research that the attribute names are not conflicting, so the key attribute names were changed to the following: Crime_Type, Crime_Date, and Crime_Location. For the sake of the mining involved in the study that demands analysis of different gradients of time. The Crime_Date attribute was expanded to create three more attributes: Crime_Month, Crime_Day, and Crime_Time. In terms of time, only the hourly was considered as doing minute by minute would not give us a great span to identify patterns. All the times were converted to military time for every dataset.

5.1.4 Data Transformation

At the conclusion of the integration we were left with 24 values for each hour in Crime_Time and types in Crime_Type. To get a more defined pattern, the data was transformed to reflect

more condensed groups. The Crime_Time was broken into 4-hour intervals. Crime_Type was condensed to six value types.

5.2 Data Analysis

As a vehicle to analyse and get a clearer view of the collected data, statistical analysis was created to reflect the attributes of the datasets. Each city was cleaned through an excel spreadsheet and then loaded up in Jupyter Notebook. Python script was used to find frequencies of the distinct values in the attributes used for the study. The graphs display percentages of occurrences based of the aspect under analysis.

Figures 4- 6 give a statistical comparison between San Francisco and Chicago crime datasets both taken from the cities respective open data portals. As it is important to keep the work current, both datasets are from the 2018 calendar year. Using the same year between both dataset also establishes consistency in the study. The numbers used are a focused on crime occurrences instead of number of types of crimes committed.

Figures 4 displays the percentage of crime occurrences from January to December in San Francisco and Chicago. The San Francisco dataset does not show any significant peaks or decreases in criminal occurrences from month to month comparison. The Chicago dataset shows significant increases in crime between the summer months of May to August.

Figure 5 displays the percentage of crime occurrences from Sunday to Saturday in San Francisco and Chicago. For both datasets, the statistical analysis shows that both cities seem to be close in consistent for the spread of criminal occurrences. There is a slight peak in numbers for Friday and slight decrease in numbers for Sunday in both datasets which is not unexpected.

Figure 6 displays the percentage of crime occurrences over the 24 hour span in San Francisco and Chicago. For both cities, it appears that the safest time of day is between the hours 4am and 8am. In the case of Chicago, the highest amount of criminal occurrences is reported between Noon to 4pm. San Francisco's highest amount of criminal occurrences is reported between 4pm to 8pm.

Figure 7 and 8 displays the percentage of all crime occurrences over different regions of San Francisco and Chicago. The areas for these graphs were selected to show the range in occurrences and establish some of the safest and most dangerous communities/districts. In San Francisco, McLaren Park appears to be the safest with minimal crime occurrences while Mission appears to be the most dangerous and saturated with crime occurrences. In Chicago, Community 9 appears to be the safest with minimal crime occurrences while Community 25 appears to be the most dangerous and saturated with crime occurrences.

Percentage of Crime over 12 Months (2018)

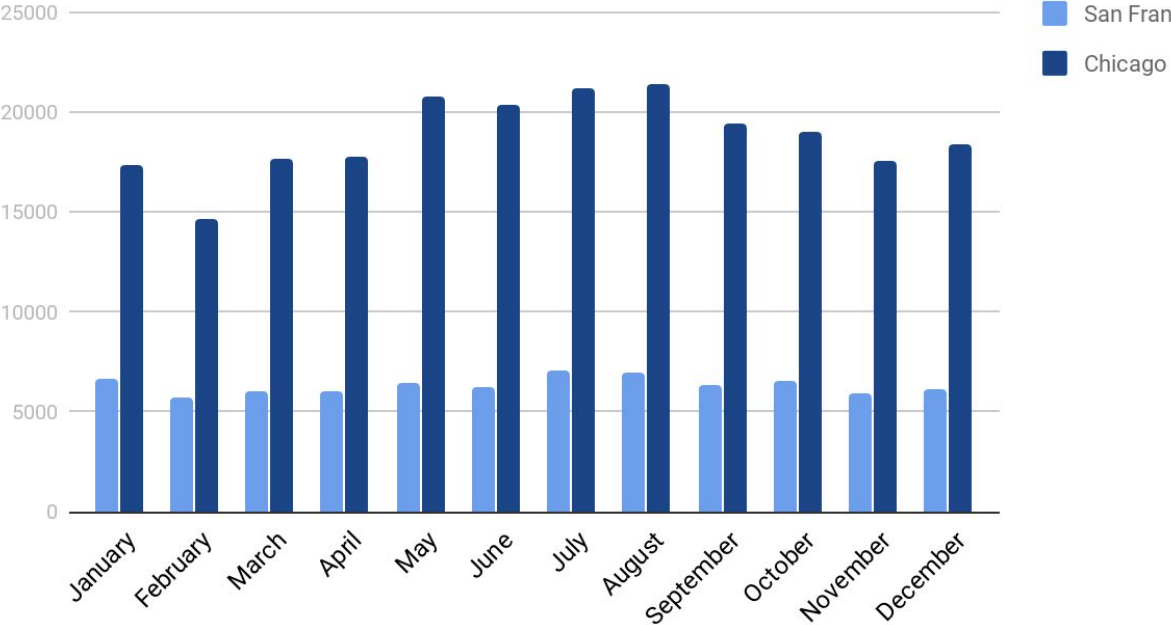


Figure 4: Crime occurrences on a monthly basis in San Francisco and Chicago

Percentage of Crime over the Week (Year: 2018)

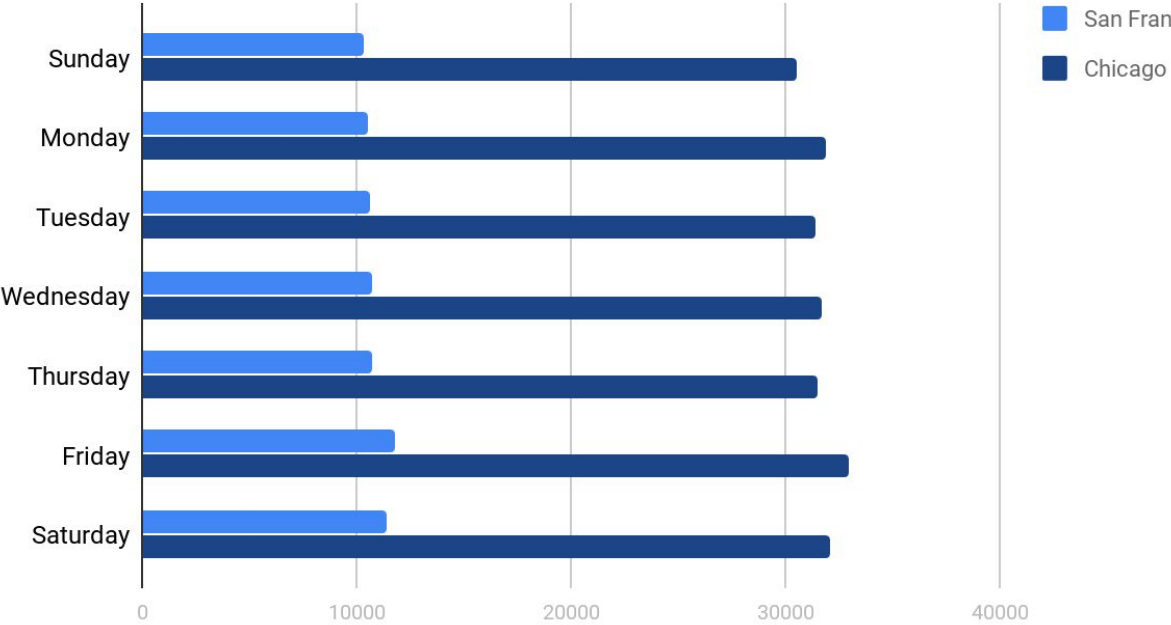


Figure 5: Crime occurrences over the days of week in San Francisco and Chicago

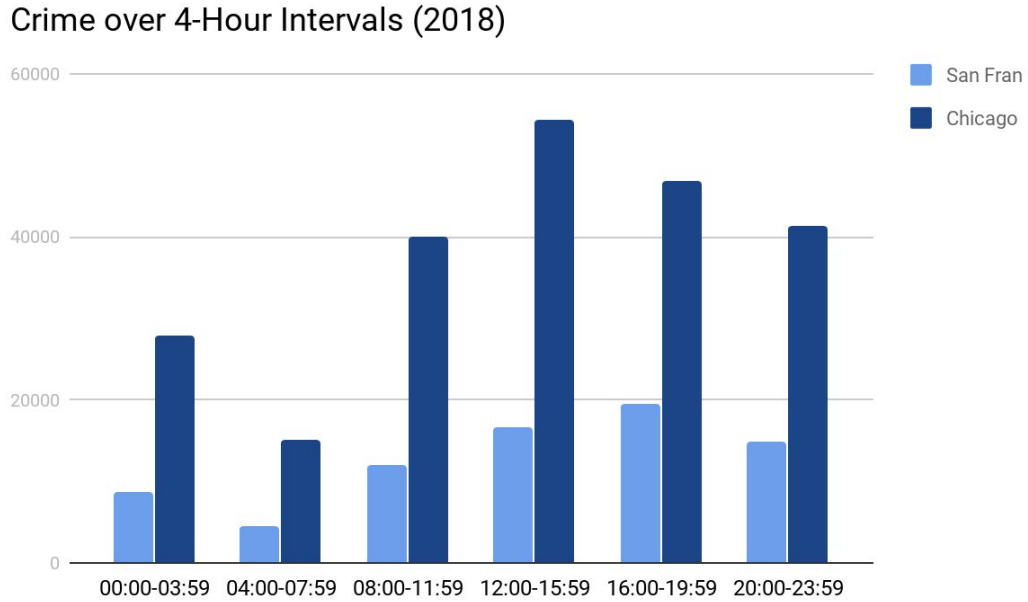


Figure 6: Crime Rate over 4-hour intervals in San Francisco and Chicago

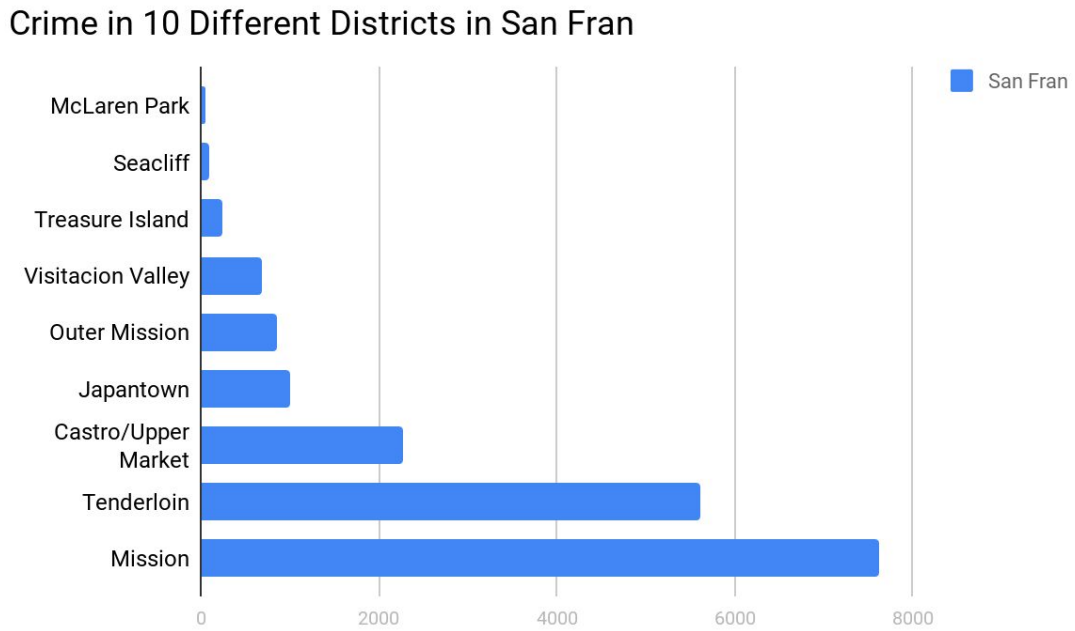


Figure 7: Crime Rate in specific districts in Chicago

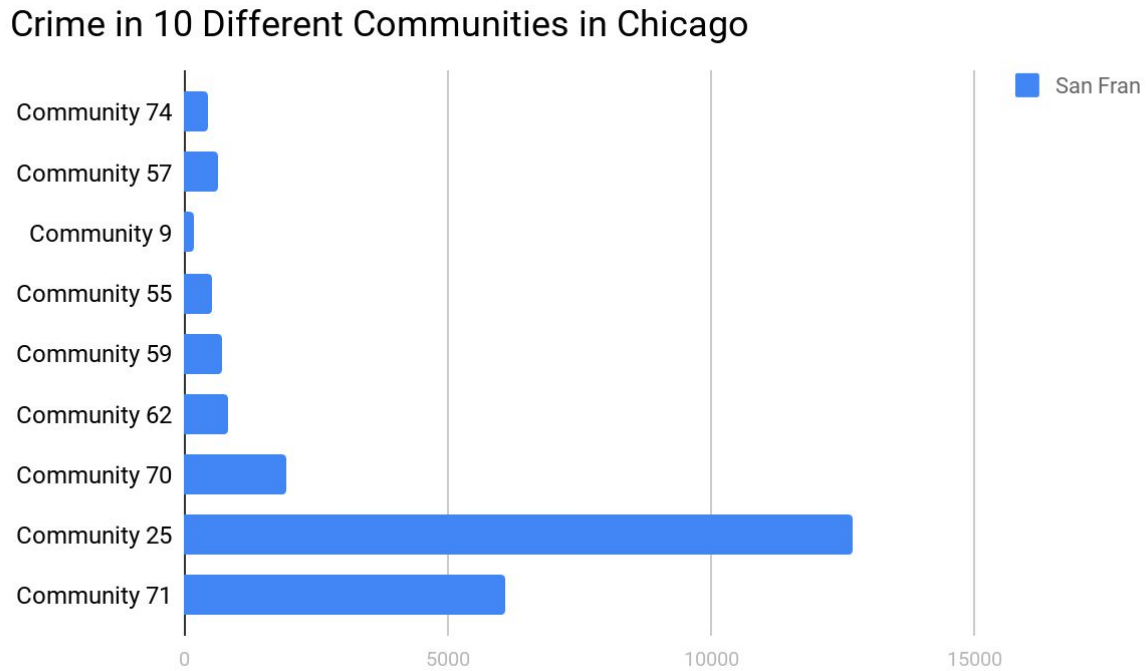


Figure 8: Crime Rate in specific districts in San Francisco

5.3 Model Construction

To pull the frequent patterns from the datasets of San Francisco, California and Chicago, Illinois crimes, the Apriori algorithm was used. These patterns are inherently used to find which combination of time, day, and location need to more heavily patrolled.

5.3.1 Apriori Algorithm

Apriori is a commonly used and fundamental algorithm used for data mining purposes. It reviews the dataset to find supports that satisfy a predetermined minimum. The desired goal was to find all of the crime patterns of high frequency without considering the types.

This model was implemented using python script in the Jupyter Notebook. A series of test were ran ann different minimum supports were applied to each dataset.

6 Results

In this section, the key results taken from the use of the Apriori algorithm on the datasets. Then, the information is combined with the demographic findings.

6.1 Hotspots

A primary goal of this research was to find a understanding in how predictive policing agencies form hotspots and how do they optimize task deployment. By applying the Apriori algorithm to the San Francisco and Chicago dataset, a support number was generated. In simplest terms, the support numbers were determined by using the formula $Frequency \div Total\ occurrences$. for the use of deployment, agencies will create a minimum support which would be the the number right above average frequency. In the case of both Tables 3 and 4, the unfiltered supports are

displayed. If a agency was looking for higher frequency patterns to patrol they might go for a min support of 0.0012 vs the low crime rated areas with a support number of 0.0001.

Table 3 shows an array of Frequent patterns found in the San Francisco dataset. As previously established through statistical analysis, it was determined that the Mission District is more likely to have criminal occurrences while districts like Seacliff is not. On the table, both districts are highlighted at the same time fame and day. The support numbers reflect which space will take precedence in patrol.

Table 4 shows an array of Frequent patterns found in the Chicago dataset. As previously established through statistical analysis, it was determined that the Community 25 is more likely to have criminal occurrences while Communities like 9 is not.

Table 3: Apriori Algorithm results for San Francisco

Frequent Pattern	Sup	Frequent Pattern	Sup
T1 Friday Bayview Hunters Point	0.0008	T6 Thursday Bayview Hunters Point	0.0012
T1 Friday Bernal Heights	0.0002	T6 Thursday Bernal Heights	0.0005
T1 Friday Castro/Upper Market	0.0006	T6 Thursday Castro/Upper Market	0.0010
T1 Friday Chinatown	0.0002	T6 Thursday Chinatown	0.0007
T1 Friday Excelsior	0.0001	T6 Thursday Excelsior	0.0002
T1 Friday Financial District/South Beach	0.0011	T6 Thursday Financial District/South Beach	0.0025
T1 Friday Glen Park	0.0000	T6 Thursday Glen Park	0.0002
T1 Friday Golden Gate Park	0.0001	T6 Thursday Golden Gate Park	0.0002
T1 Friday Haight Ashbury	0.0002	T6 Thursday Haight Ashbury	0.0003
T1 Friday Hayes Valley	0.0004	T6 Thursday Hayes Valley	0.0008
T1 Friday Inner Richmond	0.0003	T6 Thursday Inner Richmond	0.0006
T1 Friday Inner Sunset	0.0001	T6 Thursday Inner Sunset	0.0003
T1 Friday Japantown	0.0001	T6 Thursday Japantown	0.0003
T5 Wednesday Mission	0.0033	T6 Thursday Lakeshore	0.0004
T5 Wednesday Mission Bay	0.0009	T6 Thursday Lincoln Park	0.0000
T5 Wednesday Nob Hill	0.0008	T6 Thursday Lone Mountain/USF	0.0003
T5 Wednesday Noe Valley	0.0004	T6 Thursday Marina	0.0008

T5 Wednesday North Beach	0.0016	T6 Thursday Mission	0.0031
T5 Wednesday null	0.0000	T6 Thursday Mission Bay	0.0005
T5 Wednesday Oceanview/Merced/Ingleside	0.0003	T6 Thursday Nob Hill	0.0007
T5 Wednesday Outer Mission	0.0003	T6 Thursday Noe Valley	0.0003
T5 Wednesday Outer Richmond	0.0009	T6 Thursday North Beach	0.0010
T5 Wednesday Pacific Heights	0.0006	T6 Thursday Oceanview/Merced/Ingleside	0.0003
T5 Wednesday Portola	0.0004	T6 Thursday Outer Mission	0.0004
T5 Wednesday Potrero Hill	0.0007	T6 Thursday Outer Richmond	0.0004
T5 Wednesday Presidio	0.0000	T6 Thursday Pacific Heights	0.0006
T5 Wednesday Presidio Heights	0.0003	T6 Thursday Portola	0.0003
T5 Wednesday Russian Hill	0.0011	T6 Thursday Potrero Hill	0.0007
T5 Wednesday Seacliff	0.0000	T6 Thursday Presidio	0.0000
T6 Monday Potrero Hill	0.0005	T6 Thursday Presidio Heights	0.0002
T6 Monday Presidio Heights	0.0001	T6 Thursday Russian Hill	0.0008
T6 Monday Russian Hill	0.0006	T6 Thursday South of Market	0.0022
T6 Monday Seacliff	0.0000	T6 Thursday Sunset/Parkside	0.0006
T6 Monday South of Market	0.0017	T6 Thursday Tenderloin	0.0018
T6 Monday Sunset/Parkside	0.0005	T6 Thursday Treasure Island	0.0001
T6 Monday Tenderloin	0.0015	T6 Thursday Twin Peaks	0.0001
T6 Monday Treasure Island	0.0000	T6 Thursday Visitacion Valley	0.0002
T6 Monday Twin Peaks	0.0001	T6 Thursday West of Twin Peaks	0.0005
T6 Monday Visitacion Valley	0.0002	T6 Thursday Western Addition	0.0006
T6 Monday West of Twin Peaks	0.0004	T6 Tuesday Bayview Hunters Point	0.0011
T6 Monday Western Addition	0.0006	T6 Tuesday Bernal Heights	0.0006
T4 Friday Bernal Heights	0.0005	T6 Tuesday Castro/Upper Market	0.0008
T4 Friday Castro/Upper Market	0.0010	T6 Tuesday Chinatown	0.0004
T4 Friday Chinatown	0.0005	T6 Tuesday Excelsior	0.0004
T4 Friday Excelsior	0.0005	T6 Tuesday Financial District/South Beach	0.0022
T4 Friday Financial District/South Beach	0.0042	T6 Tuesday Glen Park	0.0002

Table 4: Apriori Algorithm results for Chicago

Frequent Patterns	Min	Frequent Patterns	Min
T1 Friday 1	0.0002	T3 Sunday 34	0.0001
T1 Friday 10	0.0001	T3 Sunday 35	0.0002
T1 Friday 11	0.0000	T3 Sunday 36	0.0000
T1 Friday 12	0.0000	T3 Sunday 37	0.0000
T1 Friday 13	0.0000	T3 Sunday 38	0.0003
T1 Friday 14	0.0002	T3 Sunday 39	0.0001
T1 Friday 15	0.0002	T3 Sunday 4	0.0001
T1 Friday 16	0.0001	T3 Sunday 40	0.0002
T1 Friday 17	0.0001	T3 Sunday 41	0.0001
T1 Friday 18	0.0001	T3 Sunday 42	0.0003
T1 Friday 19	0.0003	T3 Sunday 43	0.0007
T1 Friday 2	0.0002	T3 Sunday 44	0.0005
T1 Friday 20	0.0001	T3 Sunday 45	0.0001

T1 Friday 21	0.0001	T3 Sunday 46	0.0003
T1 Friday 22	0.0003	T3 Sunday 47	0.0000
T1 Friday 23	0.0004	T3 Sunday 48	0.0001
T1 Friday 24	0.0005	T3 Sunday 49	0.0005
T1 Friday 25	0.0008	T3 Sunday 5	0.0001
T1 Friday 26	0.0003	T3 Sunday 50	0.0001
T1 Friday 27	0.0002	T3 Sunday 51	0.0001
T1 Friday 28	0.0005	T3 Sunday 52	0.0001
T1 Friday 29	0.0005	T3 Sunday 53	0.0004
T1 Friday 3	0.0002	T3 Sunday 54	0.0001
T1 Friday 30	0.0002	T3 Sunday 55	0.0001
T1 Friday 31	0.0001	T3 Sunday 56	0.0001
T1 Friday 32	0.0003	T3 Sunday 57	0.0001
T1 Friday 33	0.0001	T3 Sunday 58	0.0002
T1 Friday 34	0.0001	T3 Sunday 59	0.0001
T1 Friday 35	0.0002	T3 Sunday 6	0.0004
T2 Saturday 42	0.0001	T3 Sunday 60	0.0001
T2 Saturday 43	0.0004	T3 Sunday 61	0.0003
T2 Saturday 44	0.0003	T3 Sunday 62	0.0001
T2 Saturday 45	0.0000	T3 Sunday 63	0.0001
T2 Saturday 46	0.0002	T3 Sunday 64	0.0001
T2 Saturday 47	0.0000	T3 Sunday 65	0.0001
T2 Saturday 48	0.0000	T3 Sunday 66	0.0004
T2 Saturday 49	0.0003	T3 Sunday 67	0.0006
T2 Saturday 5	0.0001	T3 Sunday 68	0.0005
T2 Saturday 50	0.0000	T3 Sunday 69	0.0004
T2 Saturday 51	0.0000	T3 Sunday 7	0.0003
T2 Saturday 52	0.0000	T3 Sunday 70	0.0002
T2 Saturday 53	0.0002	T3 Sunday 71	0.0006
T2 Saturday 54	0.0001	T3 Sunday 72	0.0001
T2 Saturday 55	0.0000	T3 Sunday 73	0.0002
T2 Saturday 56	0.0001	T3 Sunday 74	0.0000
T2 Saturday 57	0.0000	T3 Sunday 75	0.0001
T2 Saturday 58	0.0001	T3 Sunday 76	0.0001
T2 Saturday 59	0.0001	T3 Sunday 77	0.0002
T2 Saturday 6	0.0002	T3 Sunday 8	0.0007
T2 Saturday 60	0.0001	T3 Sunday 9	0.0000
T2 Saturday 61	0.0001	T3 Thursday 1	0.0004
T2 Saturday 62	0.0001	T3 Thursday 10	0.0001
T2 Saturday 63	0.0001	T3 Thursday 11	0.0001
T2 Saturday 64	0.0000	T3 Thursday 12	0.0001
T2 Saturday 65	0.0001	T3 Thursday 13	0.0001
T2 Saturday 66	0.0003	T3 Thursday 14	0.0002

7.1 Demographics

After the original goal was met to locate hotspots and their concentration, the study shifted towards the demographics that compose the areas in question. Table 5 and 6 show a population breakdown of certain areas in the two cities being reviewed. It was found that the spatial hotspots that had higher support number have significantly larger population, high density of housing units, and were majority non-white identified. In addition, these areas plagued with high crime occurrences, are commonly placed in food deserts with higher poverty rates.

Another noteworthy demographic, the areas considered more dangerous have a higher percentage of the population between that ages 20-29 and a higher percentage of males. In contrast, the safer areas have a higher population of individuals between the ages 50-59 and a higher percentage of females.

Table 5: Population breakdown of 6 San Francisco Districts in 2017

Population of 6 San Fran Districts in 2017

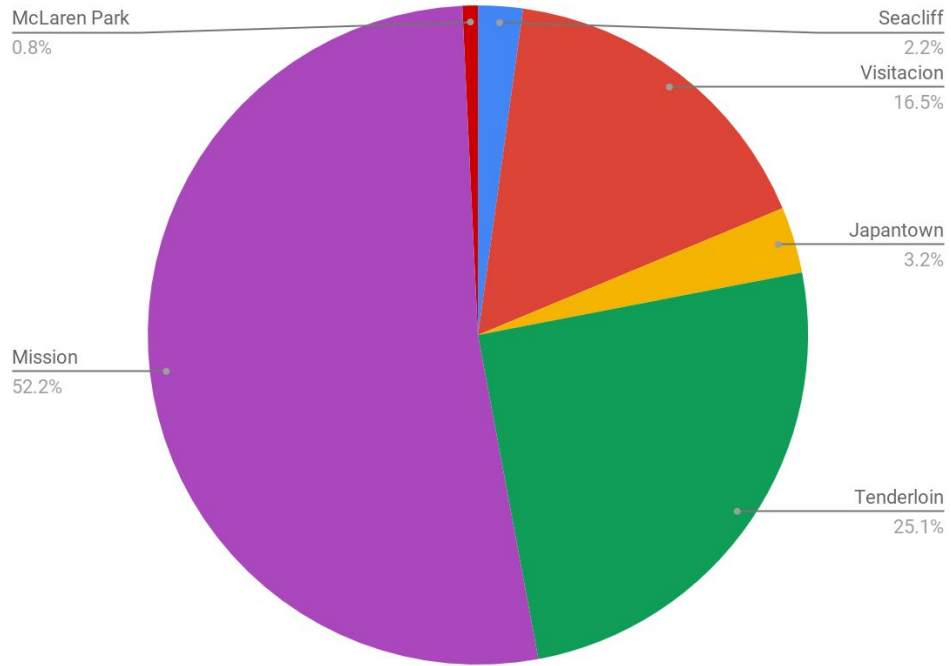
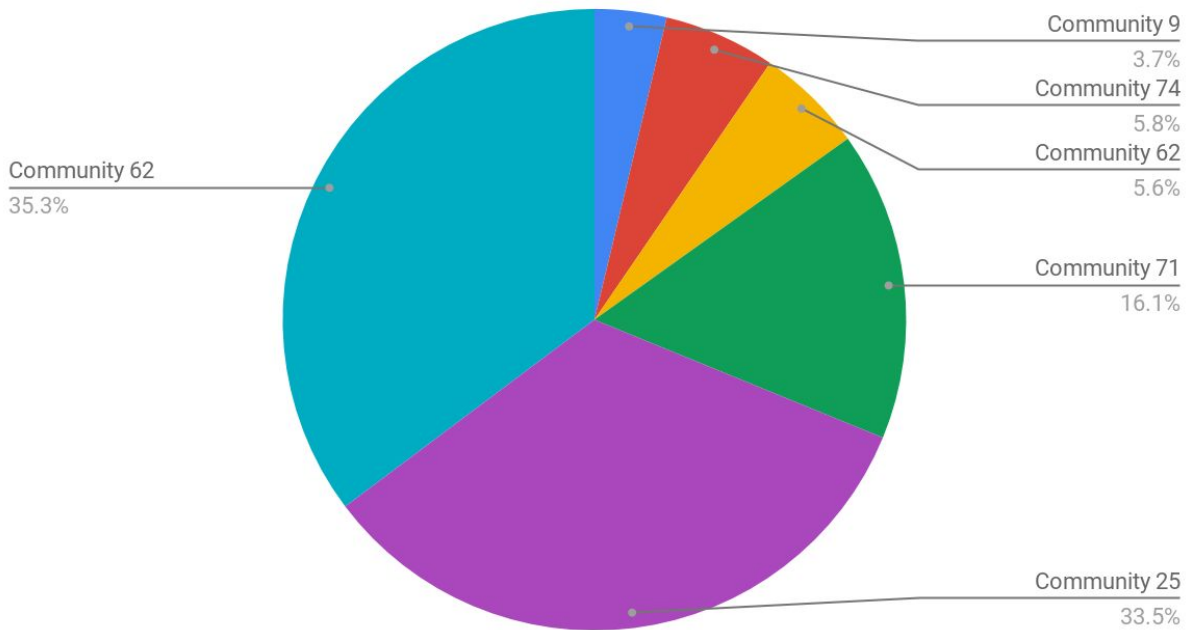


Table 6: Population breakdown of 6 Chicago Districts in 2017

Population of 6 Chicago Communities in 2017



8 Adaptation Proposal

As previously established through the study, it is to be placed in a cookie cutout and result in a universal solution. There is lack of thought when it comes to strengthening the relationship between law enforcement and the communities they serve. If one was to look at the hotspot data pulled from this study, the data will highlight the concentrated areas that have also been trialed with socioeconomic affliction

These areas are overpopulated, underfunded, neglected, and lacking the basic necessities to produce a quality of life that can lead to law abiding citizens. The data analytics have been showing all the diverse ways a community has and can be failed. With this in mind, how can the police department's stop or diminish criminal activity when the crime is a direct response to survival and the toxic environment people are exposed to?

The first proposed step is to, pull back less from the predictive accusations and lean more towards preventative care. Instead of using the algorithm and data learning agents to create a profile for a possible criminal, use the spatial and temporal data to determine which housing units or areas have are at risk. For example, first floor apartments/homes are easier to have a break-in through windows. How can we prepare occupants in high risk areas to slight there chances? Window alarms, updated locks, personal surveillance, etc.

To build from that, how can this learning agents push for a trusting relationship between the law enforcement and the people it serves. It is the idea of ‘sheep and shepard’, that the people are being lead from good intention and feel safe. It is reported that police departments actively using a predictive policing software have saved millions of dollars since its implementation. This money can be used to put back into the communities that have been overall left behind to fend for themselves. Not only would, investments towards the community from the law enforcement create a positive impact, it can result in the long term decrease in crime.

It has been reported that the power of green spaces has transformed communities drastically. Different types of green spaces have different effects on crime. If there can be a dataset created for types of green spaces and there radius effect. Machine learning agents can create hotspot maps and add layers similar to HunchLab. For example, the layers can go as followed:

1)Neighborhood map, 2)Hot spot map, 3) Monuments or geographical landmarks, 4) Food Desert Map, 5)Housing Density.

Once that map is layered up, spaces that have high concentrations of crime, density, and food deserts can look for triangular rends and have a green space or community planning in the middle to disrupt the regular flow of the space

8.1 Application Implementation

This section will provide a discription for the web based application created in accompanment with the research and proposed adaptation.

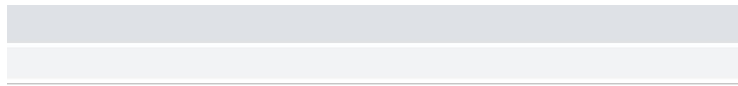
Figure 9 is the homescreen that the user can use to direct themselves to the desired portal. The options in the drop down window are: Law Enforcement, Community Planning, and Resident. Once selected, the user will be directed to their page.

Figure 10 is the Community Crime Reduction homepage. This window gives a heatmap for the user to visualize the distribution of criminal occurrences in the area and tools to work with for understanding. From there, the user can use a drop down window to pick a specific community to analyze.

Figure 11 is the analysis screen of Community 71. It is similar to the page before but is centralized to the specific area. In addition, the right side of the page has an analysis break down pulled from the data so that the user can better understand time and days that the community can most benefit from different community planning.

Figure 12 is the analysis screen of Community 71 with the filter of only viewing robbery crimes. Instead of a heat map, it displays the individual plotted occurrences. In addition, it also creates circles to represent the radius walking of students in the area after school. From that the map draws a shape that contains a space with a concentration of occurrences and schools nearby the recommend the most beneficial greenspace for that area.

Figure 9: Home Screen to Portal



Welcome to the Chicago Crime Prediction Portal

Please select the desired User Portal

Figure 10: Home Screen to the Community Portal

Welcome to the Community Crime Reduction Portal

Below is a Heat Map showing the spatial distribution of crime in the Chicago area in 2018.

Please Select the Ward for Analysis

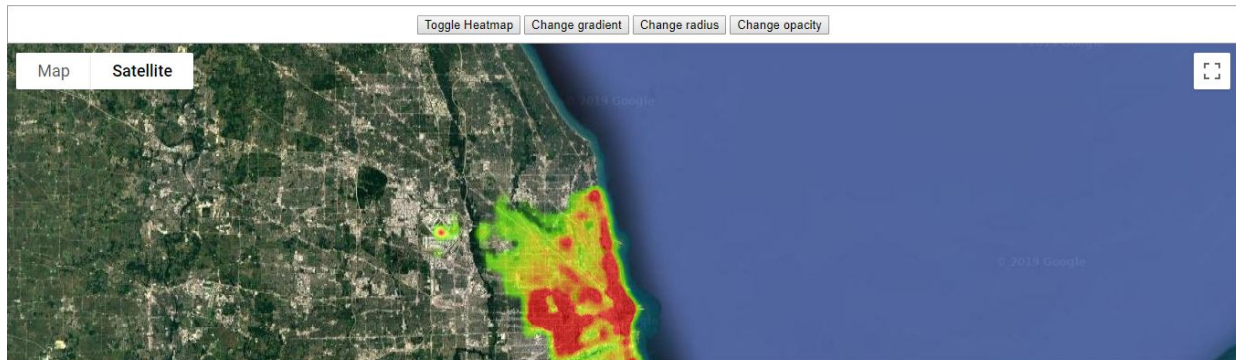


Figure 11: Analysis Screen of Community 71

Welcome to the Community Crime Reduction Portal

Below is a Heat Map showing the spatial distribution of crime in the Chicago area in 2018.

Please Select the Ward for Analysis

Select...

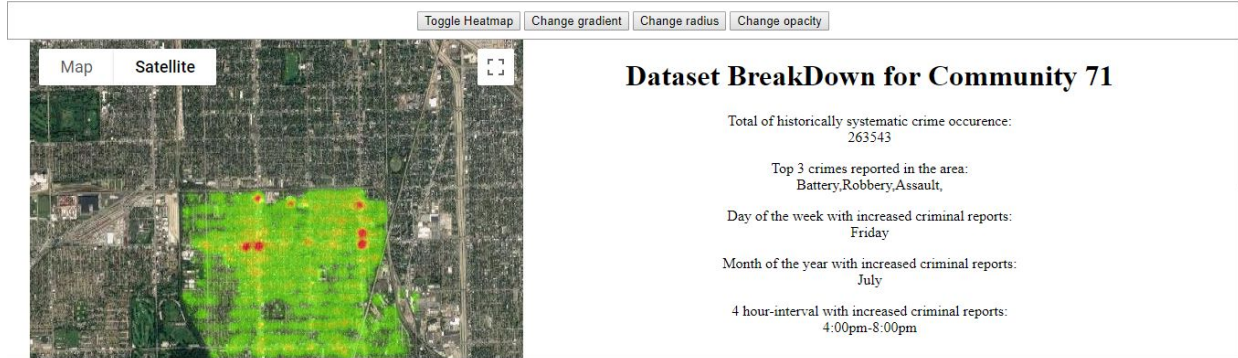
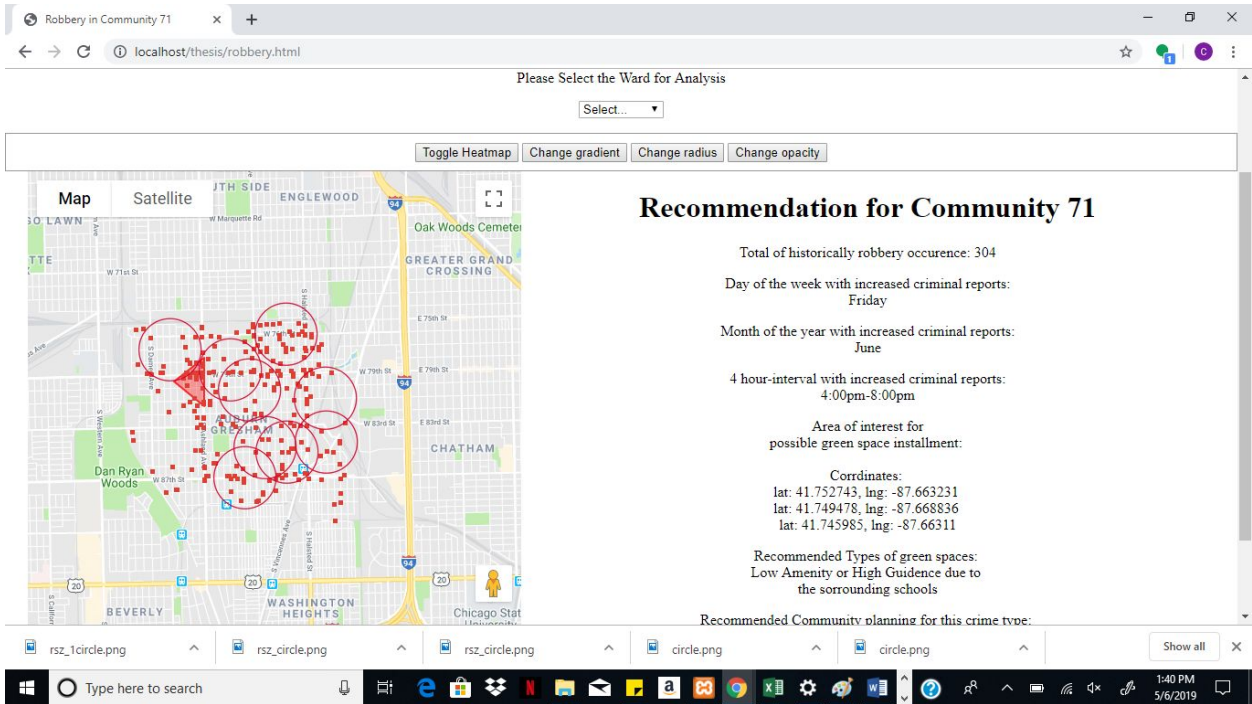


Figure 12: Analysis to find optimal spaces to place greenspace



9 Conclusion

Many interesting graphs and tables were generated and interesting statistical data was found that has given a foundation for what can be next for the roles of predictive policing. When the Apriori algorithm was applied, frequent patterns were established and a better understanding behind the concept of predictive policing was found. Analysis was provided through the manipulation of the key attributes and comparing the outputs with the demographic findings of the areas. The aim of the study was to limitations, biases, and conflict in the implementation of a Machine Learning agents and to find a proposed direction to resolve those issues.

As a future extension of this work, it is planned that a algorithm is applied to help determine radius effects of green spaces, that more models are applied to increase accuracy and to improve performance.

References

1. Moses, Lyria Bennett, and Janet Chan. "Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability." *Policing and Society* 28, no. 7 (2016): 806-22. doi:10.1080/10439463.2016.1253695.
2. Degeling, Martin, and Bettina Berendt. "What Is Wrong about Robocops as Consultants? A Technology-centric Critique of Predictive Policing." *Ai & Society* 33, no. 3 (2017): 347-56. doi:10.1007/s00146-017-0730-7.
3. "UCLA Study on Predictive Policing." PredPol. November 30, 2015. Accessed April 17, 2019. <https://www.predpol.com/ucla-predictive-policing-study/>.
4. Asher, Jeff, and Rob Arthur. "Inside the Algorithm That Tries to Predict Gun Violence in Chicago." *The New York Times*. June 13, 2017. Accessed April 17, 2019. <https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html>.

5. Greengard, Samuel. "Policing the Future." *Communications of the ACM* 55, no. 3 (2012): 19-21. doi:10.1145/2093548.2093555.
6. Kutnowski, Moish. "The Ethical Dangers and Merits of Predictive Policing." *Journal of Community Safety and Well-Being*. March 2017. Accessed April 17, 2019. <https://journalcswb.ca/index.php/cswb/article/view/36/75>.
7. Howgego, Joshua. "A UK Police Force Is Dropping Tricky Cases on Advice of an Algorithm." *New Scientist*. January 8, 2019. Accessed April 17, 2019. <https://www.newscientist.com/article/2189986-a-uk-police-force-is-dropping-tricky-cases-on-advice-of-an-algorithm/>.
8. T. Almanie, R. Mirza, and E. Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 4, pp. 01–19, 2015.
9. A. G. Ferguson, "Police Are Using Algorithms to Tell Them If You're a Threat," *Time*, 03-Oct-2017. [Online]. Available: <http://time.com/4966125/police-departments-algorithms-chicago/>. [Accessed: 17-Apr-2019].
10. "Predictive policing: How algorithms inscribe the ...," *Research Gate*. [Online]. Available: https://www.researchgate.net/publication/324210356_Predictive_policing_How_algorithms_inscribe_the_understanding_of_crime_in_police_work. [Accessed: 17-Apr-2019].
11. G. Saltos and M. Cocea, "An Exploration of Crime Prediction Using Data Mining on Open Data," *International Journal of Information Technology & Decision Making*, vol. 16, no. 05, pp. 1155–1181, Apr. 2017.
12. M. A. Boni and M. S. Gerber, "Area-Specific Crime Prediction Models," *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1–6, 2016.
13. S. V. Nath, "Crime Pattern Detection Using Data Mining," *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, 2006.
14. "Using Big Data Analytics for developing Crime Predictive Model," *Research Gate*, 2016. [Online]. Available: https://www.researchgate.net/publication/302026832_Using_Big_Data_Analytics_for_developing_Crime_Predictive_Model. [Accessed: 17-Apr-2019].
15. Y.-L. Lin, M.-F. Yen, and L.-C. Yu, "Grid-Based Crime Prediction Using Geographical Features," *ISPRS International Journal of Geo-Information*, vol. 7, no. 8, p. 298, 2018.