Project grade: 5.5/6

MGT-415: Data Science in Practice

Reducing crime in Los Angeles by Optimizing Police Patrols

Léopold Bouraux
Gaspard Debains
Jean-Baptiste de La Fage

Gabriel Mehaignerie
Tristan Trébaol
Alfonso Villegas

Directed by :
Prof. Christopher Bruffaerts
Omar Ballester

May 11, 2020

Contents

1 Introduction

1.1 Background

Los Angeles is the second largest city in the United States in terms of population after New York. With a population of approximately 4 million people and an urban area of approximately 20 million people in 2020, it is among the fastest growing cities in the United States in the last century.

In the 1930s, when the movie industry was turning into a veritable business, with the well-known epicentre Hollywood, Los Angeles began to gain international fame. This new excitement about LA made it one of the most multicultural cities in the United States, attracting people from all walks of life and all social classes (Fig. 1). At the end of the 20^{th} century, the city, often nicknamed *gangland*, had a bad reputation because of its many conflicts between rival gangs. Los Angeles has seen a significant decline in acts of violence since the mid-1990s. At that time, a study conducted by the National Drug Intelligence Center reported 1,350 gangs comprising 152,000 individuals [5]. Despite the negative stereotypes, crime has not stopped decreasing since those years. Recent statistics show that the Californian city has a lower crime index than most of major American cities.

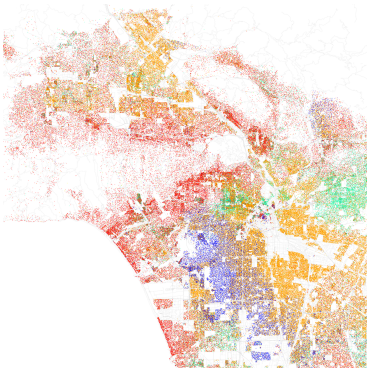


Figure 1: Distribution of ethnic groups in 2010: White, Black, Asian and Latinos

The Los Angeles Police Department (LAPD) is the police force in the city. The department was founded in 1869. Nowadays, it is the third-largest municipal police department in the US, after the New York and the Chicago Police Department: the agency has 10,002 police officers as well as 2,961 civilian staff in its ranks. All these agents must operate in an area of 1,290 km^2 depicted in red in the Figure 2. The 21 police stations in LA, called "divisions", are grouped geographically into four areas, each known as a "bureau": the Central, the South, the West and the Valley Bureau.

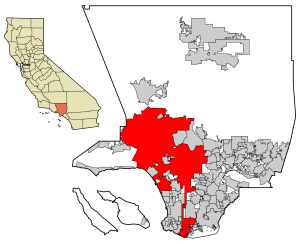


Figure 2: LAPD's jurisdictional area

Obviously, crime takes a heavy toll in our society. Not only does it brings great personal distress as well as emotional turmoil within a community, it also creates a financial cost that needs to be payed.

In the US, the annual direct cost of law enforcement is about $350 billions. This only covers the cost of police services, correctional facilities and judicial and court costs [2]. The total price of crime in the US is much greater. The direct economic costs to the victims are estimated to be at $ 310 billion every year. This estimation includes only the crimes that are reported, meaning that this cost could be very much higher. All in all, crime makes up for more than 5% of the US's GDP [2].

An efficient and effective prediction of the crime is hence crucial for every communities in order to bring down those emotional and financial tolls, especially in LA. Its size and

diversity as well as its importance to the economy of the US and its international influence make it a priority in the effective implementation of its large police force. Its financial fallout is even greater.

1.2 Problem statement in business

One of the Los Angeles Police Department's main goal in the recent decade has been to try predicting where the next crime will happen, and who will commit it. This is done by a better management of the police patrols. It has been proven that intelligently patrolling areas where crimes are most likely to be committed reduces the occurrence of those crimes. An efficient management of the patrolling can greatly increase the productivity of the police force in this task. We see below the average time required to clear a call for selected crimes (data has been collected by the San Francisco Police Department) [2]:

- Burglary – 1.9 hours
- Auto theft – 1.8 hours
- Assault – 2.6 hours
- Robbery – 2.3 hours

We see that reducing even one instance of Burglary or Auto theft can free up two hours or more for an officer's time of patrol [2]. Efficient patrolling could be as effective as increasing the patrol staff, but only for a fraction of the cost. As we can see, intelligent patrolling seems obvious. Where things start to get much more difficult is how to implement such a tool.

Our answer to that problem, and one that has proven to be effective, is predictive policing. Predictive policing refers to the use of analytical techniques by law enforcement to make statistical predictions about potential criminal activity.

The LAPD has tried to implement predictive policing for patrol optimization. They started a campaign in 2011 called Operation LASER [3], aiming at a crime prediction algorithm based on past crime data and previously convicted offenders. This operation, using new technology brought by the data analysis firm *Palantir*, uses information about past offenders and scores individuals based on their previous records. The higher the points, the more likely you are to end up on something called the *Chronic Offender Bulletin*, a list of people who are the most at risk of re-offending according to the data and need to be closely observed [1]. The problem with this approach is that it is based on data about individuals. Civil rights lawyers thus argue that such a tool is just a renewed and re-branded version of racial profiling. Additionally, this operation only uses the criminal records of previously convicted felons and can only target a very small percentage of the population of Los Angeles. To prevent a future crime with this procedure, it would be necessary to keep track of all former offenders and to observe their every move. Besides being a serious infringement of personal privacy, it is expensive in terms of money, manpower and could be terribly ineffective.

This is where our product comes in. Our goal is to propose a data-based optimized crime prediction model for a given time frame while keeping civil liberties untouched.

As opposed to the LAPD's campaign, we will make predictive policing based on crime data rather than data about the ones that have committed it. This offers a much larger visibility on the crimes and allows for a much more intricate understanding in order to make more accurate predictions.

1.3 From police patrols to data analysis

1.3.1 Police Patrolling dynamics

Police patrolling in Los Angeles is done 24 hours a day 7 days a week. About 70% of the police force is dedicated to that task at any given time. Patrolling is organized in shifts: Morning, afternoon and night. In a simple way, we aim at predicting crime based on crime data of the past years in the city of Los Angeles in order to better allocate police resources, especially in patrols. But this is a very broad statement that infers lots of different "predicting" possibilities. Do we set at predicting crime trends in the next year or in the next day ? Do we set at predicting crime in a large neighborhood like Hollywood or for a specific street block ?

We hence have to focus our prediction model in order to be the most helpful possible for the Police forces. The LAPD, like many other Police Departments, do not only work day to day, randomly patrolling around to prevent any kind of crime. They set long terms targets to emphasize on a specific crime and tackle to reduce it while also set short term targets in certain neighborhoods and communities. This means working towards goals both in the long and short run as well as in different geographical scales.

From this understanding, we will work on models of different time scopes in order to increase the reactivity and adaptability of the police forces. Geographically, we fix our model to predict "Hot-spots" of a set size. This means that we divide the city of Los Angeles in a grid with 1 km wide squared cells. This allows for a coverage of a good number of cells during a 4 hour shift for a police patrol.

1.3.2 Weekly Prediction model

Our first model will focus on the weekly time scale. We set at predicting the crime schedule throughout the week for a given week. Each week, a heat map of Los Angeles is created showing the different crime hot spots for each time shifts of each day. This helps in the organisation of the patrolling for the week by allocating the right amount of police patrols to the required blocks per shift.

1.3.3 Monthly target model

The second predictive model focuses on the monthly time scale. Each month, the heat map for a specific crime is predicted. This allows to set goals for tackling a specific crime and improve efficiency in preventing it on the long run.

1.3.4 Shift to shift reactive model

The third model works on the shift time scale. It focuses on the prediction of the crime locations based on crimes reported in the previous shift. This model is the one that is

the most reactive. It allows for corrections and improvement on a shift to shift basis from the weekly schedule outlined in the first model considering the events that just occurred.

1.3.5 Analysis Process

In order to make a police predictive tool, we will use incidents of crime in the City of Los Angeles transcribed from original handwritten crime reports dating back to the 2010s [4].

We will first implement an Exploratory Data Analysis in order to get a full grasp of the crime dynamics in the city. Then, we will implement a Feature Engineering process in order to transform our data set to make a more simple but effective analysis as well as to identify the variables that have the most impact. We will then implement the different Machine Learning methods presented above and discuss their results. 

2 Data Analysis

2.1 Data presentation

Since we want to train models in order to predict the optimal distribution of police patrol within LAPD jurisdiction area, we need both a lot of crimes data, as well as geographical data such as coordinates where crimes happened or maps to visualize our data. We found a complete set of datas on Kaggle [4].

Among these datasets we had geographical data to visualize our results on maps. In particular we used the shape file `LAPD_Reporting_Districts.shp` to draw an interactive map and apply our crime data geographically (see Fig.3). Crime data can be found in `Crime_Data_2010_2017.csv` and includes all data collected by the LAPD during arrests or complaints from 1 January 2010 to 9 September 2017 included. It consists of 1,584,316 cases from a simple altercation to car wrecking, rape or murder. Each case has 26 features which are:

- DR Number
- Date Reported
- Date Occurred
- Time Occurred
- Area ID
- Area Name
- Reporting District
- Crime Code
- Crime Code Descr.
- MO Codes
- Victim Age
- Victim Sex
- Victim Descent
- Premise Code
- Premise Descr.
- Weapon Used Code
- Weapon Descr.
- Status Code
- Status Descr.
- Crime Code 1
- Crime Code 2
- Crime Code 3
- Crime Code 4
- Address
- Cross Street
- Location

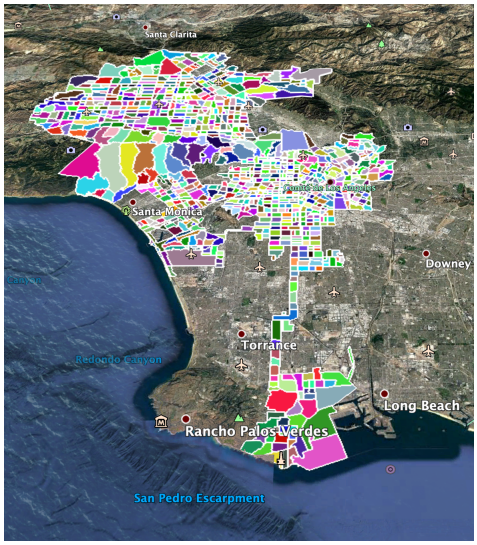


Figure 3: Shapefile districts

For the more complex ones, here are their interpretations. **DR Number** is the unique case number, **Area ID** is similar to the Reporting District, **MO Codes** and **Crime Code 1/2/3/4** are an alternative Crime Code, **Victim Descent** is the victim ethnicity, **Premise** is a short description of where the crime happened (parking, in the street...), **Status** is the type of arresting, **Cross Street** is a boolean which indicates whether or not the crime occurred at a crossroads, and finally **Location** is the geographical coordinates of the crime, which will be very useful in our case.

We will use this data to predict where the next criminal acts will occur. But first, as it can be noticed, many features are not relevant to the analysis, therefore a preprocessing is needed to clean the data.

2.2 Data preprocessing

At first sight from this dataset, and regarding the problem we want to solve, several columns could be deleted. On the one hand, invaluable features such as DR Number, Date Reported or Status Description/Code could be removed. On the other hand, redundant data is deleted: we only keep Crime Code Description to describe crimes, the geographical coordinates and the district number at the geographical information level and both Premise Code and Weapon Used Code could be removed. Also, in order to avoid any problem of discrimination, which is often reported to the US police, we decided to get rid of the Victim Descent.

Among the remaining 9 features that we still need to clean up, only 4 have missing values to handle: Weapon Description, Victim Sex, Victim Age and Location have each respectively 66.9%, 9.2%, 8.1%, 0.4% of missing values.

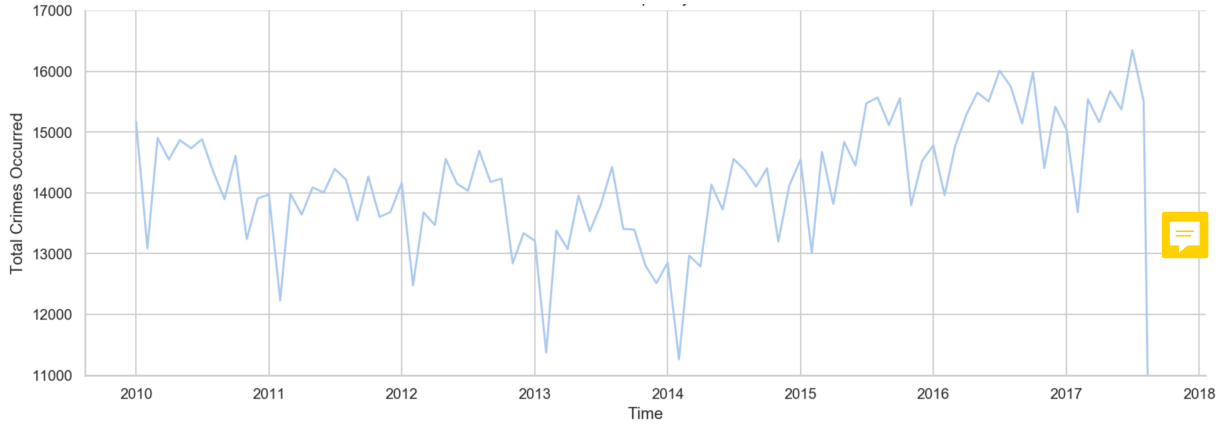
We decided to delete only the `NaN` sex data because these crimes were bizarre, the victims were teenagers, and the typical crime didn't match them. Concerning Date and Time Occurred, we just extract year, month, day and week day from the date, to make more advanced statistics. We also group crimes by 4 hours shifts which is a reasonable amount of time for police patrol shifts. Then, we reduced Crime Code Descriptions by gathering similar crimes and by deleting crimes that are not interesting to us, such as crimes where the police cannot operate directly on. We also reduced Weapon Description categories and replaced missing values since they accounted for 'no weapon used'. Missing locations have been replaced by the centroid coordinates of the district in which the crime was committed. Finally, we draw a grid over the LAPD jurisdiction zone with 1 km wide square cells that are a kilometer wide, and add a 'Cell Nb' feature for each crime.

This cleaning removed 17.44% of the initial dataset, but still keep more than 1300,000 crime data, which is a significant number of event to be able to train our prediction models. Before we get into that, let's analyze this data through univariate and multivariate feature analysis.

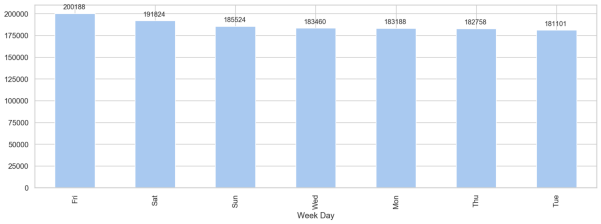
2.3 Data Analysis

2.3.1 Date and time statistics

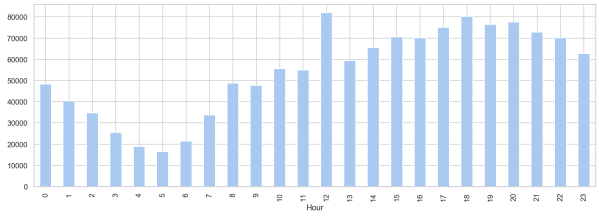
We have a lot of data on when the crimes occurred. It is interesting to see the evolution of the number of crimes over the years, but also to see the distribution of crimes over a year, a



(a) Evolution of crimes number through months



(b) Number of crime per week day



(c) Number of crimes each hour

Figure 4: Statistics through time

week or even a day. Figure 10, shows us these types of distributions. In particular, Figure 4a, depicts the evolution over months of total number of crimes registered by LAPD. We can see that it decreased from January 2010 to February 2014, reaching its minimum. Then the trend went upwards. That's why our analysis is important here. We want to bring the curve back down with crime location prediction. Month-to-month variations are due to the number of days per month, which explains why February always has fewer crimes than others. However, when the number of days per month is normalized, the distribution over the months in a year remains stable. The number of crimes during the week seems fairly constant (Fig.4b), but still higher on Fridays. This could be explained by the fact that people go out more often on Friday nights, and there is on average more crime in the evenings (Fig.4c). What jumps out is the number of crimes committed at 12:00. This might be due to crimes that cannot be precisely dated and which are registered by default at noon. For instance, it could be the case for online crimes, or some robberies and burglaries.

2.3.2 Victims Age

The average age of a victim is 37.4 years and the median age is 35 years old. Figure 5 shows that young people and older people in general are not very affected. After the age of 20, the number of victims increases quickly. It can also be seen that the most affected target is 25-year-old women with 19,000 crimes against them. We can also see that the maximum age is 99, and that there are more crimes committed against them than against 98-year-olds and youger. This is because the LAPD includes all the centennial victims at that age. This graph, depicts as well that the majority of unidentified sex people are minors.

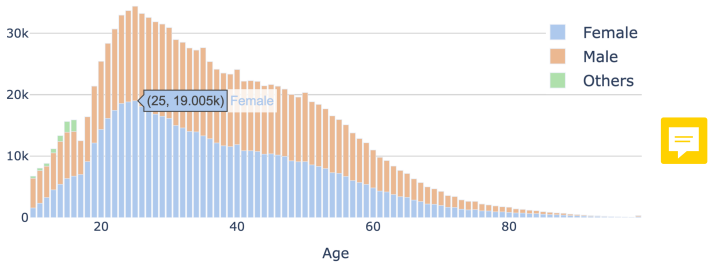


Figure 5: Repartition of sex victims depending on the age

2.3.3 Different types of crimes

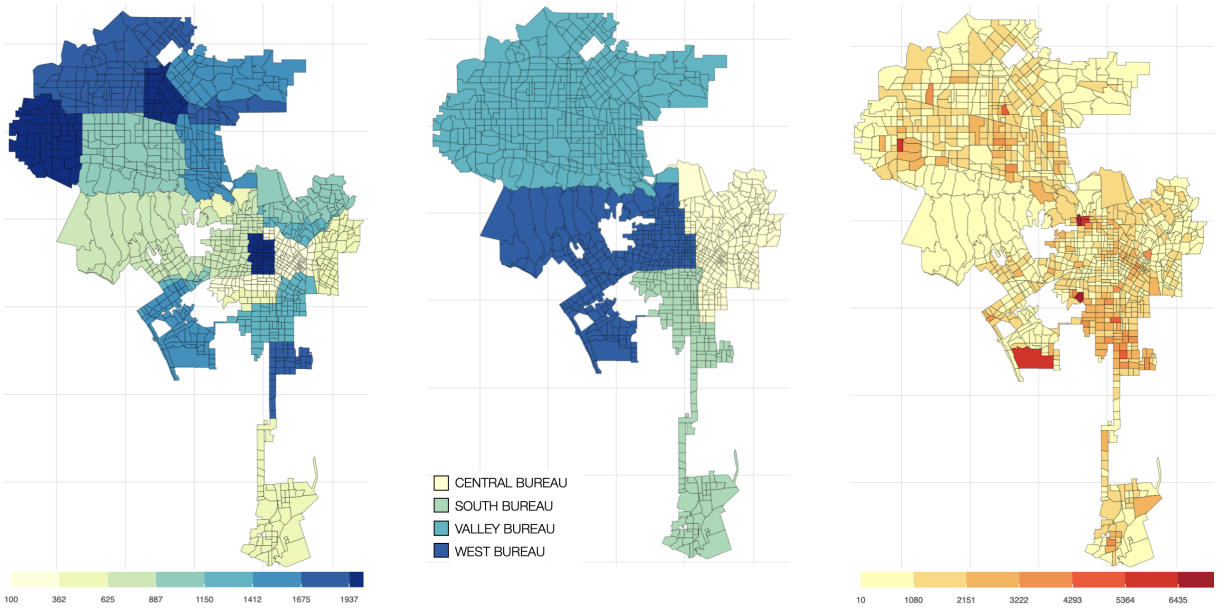
After reducing categories by grouping several of them together, we can see a wide spread between the most and least common crimes. The most common ones are burglaries, thefts and simple quarrels (something we often see), while the least common one, which happened only once, is the destruction of a train. Burglary is the most common crime committed against women and men. However, the type of crime depends deeply on the victim age group as well as the victim sex. Women are more victims of Simple Assault than men, and especially often victims of Intimate Partner Assault. On the other hand, men are more victims of theft. For the category concerning the other sex, the majority of crimes committed against them are shoplifting, this is due to the high representation of young teens in this category. Some of these crimes are recurrent. This is the case of simple assault, burglary, or theft plain which are in the top 5 most committed crimes for each age group. But quite surprisingly, Assaults with a deadly weapon are very common for youngest and not for others categories, while intimate partner assault is more common in the 20-39 age group than in others.

2.3.4 Weapon Analysis

As seen previously, about $\frac{2}{3}$ of offenders do not use a weapon in their crimes. That leaves approximately 430,000 crimes committed with weapons. Among those crimes, 318,408 are perpetrated with natural strong arms such as hands, feet, fits, etc. The second most represented category is firearms of all types (from hand guns to automatic weapons) with 54,385 representation. In total, we have data for 2809 days, i.e. 7 years, 8 months and 9 days. So on average in Los Angeles, there are 19.36 crimes involving guns daily. Figure 6, shows that the number of gun crimes tends to increase in recent years, and generally follows the upward trend of the total number of crimes (Fig.4a).

2.4 Data Visualization

For a better appreciation of the geographic location of crimes within the LAPD's jurisdiction, a map view is preferable. First of all, as mentioned earlier, each district is assigned a district number (Fig.7a), as well as one of the 4 Bureau (Fig.7b).



(a) Districts ranked by their district number

(b) District affiliation to a particular Bureau

(c) Districts ranked by total crime number

Figure 7: Simple maps by district

In Figure 7c, the total number of crimes committed by district is represented through the heatmap. It can be seen that crime is not uniformly distributed across the districts. Some are more affected than others. Therefore, patrols optimization makes complete sense here.

2.4.1 Most affected districts

In this section, we focus on the 5 most affected districts. In Figure 7c, they are the red one in the north-west of the city (District 2156), the big one in the center (District 1494), the one northeast of the latter (District 363), and the two red ones above it (District 645, 646).

In District 2156, there is a large shopping mall and a heliport that may account for the high crime rate. The Los Angeles International Airport in District 1494 is clearly behind the high incidence of crime in this district. Within District 363, again a mall might be responsible for the high delinquency whereas in Districts 645 and 646, tourism is the

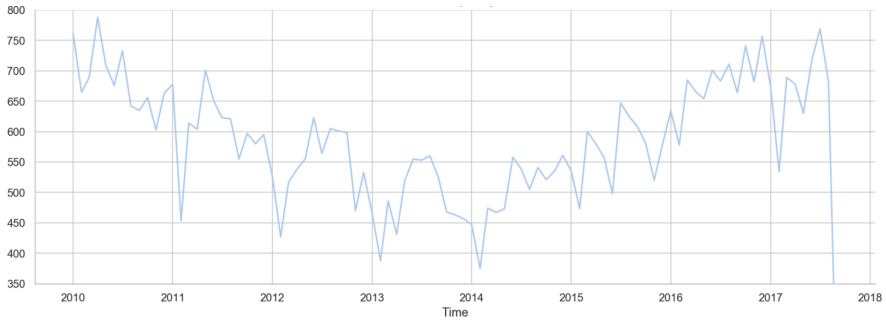


Figure 6: Evolution of number of crimes involving firearms through months

main reason. Indeed, it is the area of the famous Hollywood Boulevard where thousands of tourists (frequent target of delinquents) travel every day.

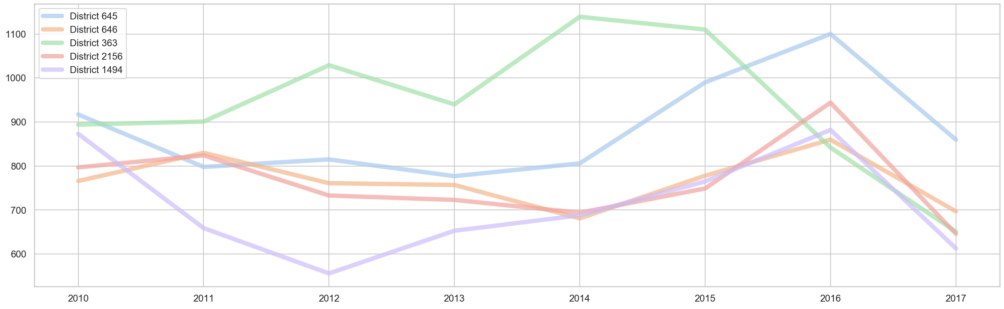
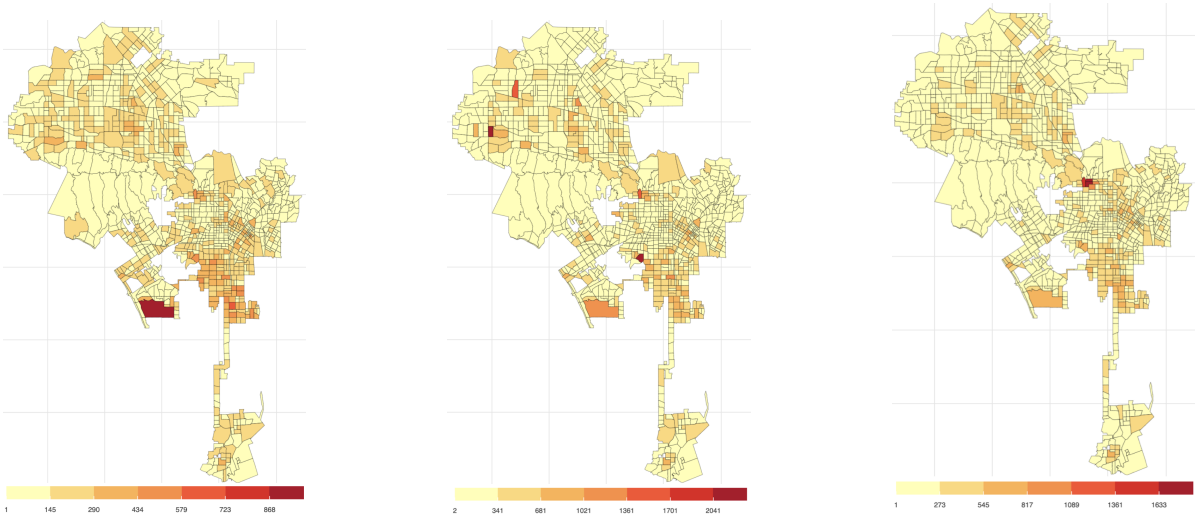


Figure 8: Evolution of number of crimes in the 5 most affected districts

In figure 8, we can see the evolution of the number of crimes per year over the 8 years available in our dataset. The upward trend since a few years ago in these neighbourhoods (except District 363) is striking. Especially in the touristic district 645, where the number of crimes committed on September 9 is already higher than all crimes committed in 2014 in the same neighbourhood. It is in this type of district, where crime is rising, that we need to maximize police patrols mobilized to fight this effect.

2.4.2 Most affected districts given time in the day



(a) From 6h to 10h (b) From 14h to 18h (c) From 22h to 2h

Figure 9: Most affected districts for a given time slot

Based on the most important locations indicated in the previous subsection for the 5 most affected districts, the crime distribution according to time of day, depicted in Figure 9 becomes more meaningful.

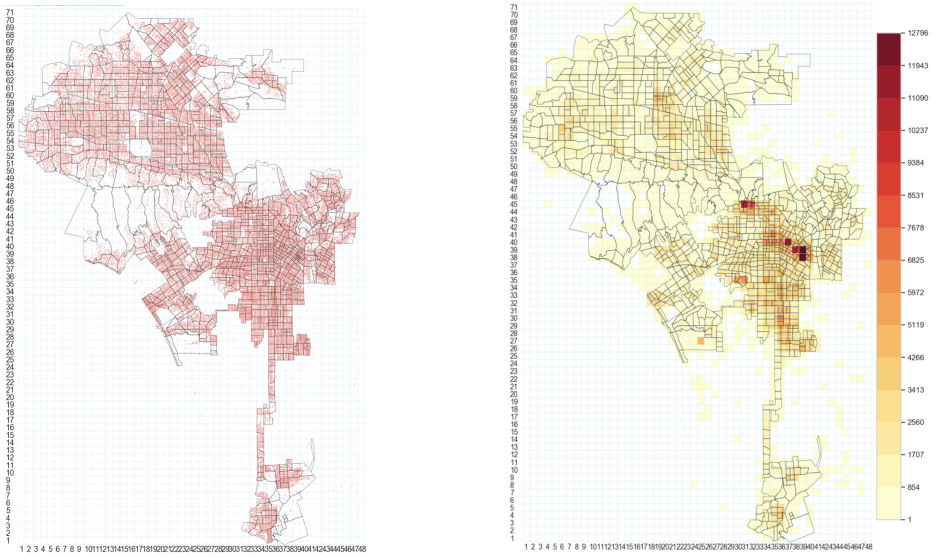
Thefts in the airport are more frequent in the morning when people are a bit dizzy and tired, or for instance when travellers get out of the plane and wait for a suitcase. Shopping mall criminal acts take place in the afternoons when stores are open and crowded. Crimes

in tourist areas take place a lot at night, after dark, when people are distracted or drunk during wild parties.

However there is something wrong with these visualizations: each district does not have the same surface area. There are large districts with very little delinquency and, on the contrary, there are much smaller districts where there is significant criminality.

2.4.3 Map without states but with grid mapping instead

Which is preferable for machine learning reasons is to represents crime occurrences in equal geographical areas. We then draw a 1 km^2 grid over the map and draw the same heatmap than in Figure 7c. In this new map (Fig.10b), new neighbourhoods that did not necessarily appear to be important to deal with when visualizing whole districts, appear to be highly affected by crimes. This is the case of the historic centre, the business district and the surrounding areas (east of the map). Indeed, these neighbourhoods are divided into many very small districts, which previously biased the analysis. This grid normalizes the number of crimes according to the size of the area. It is on these equal sized regions that we will apply our prediction algorithms.



(a) Crimes precise location with 1 km^2 grid

(b) 1km^2 cells ranked by total crime number

Figure 10: Grid maps

2.5 Feature Extraction

Table 1: Data set after preprocessing

Feature	Description	Type
Reporting District	District in which crime has been reported	Numerical
Victim Age	Age of the Victim	Numerical
Victim Sex	Sex of the Victim	Categorical (Multiclass)
Premise Description	Type of location where the crime happened	Categorical (Multiclass)
Latitude	Latitude of the crime coordinate	Numerical
Longitude	Longitude of the crime coordinate	Numerical
Year	Year it took place	Numerical
Month	Month it took place	Numerical
Day	Day it took place	Numerical
Week Day	Week day it took place (Monday for ex)	Categorical
Hour	Hour at which it took place	Numerical
Shift Hour	Day divided into 3 shift hours	Categorical (Multiclass)
Victim Tranche Age	Tranche of age of the victim	Categorical (Multiclass)
Crime Code description	Description of the type of crime	Categorical (Multiclass)
Cell Nb	Cell in which the crime occured	Numerical

2.5.1 Removing data

After the data pre-processing, we have the data set as it can be seen in Table 1. Here, we see that data associated to the victims (Victim Age, Victim Descent, etc..) has been kept. In order to create a model which ensures that no personal information is used and that no biases can be created, we need to remove certain categories straight away. We remove **Victim Age**, **Victim Sex**, **Victim Descent** and **Victim Tranche Age**.

The second step is to change all the categorical variables in order to make them meaningful and exploitable. This is done using *One-Hot Encoding*. This way, the data is translated from categorical to binary.

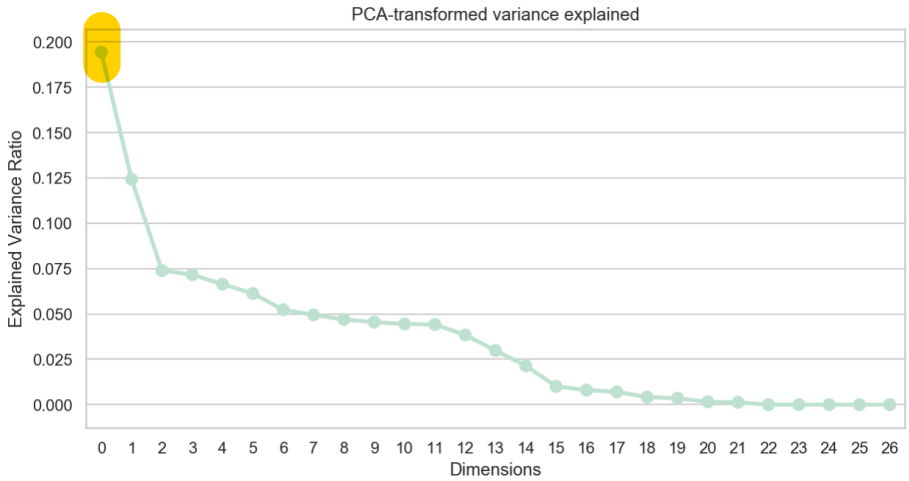
By doing this *One-hot encoding* here, we create a lot of new variables in our data set. There is 79 different entries for **Weapon Description**. This is normally something that we want to avoid. We can solve this issue by grouping into larger categories the classes of one variable. Here are the new broader categories for these variables :

- Weapon Description : No weapon - Weapon
- Premise Description : Inside - Outside
- Crime Code description : Assault - Theft - Robbery - Burglary - Motor Vehicle - Arson - Homicide - Vice - Narcotics - Others

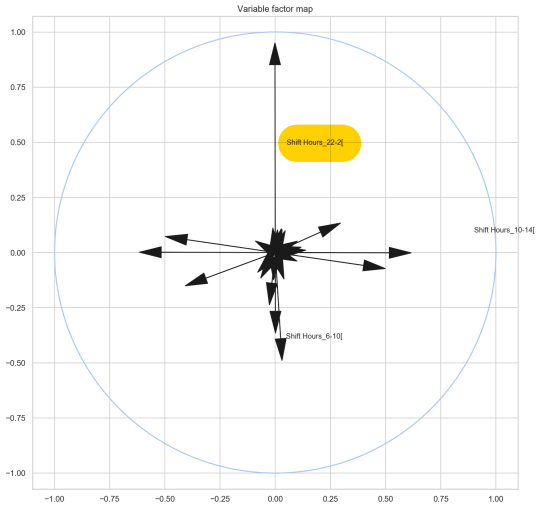
We then put all the categorical data into *One-hot encoding*.

2.5.2 Principal Component Analysis

We can then run a PCA in order to determine the relative importance of the variables in the data set and get some insights as to which variable we could drop without loosing significance.



(a) Cumulative percentage of explained variance



(b) Representation of Eigenvectors

Figure 11: PCA graphs

From Figure 11a, we see that 75% of the variance is explained by the first 3 components. We see in Figure 11b that these 3 components are different time shifts in the day. We understand that these are the variables that have the most impact on the crime distribution in the city of Los Angeles.
From this PCA, we see that the first three components could be kept and the rest removed. We are here with a trade-off between the best possible prediction accuracy for our model or the simplification of the analysis and the reduction of the computational time.
Looking at the PCA results, we choose to implement a model with less features so as to have a better handle and comprehension of the model dynamics. We remove **Weapons Description** as well as **Premise Description** because of their relatively low variance

explanation. Considering that the implementation of our model will be difficult, it is safer to remove these variables. Obviously, keeping them would improve the strength of our predictions. We keep **Crime Code Description** because it is an essential aspect of the crime prediction. If we wish to set specific targets on reducing a given crime, this feature is necessary in our model. This PCA will help us in the implementation of our model. We know that the shift hours play an important role and will be essential in the accuracy of the prediction model that we establish.

2.5.3 Other modifications to the data set

Implementing a prediction algorithm for a specific location would be extremely difficult and useless. Indeed, the granularity for a specific location is very small and any Machine Learning algorithm would fail to make a relevant prediction. To solve this problem, we increase the granularity of the crime location. We achieve this by dividing the map of Los Angeles into a grid with a granularity of 1 kilometer. Each crime location is associated to the grid cell that covers it.

We are then left with around 3000 cells of 1 km by 1km that cover the surface of the city of Los Angeles. We will now focus on establishing a prediction for a given cell, rather than a specific location. This would increase the generalization possibility and make for a more accurate model later.

2.5.4 Summary of Data Analysis

Finally, we have the following data set that will be used for the training of our model (see Table 2). It is reduced compared to the numerous features that were initially present. We have seen that the crime dynamics is heavily influenced by time especially week days shift hours and months. They are the main trend setters in the crime distribution. Our model will focus on this for the crime location prediction.

Table 2: Data set used for the Machine Learning

Feature	**Description**	**Type**
Latitude	Latitude of the crime location	Numerical
Longitude	Longitude of the crime location	Numerical
Year	Year it took place	Numerical
Month	Month it took place	Numerical
Day	Day it took place	Numerical
Week Day	Week day it took place (Monday for ex)	Multiclass
Shift Hour	Part of the day the crime occurred	Categorical (Multiclass)
Crime Code Description	Coordinates of the location of the crime	Categorical (Multiclass)
Cell Number	Cell in which the crime was located	Numerical

3 Prediction models

3.1 Prediction of a typical weekly schedule

In this model, we assume that the crime prediction is deterministic, i.e. no crime will be predicted in the places where no crime has ever happened.

3.1.1 Objectives

In order to optimize police patrols localisation and itinerary, the best case scenario is be to have in advance the crime repartition and density within the LAPD jurisdiction, from one day to another.
Thus, taking advantage of the size and variety of the datas at our disposal, we tried to extract from the data the most typical possible distribution of a week of crime in LA.
More precisely, the goal was to get different heat maps illustrating the occurrence probability of any type of crime to happen for each days of the week, divided into 6 time slots.
The choice to reduce the analysis and the projection to specific time slots and days of weeks is justified by the preliminary data analysis, where it seemed that those elements where two important factors both strongly influencing the crime distribution in Los Angeles.
Indeed, the PCA performed in the section above further justify our choice of focusing only on the time slots and the days of the weeks, as there are parameters that explained a consequent part of the data set variance.

3.1.2 Methodology description

After the preliminary data analysis and processing, we further reduced the complexity of the data set, by deleting any information concerning the victim or the type of crime, as our objective was just to know if a crime of any type was suspect to happen for a given day and time slot, at a given location on the map.
Then, we replaced the year, month, and day information by the week number, starting from week 1 to week 401, and we deleted the first 18 lines of the data set to begin with a Monday.
Also, we resumed any information regarding the localisation, such as the district name or the latitude/longitude, and used only the cell number , which was then corresponding to the resolution of our map.

Table 3: Data set used for the first model

Feature	Description	Type
Week Day	Days from Monday to Sunday	Categorical, used in one hot encoding
Time Shift	Slot of 4 hours patrol shift (6 shifts)	Categorical, used in one hot encoding
Cell Nb	Cell number in which the crime occured	Categorical, used in one hot encoding

The objective of the regression is to fit the shifts over one full week with the spatial distribution of the crimes. It was performed with a RandomForestRegressor with the following data:

- Spatial data: matrix y contains a one hot encoded version of [Cell Nb], the reference number for each cell of the grid mapped over LA. It total there are 1428 grid slots where at least one crime was recorded. Hence, y contains 1428 columns. It has 16815 rows, because it is the number of shift records over the 401 weeks.
- Time data: the matrix X contains a one hot encoded version of [Week Day] and [Shift Hours]. X has therefore 13 columns for the 7 week days and the 6 patrol shifts per day; and 42 unique rows repeated 401 times to match the number of rows of y.

The input contains a one hot encoded version of [Week Day] and [Shift Hours] information is a one hot encoding of X corresponding to a 2 column matrix [Week Day, Shift Hours]. The output Y to be returned was then a 1428 column matrix, with a value ranging from 0 to 1.6 in each cell. This value characterize the estimate frequency density for a any type of crime in a given cell, for a certain time slot during the week.

3.1.3 Results

Figure 12 illustrates the data after the regression. The columns named 34 to 3375 represents the cell number for which a crime happened during the corresponding shift.

Cell Nb	Mon 10-14[	Mon 14-18[	Mon 18-22[	Mon 2-6[	Mon 22-2[	Mon 6-10[	Tue 10-14[	Tue 14-18[	Tue 18-22[	Tue 2-6[	...	Sat 18-22[	Sat 2-6[	Sat 22-2[	Sat 6-10[	Sun 1 1
34	0.000000	0.000611	0.000000	0.001197	0.002368	0.000000	0.002738	0.002182	0.003282	0.001195	...	0.002738	0.003526	0.008402	0.001198	0.0027
35	0.024893	0.016871	0.007899	0.001526	0.011233	0.011907	0.010324	0.019617	0.028160	0.006026	...	0.018034	0.019533	0.022733	0.002222	0.0100
36	0.009963	0.012000	0.029602	0.002286	0.008718	0.013263	0.019970	0.038665	0.018030	0.014688	...	0.039238	0.008229	0.024931	0.002965	0.0246
80	0.007628	0.000000	0.000853	0.000000	0.004090	0.001821	0.001300	0.001348	0.000000	0.000000	...	0.003402	0.006329	0.002721	0.000000	0.0010
81	0.004470	0.003391	0.006965	0.002718	0.004170	0.004453	0.004756	0.003719	0.011113	0.006760	...	0.005050	0.000000	0.006144	0.000000	0.0077
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3335	0.010817	0.016416	0.009493	0.005414	0.015225	0.013173	0.006162	0.005030	0.012855	0.001354	...	0.010588	0.002255	0.010527	0.010539	0.0025
3336	0.003486	0.001223	0.001474	0.001791	0.001755	0.005699	0.004388	0.006459	0.012202	0.003225	...	0.003663	0.008190	0.002139	0.001338	0.0028
3337	0.000000	0.000000	0.001628	0.000897	0.002505	0.000000	0.002123	0.001829	0.000000	0.000000	...	0.000000	0.000000	0.003746	0.000000	0.0000
3375	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.001227	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0000
3376	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0000

1428 rows × 42 columns

Figure 12: Result of the regression

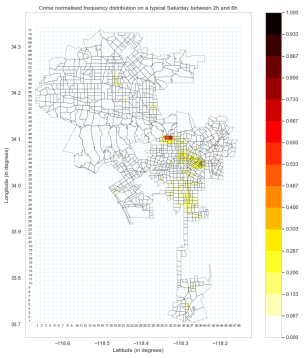
As stated before, the frequency distribution of the crimes ranges between 0 and 1.636. The maximum value is obtained on a the evening shift (18h-22h) typical Sunday, see figure 12. On this figure one can observe that the crime activity on the grid is very low.

Maps of crime normalised frequency distribution over a typical week in Los Angeles

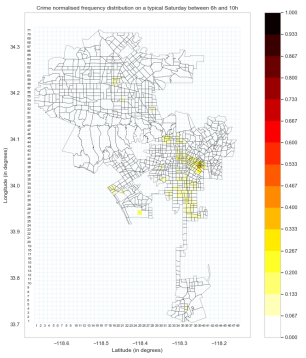
Since each row of the table in figure 12 corresponds to the crime frequency distribution on the grid during a specific shift, can now produce the heatmaps with the crime distribution in Los Angeles.

In order to produce those maps, we decided to linearly normalise all the data. Indeed, we thought that values between 0 and 1 is more understandable than between 0 and 1.636, as we can assimilate them to an occurence probability. Also, the loss of information is not too large.

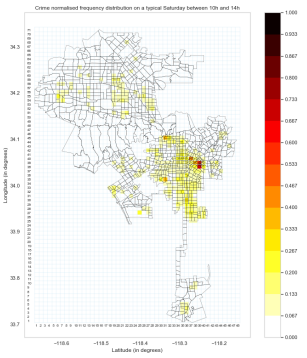
For readability purpose we display the results for only two days. We chose the days which where the most different: Wednesday (day with low crime frequency) in Figure 13, and Saturday (day with high crime frequency) in Figure 14.



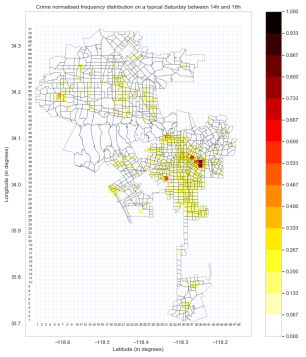
Time slot 2h/6h



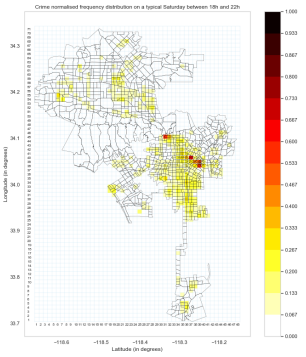
Time slot 6h/10h



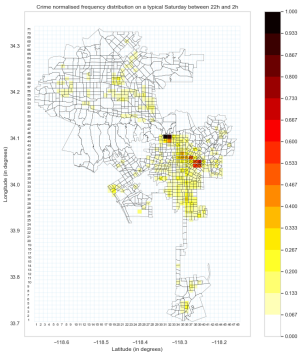
Time slot 10h/14h



Time slot 14h/18h

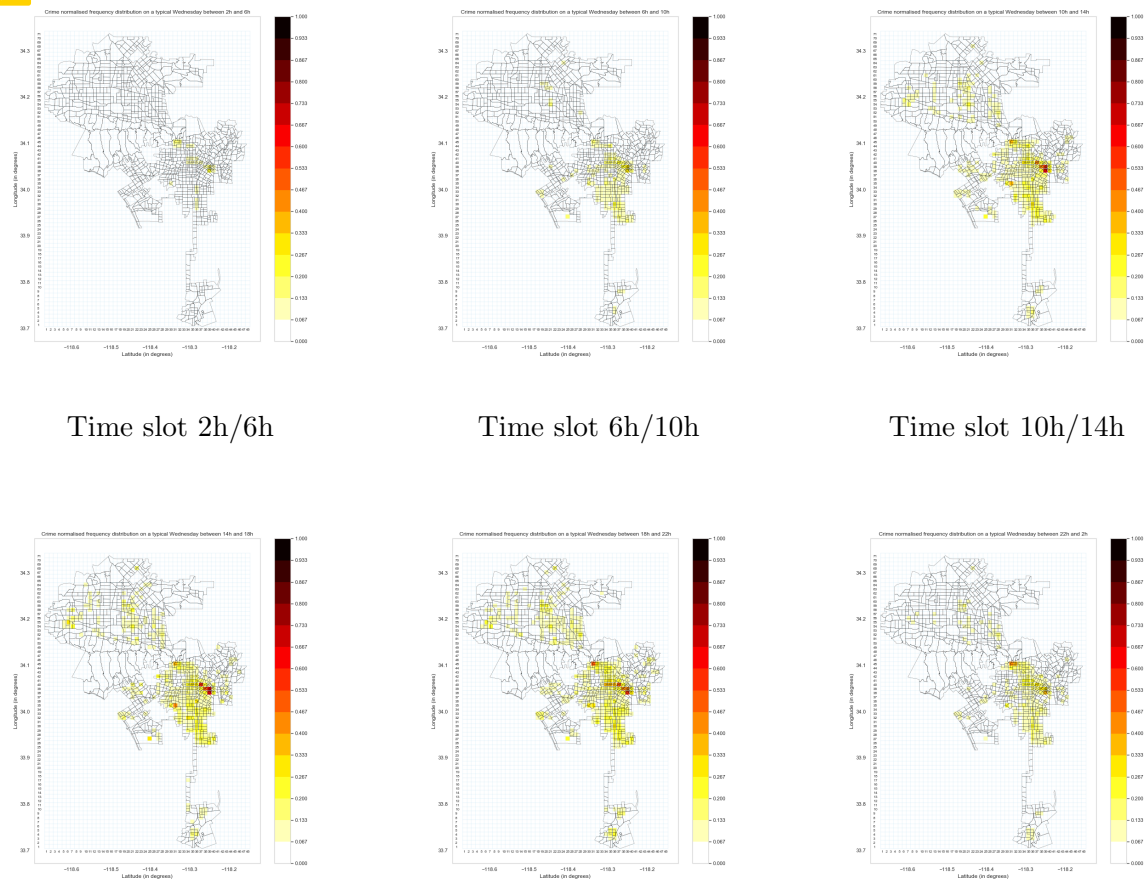


Time slot 18h/22h



Time slot 22h/2h

Figure 13: Predicted crime repartition for the different time slots of Saturday



Time slot 2h/6h | Time slot 6h/10h | Time slot 10h/14h

Time slot 14h/18h | Time slot 18h/22h | Time slot 22h/2h

Figure 14: Predicted crime repartition for the different time slots of Wednesday

These two days represent well the change in crime repartition between different time slots in the first hand, and between week and week-end in the other hand.
Indeed, as expected, for a same time slot, the changes in crime repartition between week days is non-significant. However, the modification is clear when it comes to the comparison of week days and week-end days.
The time slot 22h/2h and 2h/6h heat maps present a strong density of crime in cells corresponding to the walk of fame area and the downtown on Saturdays, whereas on Wednesdays it is really less significant.
The other interesting feature of these heatmaps is the apparition or disapparition of different key hot spots when switching to different time slots of the same day, an element that could be used to better organise and plan different routine police patrols, and direct them to where they are most needed.

3.1.4 Advantages and Limitations

The main advantage of this model is it simplicity. Based on crime data records over almost 8 years (401 weeks), it gives the most probable distribution of crimes for each day

and time shift of a week. The police patrol could use it as a solid base to distribute its unity, as it gives the principal criminal hot spots that have arisen for the last 8 years.

However, a non negligible limitation concerns the steady state aspect of this analysis. As it is solely based on the 8 previous years, it does not take into account nor give more importance to the crimes that has happened recently. It only gives a fixed representation of the weekly criminal distribution of Los Angeles, but does not adapt to changes in time. An another inconvenient is linked to the repartition of the crime. As the crime is concentrated into hot spots, the other cells, even if a chance exist there for a crime to happen, are diluted, and the identification of a precise pattern to follow for the police patrol is difficult.

The two models that follow take the fundamental time aspect into account, and propose a real prediction based and trained on what has happened before.

3.2 Monthly Target Model

3.2.1 Objectives

With this model we want to identify hotspots for a specific crime for the following month. This would help the police select the top priority areas to deploy police resources. A hotspot is defined as an area where there is at least a occurrence (or more, depending on the type of crime and on the police requirements)of the targeted crime. This threshold can be adjusted.

The approach we took was to try to create a model that could adapt to the type of crime we want to prevent. It seems important for the police department to be able to differentiate predictions according to the crime. Deployment possibilities of law enforcement resources are enormous: officers can be by foot, by bicycle, by car, they can use high visibility and low visibility patrols strategies, etc. These methods depend highly on the type of prevention or control the police department needs and therefore they depend on the type of crime targeted. Also, by focusing on one type of crime at a time in the model, we allow for a better identification of crime occurrence patterns and therefore a better prediction. We can always combine predictions to have a broader all-crime-types prediction if necessary later.
Burglary is a typical crime that is targeted by police patrols so we focused our study on this type of events.

3.2.2 Methodology description

The fact that we focus on only one crime type for the target doesn't mean that we should only look at previous events of the same crime type when looking for patterns. Indeed, certain related crimes can signal that the neighborhood is at risk and can be used to predict imminent events. The first step was thus to group all the different crime categories into ten different groups: Assault, Theft, Robbery, Burglary, Motor Vehicle, Arson, Homicide, Vice, Violence and Others. Then we updated our table describing each event with only columns for the year, the month, the cell number and the different crime categories (a

1 indicates which type of crime was committed in the corresponding column) as shown below.

Year	Hour	Month	Cell_Nb	Crime_Arson	Crime_Assault	Crime_Burglary	Crime_Homicide
2015	0	2	7392	0	1	0	0

Figure 15: Illustrative dataframe sample

In the previous model, we were looking at the whole dataset in order to understand a general pattern and estimate when and where are the typical crime occurrences. It has been shown[i] that crime follow spatio-temporal patterns and we wanted to really use these patterns to build a much more accurate predictor.

We used machine learning to go a step further in exploiting the temporal patterns. We made the assumption that future crimes occurrences can be foreshadowed by recent events. It seems evident that crimes committed a long time ago are less relevant than recent ones to predict what will happen in the near future. Nevertheless, it is primordial to take advantage of all the past data we have in order to construct an accurate model. For the learning process we can use crime occurrences during January to predict events in February. Then we use events of February to predict crimes in March. And so on. That way, the model learns the underlying patterns that describe upcoming events, but only uses recent data to make predictions.
Again, our model should adapt to the different crime categories. The temporal patterns of crimes may vary along with the type of crime. For this reason we implemented a method which allows us to choose the duration of the time period we consider to make predictions. We can for example look at crime occurrences from the past three months instead of only the past month, and determine which time period works best for different crime categories and different the machine learning algorithms.
To implement the model's learning architecture, we had to group the events by months of each year and count all the different crimes that were committed in the cell during the chosen time period. The total number of occurrences for each type of event represent the features. The occurrence of at least one of the targeted crime during the following month represents the target.

Figure 16: Example of how to construct features to predict hotspots with a 2-month interest period

It is important to take into account spatial patterns as well. As explained above, when training the model we are looking at each geographical area individually. We count the previous events that happened on the cell and predict if the cell will be a critical area in the following month. To maximize the use of spatial knowledge and count for the influence that neighboring cells can have on each grid cell, we perform a Moore neighborhood average, meaning that we also count the previous crimes committed on all of our cell's neighbours, and divide the total by the total number of neighbours plus one.

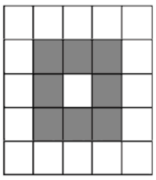


Figure 17: Moore Method

However, when implementing the moore neighborhood method, we obtain lower performances from the model. This is possibly due to the fact that some of the cells where a lot of crimes are committed are close to the geographical limits of the LAPD's jurisdiction. Therefore we are often when taking into account neighboring cells that are only empty because of geographical constraints and this lowers the average when it shouldn't.

Morerover, we observed disproportionate ratios in the binary classified data. This is a common fact in the literature. The unprecision during the fitting can be addresse by rebalancing the training and testing data with a resampling method. As crime records would have been deleted with undersampling, we used an oversampling method. We eventually choose to a RandomOverSampler from the imblearn library, to randomly duplicate input vectors belonging to the minor until the data has equal amounts of each class (50% of zeroes and 50% of ones). The use of oversampling makes the model more robust to

variations in the distribution of zeroes and ones in the input training and testing sets. The robustness of the model is important because it be enaable the LAPD to extend it with future crime records.

3.2.3 Results

In the two following tables are presented scores for two different periods of interest: the 2 previous months and the 8 previous months.
We can observe that for this crime category (Burglary) we have almost always better performances with the Decision Tree Classifier (using the gini indicator and a maximum depth of 5). Overall, we get better results looking at the 8 previous months than the two previous ones except for the Random Forrest's recall score which is the highest. This is important to remark because it means that it is the best model to identify a maximum of hotspots, which is the main interest of the police department.

Table 4: Accuracy of the different Machine Learning Algorithm for an interest period of 2 months

	Decision Tree	KNN	Log Regression
Accuracy	82 %	81 %	78 %
Recall	86 %	89 %	72 %
Precision	83 %	80 %	89 %

Table 5: Accuracy of the different Machine Learning Algorithm for an interest period of 8 months

	Decision Tree	KNN	Log Regression
Accuracy	83 %	82 %	81 %
Recall	84 %	88 %	77 %
Precision	86 %	83 %	89 %

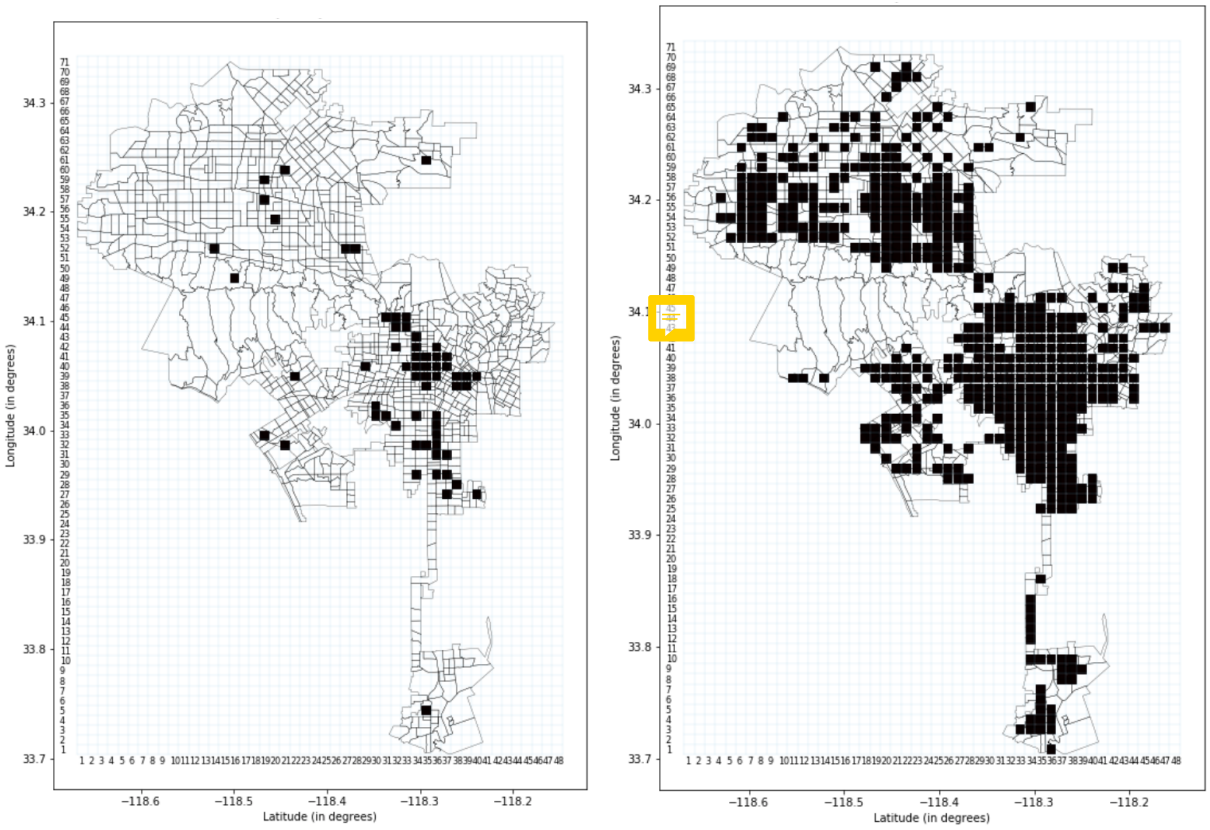
Table 6: Accuracy of the different Machine Learning Algorithm for an interest period of 2 months and with the Moore neighborhood method

	Decision Tree	KNN	Log Regression
Accuracy	77 %	77 %	77 %
Recall	79 %	82 %	74 %
Precision	81 %	79 %	83 %

Here below are presented the scores with the Moore neighborhood which are a bit lower, as described before.

In the two pictures below we can see examples of the final output of the model. They show a map of the LAPD's jurisdisction highlighting all the cells that are predicted as hotspots (with the same "predicted occurences threshold" of 3. We can observe here a

very important thing: crime do vary with respect to the time of the year, something that we weren't able to see during the Exploratory Data Analysis. The police department could base the planning of their monthly activities on the model outputs



Hotspot map for burglaries - March 2016 Hotspot map for burglaries - September 2016

Figure 18: Hotspot map for burglaries

3.2.4 Advantages and limitations

The advantages of this model are multiple. First, this model identifies and uses temporal patterns to predict future crimes by focusing on recent data. It is therefore smarter than a "steady-state" model. Second this model is very adaptable. It can predict different types of crime, the interest period during which the model is looking at data can be modified and the threshold for the number of occurrences from which it determines that the cell is a hotspot or not can be tuned. This allows to fine tune the model depending on the dataset, the environment we are assessing and the needs of the police department, which is a real added value.

However this model can be further improved. It would be interesting to implement the Moore neighborhood method with a way to disregard cells that are outside of the jurisdiction and correctly assess if looking at neighboring cells can increase the performances of the model. Also, this model only allow for a binary classification of the cells. Even if the amount of hotspots given by the model can be regulated by changing the threshold,

it would be better to have a way to rank the different cells in function of the number of predicted crime occurrences using a regression model for example.

3.3 Shift to Shift Model

3.3.1 Motivation

Here we wanted to help patrol that are on the field to react quickly if any pattern happens.
The basic idea of the model is to be able with current crimes happening at time period t to be able to predict if a crime is going to happen at time period t+1.
We also wanted to adapt the model for deadly patrol and just perilous one.

3.3.2 Methodology

We use the same methodology as the previous model, but as the data size is much bigger here, we trained our model on only 1 year : 2016, as it is the latest that we have. We aggregated the different type of crime as before. The difference here is that we don't want to predict a precise type of crime, but the two different model predict if a "deadly" or "perilous" crime is going to occur.
So our inputs on what each model trains are those features, and the output that each model should predict is if a crime of the category "perilous" or "deadly" is going to happen.

Table 7: Data set used for the Machine Learning

Feature	**type pf crime**
Arson	"perilous"
Assault	"deadly"
Burglary	"perilous"
Homicide	"deadly"
Others	"perilous"
Robbery	"perilous"
Theft	"perilous"
Vice	"perilous"
Violence	"deadly"

An example of our data processed and almost ready to be trained can be found below.
Each row represent a different cell at different time shift, and the 10 first columns are the input X of crimes happening at time t, and the last column is the output Y, representing, here in the "perilous" patrol model, how many crime of the category "perilous" happened at time t+1

Arson	Assault	Burglary	Homicide	Motor Vehicle	Others	Robbery	Theft	Vice	Violence	y
0	0	0	0	0	0	0	0	0	1	1
1	1	0	0	0	0	1	1	0	0	2
2	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	0	0	3
0	1	2	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	5
0	0	0	0	0	0	0	1	0	0	2
0	1	1	0	0	0	0	0	0	1	1
0	1	1	0	0	0	0	0	0	0	2
0	0	1	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	1

Figure 19: Example of the dataframe before used to train our ML model

Finally, to simplified our study , we transformed the number of crime happening at t, with 1 if number of crime > 1, and 0 if number of crime $= 0$. So our output y can take value {0,1}

Before we start the training, we observed that there were significantly more output positive than negative. To overcome this, we oversampled. Looking at the distribution of the minority output, we generated new samples to balance the output.

We trained our data using the rule of 85/15% training/testing split. We used 3 different types of ML algorithm, Decision Tree, KNN and Logistic regression. We optimized the parameter to obtain the best accuracy.We also looked at some factors like recall and precision to be sure that our model is relevant and not overfitting.

3.3.3 Result

Table 8: Accuracy of the different Machine Learning Algorithm for the "perilous" patrol

	Decision Tree	KNN	Log Regression
Accuracy	77 %	76 %	70 %
Recall	100 %	80 %	69 %
Precision	72 %	76 %	78 %

Even if the decision tree has the best accuracy, we will go for the KNN method as it's recall value seems more reasonable.

Table 9: Accuracy of the different Machine Learning Algorithm for the "deadly" patrol

	Decision Tree	KNN	Log Regression
Accuracy	72 %	71 %	67 %
Recall	61 %	58 %	65 %
Precision	66 %	65 %	59 %

Here we chose the decision tree because of its high accuracy.

The "perilous" patrol model use the KNN algorithm with an accuracy of 76%.
The "deadly" patrol model use the decision tree algorithm with an accuracy of 72%..

3.3.4 Advantages and Limitations

The main advantage of this method is that it is very reactive during the day and should be able to help the patrol that are already on the street to decide of their next move. It is a very simple model and choose fit easily with the daily routine of the police. This model is the great complement of the 2 models explained before. For the limitation, the model is very basic and it could be interesting to incorporate in our training the trends during the day, or the seasonality. We could talk of our model as a naive model in term of forecasting.

4 Implementation

4.1 Deployment

4.1.1 Tailoring to the LAPD

As explained before, the three models that we have developed are not distinct but rather work as a unit and are complementary. We could include it in a single software to be sold to the Police Department of the city of Los Angeles. The effectiveness of our model relies on an implementation of this software as a whole in the Police Department. Because a lot of pre-processing has been done and because our model is particularly fitted to the prediction of crime in the city of Los Angeles, its effectiveness relies on the close collaboration of the police department at first, in order to set it up. This means that our software would be tailored to the needs of the LAPD. For it to be effective, we need to follow these steps for implementation :

- How it works : We need to explain clearly how our model works to the Police officers and what it can accomplish and what it cannot.
- Set the goals : The LAPD needs to work with us in the definition of its long term goals and targets in order to set an adapted model.
- Fit the results : we need to discuss the possibilities of output that would best fit the LAPD and what would be most beneficial for them.
- Implement the model : the model can be run and the first predictions put into test.
- Assessment of the results : discuss the possible modifications from the first results to improve the fit further for the LAPD.

The Police department is involved in every step of the development of our tool in their work. This procedure allows for a progressive implementation of the tool in order for the Police Department to use it autonomously in the near future.

4.1.2 Steps further

The tool that we developed here could be seen as the first step to a more global and autonomous Predictive Policing program. A new platform could be implemented that provides resource and location management as well as mission planning using the results from our tool. This means that the amount of time officers spend inside the hot areas would be optimized. Also, an analytic and reporting module could be added as en extension to quickly update the model and make it more efficient. This module would make it much easier for police officers to report a crime and enter it in the Police Department database.

Finally, the tool could be made more general in order to be adaptable and usable in other cities. This could be easily implemented as crime data exists and is frequently updated in many cities around the globe. We would have to make a more efficient pre-processing and make the tool much easier to use in order for it to be usable by a wide range of Police Departments in cities of different sizes and in different countries.

4.2 Return on Investment

As previously stated, our predictive tool allows to identify where and when crime is most likely to happen, enabling the Police Department to effectively allocate resource for crime prevention. It is difficult to precisely quantify the scope of the value that could be generated. But it will help insure that each officer's time and means are used in a more efficient way.

One way of getting an estimate of the Return on Investment is by calculating the cost savings from the crime reduction with the help of our tool. It is possible to evaluate the average cost per crime by summing the costs to the victim, the cost for the offender and the cost for the criminal justice system [2]. Looking at Table 10, we have an estimate of the potential savings for one year with the implementation or our tool for the LAPD. The number of crime prevented per week by our tool are fictional values used in order to get an estimate of the total savings.

This return on investment calculation does not cover the better allocation of resources as well as the increase in patrol time that our tool offers because these are much more difficult costs to assess.

Table 10: Potential Yearly LAPD Savings with our predictive Tool

Crime Type	Cost per Crime in $	Crimes per Year	Costs per year in $	Crime prevented per week	Savings per Year in $
Murder	8 649 216	312	2 698 555 392	0,1	44 975 923,2
Rape	217 866	903	196 732 998	0,2	2 265 806,4
Robbery	67 277	12 217	821 923 109	1,7	5 947 286,8
Aggravated Assault	87 238	10 638	928 037 844	1,4	6 350 926,4
Burglary	13 096	18 435	241 424 760	2,1	1 430 083,2
Motor Vehicle Theft	9 079	18 391	166 971 889	1,6	755 372,8
				Total Savings in $	**61 725 399**

It is important to note that the tool that we developed does not replace the experience and intuition of police officers, but is rather an invaluable added tool that allows our police force to use their patrol time more efficiently and helps stop crime before it happens [2].

5 Conclusion & Discussion

Our ambition was to develop a predictive policing tool that could be use by the LAPD to prevent crime and optimize the Police Department resources without using any sort of personal information. To do that, we have developed three complementary models that work on crime prediction at different time scales in order to offer a trustworthy yet reactive crime prediction tool.
Indeed, the first model would allow the police officers to select the critical hot spots that have arisen from the past years, and to build a typical weekly road map. The second one, flexible and adaptive as it can select a particular type of crime, and be trained on a specific amount of month, could improve and guide the prediction. More importantly, it could redirect certain type of police units where they are needed the most, filtering the type of crime within the map. The third one, with a similar methodology but a different training time frame, would finally allow the police patrol to be as reactive as they need to, and to adapt their localisation based on what has happened on the previous time slots of the day.
From the obtained results of our models, it appears clearly that some improvements could be made in order to further enhance efficiency and robustness. It is also important to note that such a tool, while it offers a good insight into the crime dynamics of the city of Los Angeles, is still at a starting phase and lots of additional developments are to be done after it is set in place.
Additionally, while we have made our model without the use of any information about the victims or the offenders, biases can still be present as more crimes are reported in certain communities compared to others. This tool should not be used alone and should be coupled with the experience and knowledge of the police officers on duty. This is made especially clear as the race distribution is very clustered. This is clearly shown in the figure 1, presented at the beginning of this paper. Spacial segregation remains to this day in Los Angeles as Latinos, African Americans, Asians and white non-Hispanic live in separated neighborhoods. Hence, we need to be very careful with the implementation of our tool and the pertinence of the datas we used, as more crimes have certainly been reported into specific areas partly due to the influence of spacial segregation in the patrol routes taken by the police for the eight last years.
Finally, the results and methodology of the different crime detection tools presented in this paper, yet to be improved, are still quite relevant and could be used in a efficient way by the LAPD police to predict the crime and to reduced its occurrence in Los Angeles.

References

[1] Carole Haskins. *Dataset.* URL: https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software.

[2] Predpol. *ROI for predictive policing.* URL: https://www.predpol.com/roi-of-predictive-policing/.

[3] Mark Puente. *LAPD ends another data-driven crime program touted to target violent offenders.* URL: https://www.latimes.com/local/lanow/la-me-laser-lapd-crime-data-program-20190412-story.html.

[4] XXX. *Dataset.* URL: https://www.kaggle.com/cityofLA/crime-in-los-angeles.

[5] XXX. *U.S. Department of Justice National Drug Intelligence Center (NDIC).* URL: www.usdoj.gov.