Crime analysis and trends forecasting

Ghada AMAIRI
Celine DJEDDI
Hala DJEGHIM
Mohamed AZIKIOU

April 2021

1 Introduction

Crime has long been a ubiquitous social problem in human society. Crime prediction is of great significance to the formulation of policing strategies and the implementation of crime prevention and control. In the last few years, Spatio temporal data related to the public security have been growing at an exponential rate. Their use would therefore be a real advantage to reduce criminal tendency by looking for potential factors and trends using prediction algorithms.

2 Related work

There are several Big Data Analytic methods that can be used for crime analysis, authors of [10] chose a semantic approach by clustering the terms frequency and comparing there occurrence and co-occurrence (pattern mining technique). The analysis has been evaluated manually which is time and effort costing.
[5] shows how machine learning techniques can be used on data extracted from geographical information systems in order to make urban crime risk prediction using point of interest data. Results show that point of interest attributes are very useful for crime prediction,making it possible to accurately discriminate between high-risk and low-risk areas.
[11] proposed a model of charge prediction via the judicial interpretation of crime which is considered as a text classification problem. The CPJIC architecture relays on the information about the crime that is extracted, mapped and then extended to give a final prediction. Despite the data imbalance and illegibility, their framework gives better results than the baselines models used for comparison.(CNN, RNN+attention, Fact-Law Attention Model and Few-Shot+Attributes Model)

Articles [7] and [12] discuss the use of learning algorithms to predict crime rates. In article [7] the authors used data from 3 cities (Chicago, Philadelphia, San Francisco) to visualize the data and then compare different learning algorithms. They used the LSTM, the Prophet model and a neural network. Finally they concluded that the LSTM and the Prophet model are the

best for predicting crime rates.
In article [12] the authors also looked at other algorithms such as SVM or the random forest but they also used LSTM. In addition, they have the same conclusion,LSTM was the best algorithm using criminal data from a coastal city in China.
The algorithms established in these two articles are based on the use of crime counts in major cities. Article [1] explains that it is not always easy to obtain data, especially for small towns. It proposes to use a gradient boosting model with predictors of crime. The results obtained predict crimes with an accuracy of about 70%. This may be a new way of predicting criminal tendencies.
Article [13] analyzes the influential factors of crime activities and experimented it on New York City.
Performance of gradient boost decision tree (GBDT), logistic regression (LR), support vector machine (SVM), artificial neural network (ANN) and random forest (RF) are compared to generate an optimum model and an optimal feature set.
1515 different factors ranging from demographic, housing, education, economy, and social of each sample were considered and analyzed.
A recursive feature elimination framework was implemented to support feature selection and feature ranking.
The experimental results demonstrate that the combined GBDT and GIS model can find out the most important factors of crime rate with high efficiency and accuracy.
Finally, An integrated model combining gradient boost decision tree (GBDT), recursive feature elimination (RFE) and geographical information system (GIS) is used to filter, rank and analyze the important features.
results showed that five kinds of features are found to be the most important factors of crime rate of felony assault. They are marital, Black or African American, economic, education and place of interest.
Authors of [8] developed a fast and practical model-selection approach for spatial regression models, focusing on the selection of coefficient types that include constant, spatially varying, and non-spatially varying coefficients. A pre-processing approach, which replaces data matrices with small inner products through dimension reduction dramatically accelerates the computation speed of model selection. The simulation experiments suggest that this approach accurately selects the true model and quantifies numerous hidden effects behind geographic phenomena.
Article [9] summarizes methodologies used in extracting entities and topics from a database of criminal records and from a database of newspapers. The goal of the research was performing statistical and natural language processing methods to extract entities and topics as well as to group similar data points into correct clusters, in order to understand public data about U.S related crimes better.
Article [3] presents a novel approach to better visualize the threats, by identifying and predicting the highly-reported crime zones in the smart city. To this end, it first investigates the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to detect the hot-spots that have a higher risk of crime occurrence. Second, for crime prediction, Seasonal Auto-Regressive Integrated Moving Average (SARIMA) is exploited in each dense crime region to

predict the number of crime incidents in the future with spatial and temporal information.
The experimental results show that the proposed system outperformed compared to state-of-the-art systems with an average Mean Absolute Error (MAE) of 11.47.
Authors of article [4] summarized, and evaluated the state-of the-art for spatio-temporal crime hotspot detection and prediction techniques by conducting a systematic literature review. This study is the premier attempt to critically analyze the existing literature along with presenting potential challenges faced by current crime hotspot detection and prediction systems. This research showed that DBSCAN and Random forest are found to be useful and efficient in terms of accuracy and efficiency. However, several limitations were identified during the SLR that crime hot spot detection algorithm should be; scalable, can deal with sparsity, underlying population,and demographic factors, etc.

3 Problem formulation

In this report our problematic is how can big data techniques give new insights for crime analysis ?. We will therefore visualize certain attributes from a criminal database to search for potential factors influencing crime rates. In addition we will use prediction algorithms to try to predict the number of crimes per day in the city of Montgomery.

4 Solution

4.1 Data analysis

4.1.1 Data management

In order for the the data to be usable, it needs to follow a predefined structure. The one that we are using consists of a list of data records. Each one will have a set of attributes that will be identical in every data record. A set of records represent a crime. We need to clarify that there is some crime records that have several data records. One difference between its records may be the crime type. For instance: the crime with the incident id 201182033 has two records: one for selling drugs the other is for buying them. The attributes of a record describe the place of the crime, its time, criminal type, victim type and the authority that is handling it.

We opted for a relational database management system to implement this data structure. The one that we used is **SQLite3**. It a C-based variant of SQL (short for: structured query language) based database management engine. This choice is due to various reasons which we will discuss next.

NoSQL systems are preferred over SQL systems for their data flexibility. Every record can have have different structure. This saves the time of implementing the data structure in the system because at the beginning they don't possess one. So NoSQL suits best dynamic data.

Our data is static, so NoSQL systems don't find there main benefit of flexibility in our situation. But adds more complexity for retrieving data as their lack of order makes them more costly in term of data query.

In the other hand SQL based sys-

tems are proposing richer and faster query system. And they are a better match for our case as the structure of our data is determined in advance. Its query system makes it easier to make complex queries and to filter the data without having to load them first.

After explaining why the system is SQL based we will cover now why it is the SQLite3. The main reason for this is that it server-less. Because the data is processed locally the data is stored locally too. This will reduce the data loading cost. None of MySQL, SQL server, Oracle, ... are server-less. In addition SQLite3 is a basic library of python, do it is easier to use SQLite3 than any other system.

4.1.2 Data pre-processing

This phase consists of transforming the USA governmental crime database into a usable one for the other phases. It was relatively well-structured before this phase, but can't be used yet.

We developed for pre-processing data a module responsible for mainly three tasks. The first one is data acquiring and parsing data from the US government website into SQLite3 databases. We store all relevant the available including the gaped records. The second role is data loading, the module loads the data from an SQLite3 database, it makes querying easier. The last task is to select data into nparrays. The data can be processed in different ways according the arguments given. It may filter gaped data in selected columns or keep them. This module is highly customizable and non dependent of the database.

We used the previous to remove gaped data to make the prediction and visualisation more reality representing and accurate. The removed data do not exceed 1% of the global data.

Also in data prediction we grouped the data into by time as our data is time-related.

4.1.3 Visualization

figure 1 summarized crime incidents in each year.The number of crime incidents seems to soar since 2017 and reaches its peak in 2018, following certain patterns.

Fig. 6,7,8,9 show the number of crimes by type for each year. As seen, crimes again property are major problems which really increased in 2020 (51,5 % comparing to other crimes)

The hourly trends unravel some interesting crime facts as shown in Fig. 10,11,12,13. As seen, crimes by hour during the four years share similar patterns, where 4-5 am is the safest part of the day whilst 4 am is the most dangerous hour with the most crime incidents reported. Also, 12 pm is also very dangerous during a day. This actually is reasonable as it is the time when more people go out for lunch . As a result, more police resources should be allocated in the shift from noon to midnight.

Fig. 14, 15, 16, 17 are box plots showing the distribution of crimes for each month day, by month.

4.2 Prediction models

A time series is a sequence of data that occur in a successive order over some period of time.

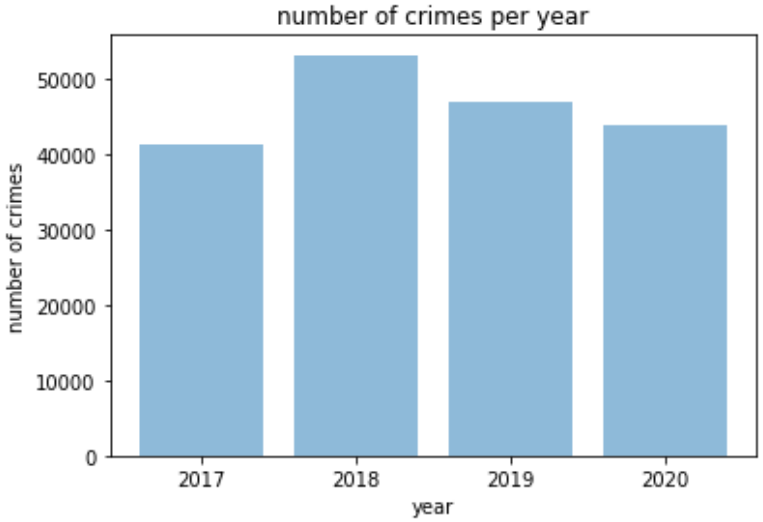

Figure 1: Number of crimes per year

Data visualization phase shows how the number of crime incidents changes over time, which made us think of a potential trend and seasonality.
A trend exists when there is a long term decrease or increase in the data, while a seasonal pattern occurs when a time series is affected by seasonal factors.

4.2.1 Prophet model

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. [6]

It can be considered as a non linear regressive model, defined as :

$$yt = g(t) + s(t) + h(t) + \epsilon t \quad (1)$$

where g(t) describes a piecewise-linear trend (or "growth term"), s(t) describes the various seasonal patterns, h(t) captures the holiday effects, and t is a white noise error term.

Facebook Prophet predicts data only when it is in a certain format. The data frame with the data should have column saved as ds for time series data and y for the data to be forecasted (number of crimes)

	ds	y
0	2017-04-02	89
1	2017-04-03	114
2	2017-04-04	138
3	2017-04-05	137
4	2017-04-06	149

Figure 2: Data format for prophet model

We wanted our model to predict for the next year, so we used
periods = 356 as a parameter.

4.2.2 LSTM model

For the prediction of criminal tendencies, the long-term memory algorithm (LSTM) has been captured in different articles [1,3], indeed it is very suitable for data composed of time series.

LSTM is a member of the neural network family, that can remember past information.
Classical neural networks have difficulty knowing what information to consider when the processing sequence is too long. Therefore, recurrent neural networks were put in place and LSTM falls into this subcategory. They can learn long-term dependencies using a 3-gate, 2-state dialing system.

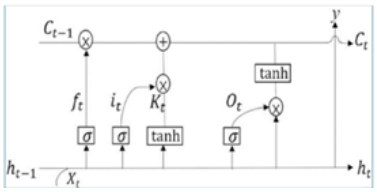

Figure 3: Example of LSTM's cell

The first gate is the Forget Gate, it allows you to select only the important information from the previous step and to concatenate it with the new data. This gate uses the formula:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Then we have the Input Gate, which is used for adding new information, this gate will also sort to keep only the useful data. This sorting phase is done using two functions:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

This gate will then update by adding the data in the memory cell.

Finally, the third gate is the Output Gate, it represents the output of the cell. The latter will use the data in the memory cell and those of the preceding gates to return the new information thanks to the following formulas:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \times tanh(C_t)$$

For all formulas used in a cell ht - 1 and xt are the input values. Wo, Wf, Wi, and WC are the previous value in each layer. bf, bi bo, and bc are degrees of freedom (constant). Ct is the updated cell state.

5 Experiments

5.1 Dataset

To make the predictions, we retrieved the dates of each crime and the identifier in the database. We first delete the identifier present in several times then we counted the number of crimes per day.

This gives us data that goes from 2017 to 2021. We decided to use the data from 2017 to 2020 for training and the data from 2020 to 2021 for the test.

5.2 Evaluation

We have explored LSTM and prophet models to predict crime trends. In order to explore the results of our models, we need to evaluate their performance, we used the Root Mean Square Error (RMSE) for both models.

The RMSE is the standard deviation of the residuals (prediction errors), it is a measure of how spread out these residuals are.

$$RMSE = \sqrt{\sum_{i=0}^{n} (Yi - YPi)^2 / N} \quad (2)$$

Where Y are the true values and YP are predicted ones.

5.3 Prediction

Before starting the comparison between the LSTM and the Prophet model, we need to find the optimal hyper-parameters.

5.3.1 Prophet

For prophet trends forecasting we used the fbprophet python's library. Holidays and changepoints are the most important hyperparemeters that may optimize the prophet model.
Holidays are periods of time where the days have the same effect each year.
We added manually the 10 top dates with the most and least crime numbers, we can notice that this choice of dates isn't optimal for our model.

Holidays	Without holidays
31.25	**30.99**

Table 1: RMSE error comparaison

Changepoints are the points in the data where there are sudden changes in the trend.

Changepoint prior scale parameter allows modulating the flexibility of the automatic changepoint selection. Large values will allow many changepoints, small values will allow few changepoints.

Prophet model by default uses the value 0.05, for our model the best scale value is 0.01.

Scale	RMSE
0.01	**23.06**
0.02	27.65
0.05	34.19
0.1	28.35
0.5	75.04

Table 2: RMSE error comparison of different changepoints scale

5.3.2 LSTM

To make a prediction with the LSTM we used Python's Keras library which already offers an implementation of the algorithm.[2]
For the LSTM, we first look for the optimal epoch, this number represents the number of times the training data has been completely passed through the algorithm. In our case we found a number equal to 200.

Epoch	RMSE
50	22.88
100	22.99
150	21.72
200	**21.61**
250	21.77
300	21.92
350	22.04
400	23.89

Table 3: Comparison RMSE errors for different epoch

The second hyper-parameter we looked for is the number of neurons in the LSTM. The optimal number of neuron is 35.

Number of Neuron	RMSE
5	24.38
15	22.03
25	24.00
35	**21.45**
45	21.67
55	21.87
65	21.92
75	22.69

Table 4: Comparison of different number of neuron

5.4 Training dataset

The efficiency of the algorithms also depends on the training period, which can be longer or shorter. We tested 3 training periods for each algorithm. The results are presented below.

	1 year	2 years	3 years
LSTM	23.57	23.06	21.4
Prophet	31.64	36.42	23.06

Table 5: Comparaison of training period for LSTM model and prophet model in term of RMSE error

We observe that each algorithm has an optimal training period. For the LSTM and the Prophet model the lowest RMSE is when the training period is 3 year.

We also visualized the results of the predictions for the two models for a training period of 3 years.

The figure 4 shows the results of a prediction with the LSTM. In blue we have the whole dataset,in orange we have the prediction with the training data and in green we have the prediction for the test data.

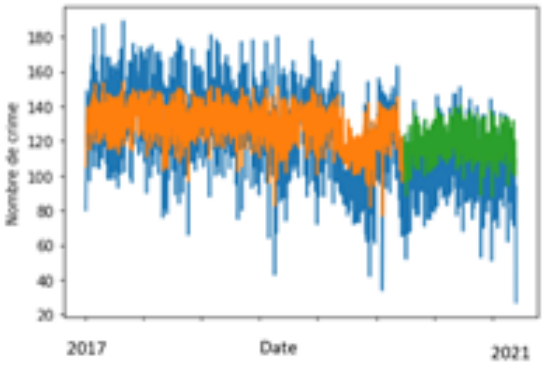

Figure 4: Crime prediction with LSTM

The figure 5 shows the forecast result for the prophet model. The light blue is the uncertainty level, the dark blue is the prediction and the black dots are the original data.
From these results it can be concluded that the LSTM model is more suitable for predicting crime rates in the town of Montgomery.

6 Conclusion

To conclude, we can say that big data is very interesting in the field of crime. First of all, thanks to them we were able to discover various very useful information on criminal tendencies. For example, we can observe that there is more crime at certain times of the day, by having this information it is possible to mobilize more security personnel to manage this and therefore to reduce crime rates.
Then we saw that it was possible to use prediction algorithms to predict the number of crimes. On the criminal data of the city of Montgormy, we observe predictions that are quite close to reality. So getting predictions directly is also possible thanks to big data. To take this project further, it is possible to test the algorithms used on criminal data from other cities.

References

[1] Predicting spatial crime occurrences through an efficient ensemble-learning model. *ISPRS International Journal of Geo-Information*, 9(11), 2020.

[2] Jason Brownlee. Time series prediction with lstm recurrent neural networks in python with keras, Aug 2020.

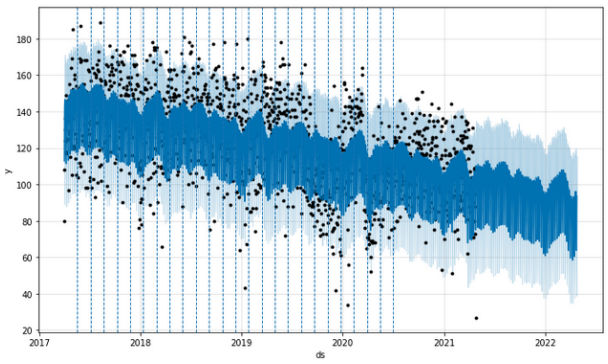

Figure 5: Crime trends using Prophet model

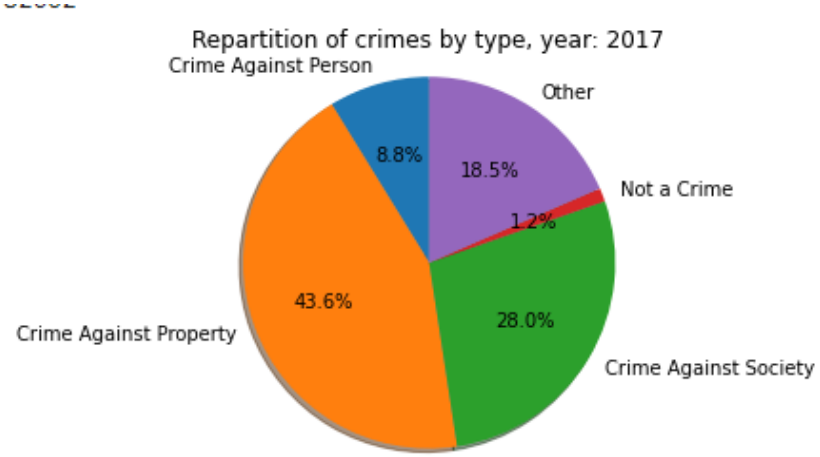

Figure 6: Crimes by type: 2017

[3] Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir, Tieng Wei Koh, and Hafiz Husnain Raza Sherazi. Spatio-temporal crime predictions by leveraging artificial intelligence for citizens security in smart cities. *IEEE Access*, 9:47516–47529, 2021.

[4] Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan, Mubashir Ali, Anees Baqir, and Hafiz Husnain Raza Sherazi. Spatio-temporal crime hotspot detection and prediction: A systematic literature review. *IEEE Access*, 8:166553–166574, 2020.

[5] Paweł Cichosz. Urban crime risk prediction using point of interest data.

[6] Facebook. Forecasting at scale.

[7] Mingchen Feng, Jiangbin Zheng, Jinchang Ren, Amir Hussain, Xiuxiu Li,

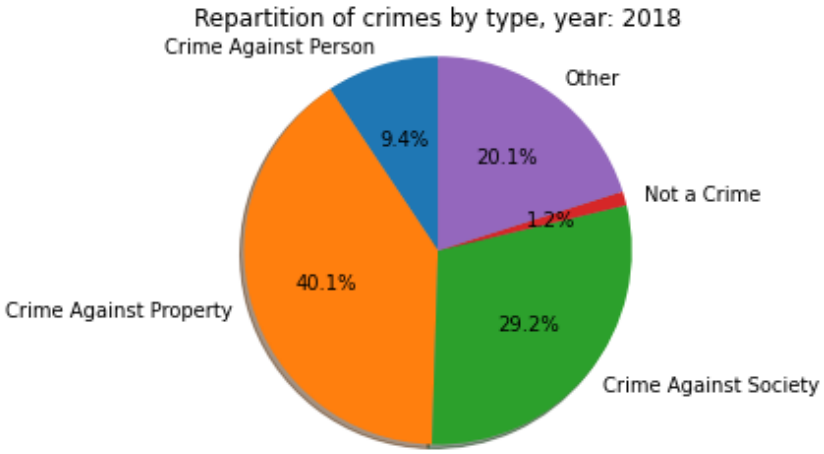

Figure 7: Crimes by type: 2018

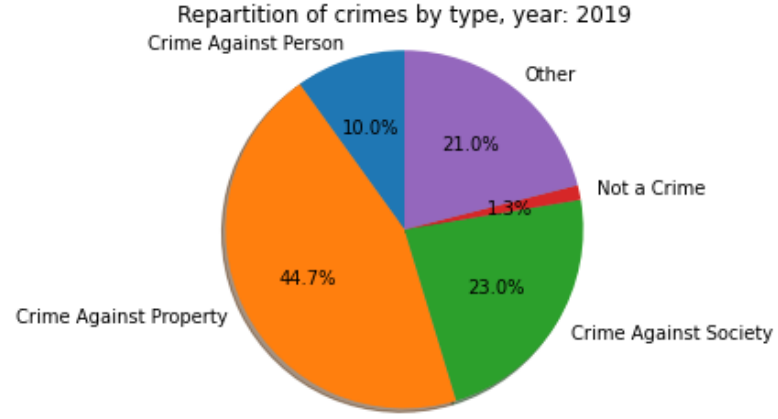

Figure 8: Crimes by type: 2019

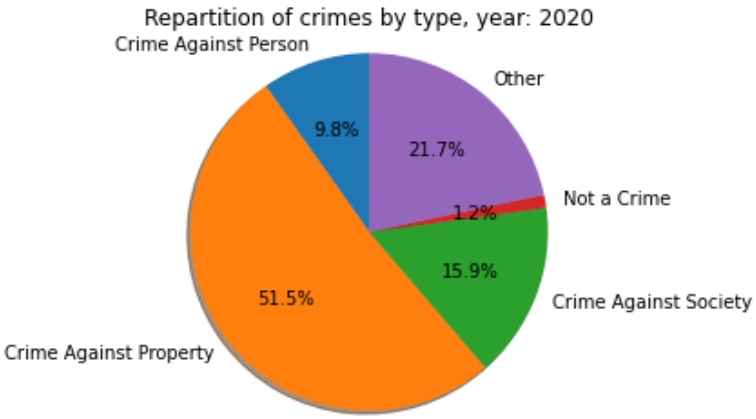

Figure 9: Crimes by type: 2020

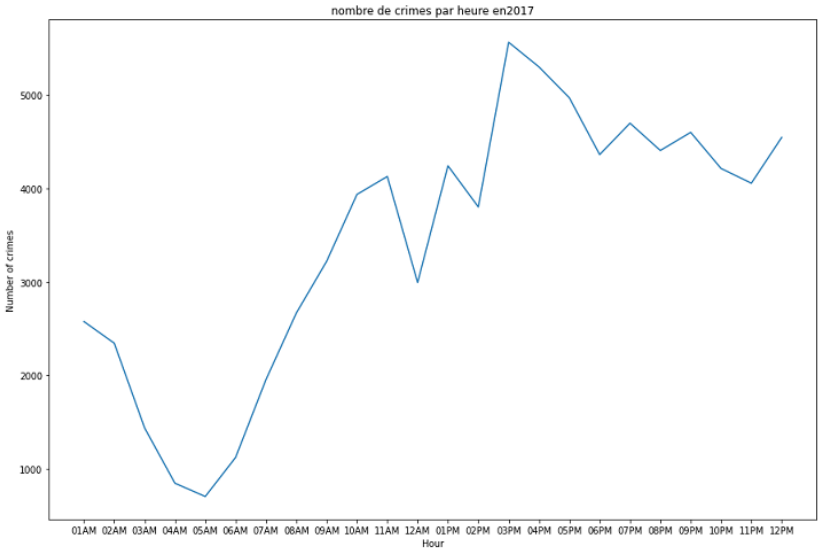

Figure 10: Number of crimes per hour:2017

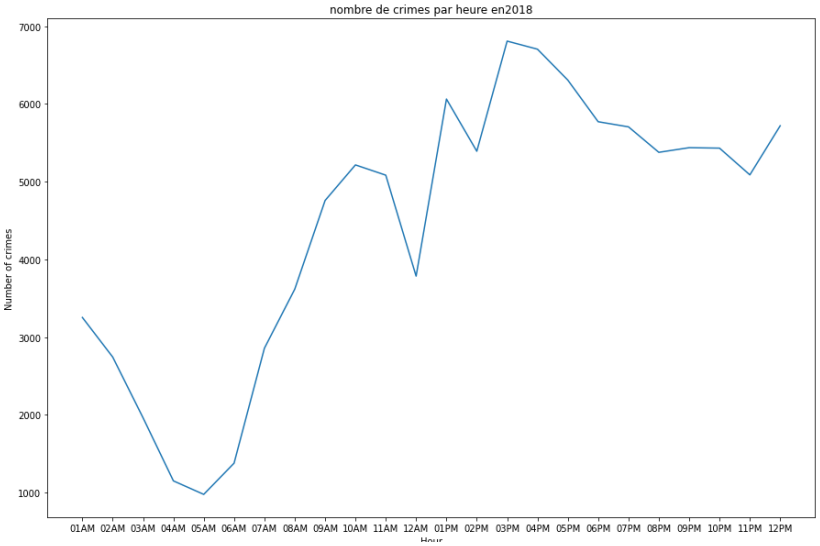

Figure 11: Number of crimes per hour:2018

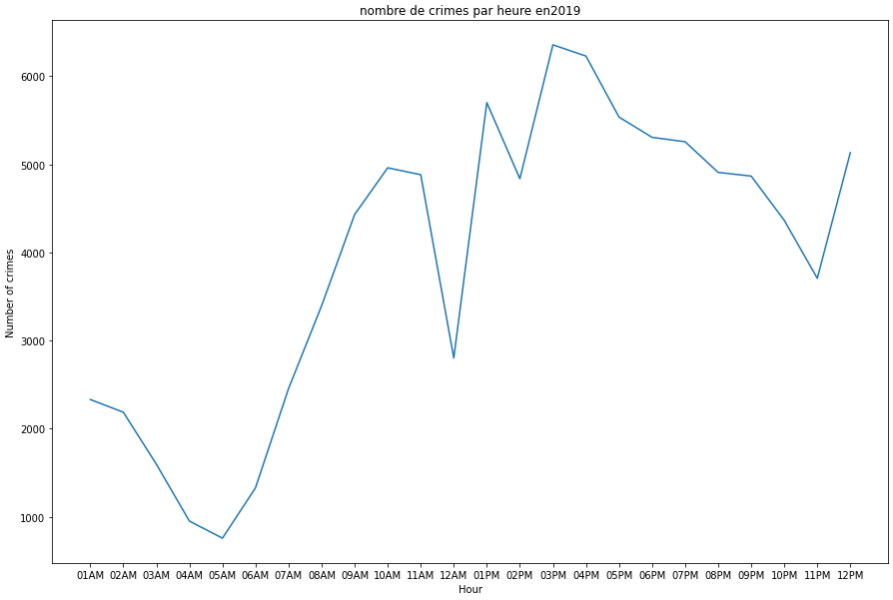

Figure 12: Number of crimes per hour:2019

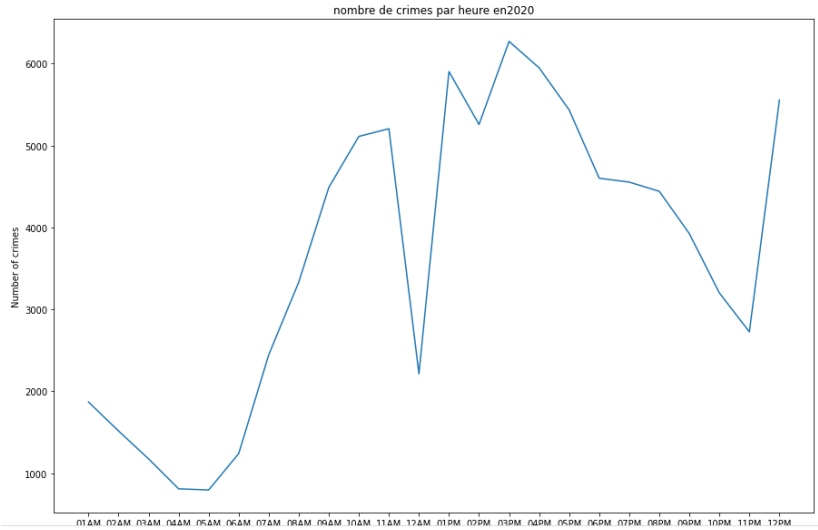

Figure 13: Number of crimes per hour:2020

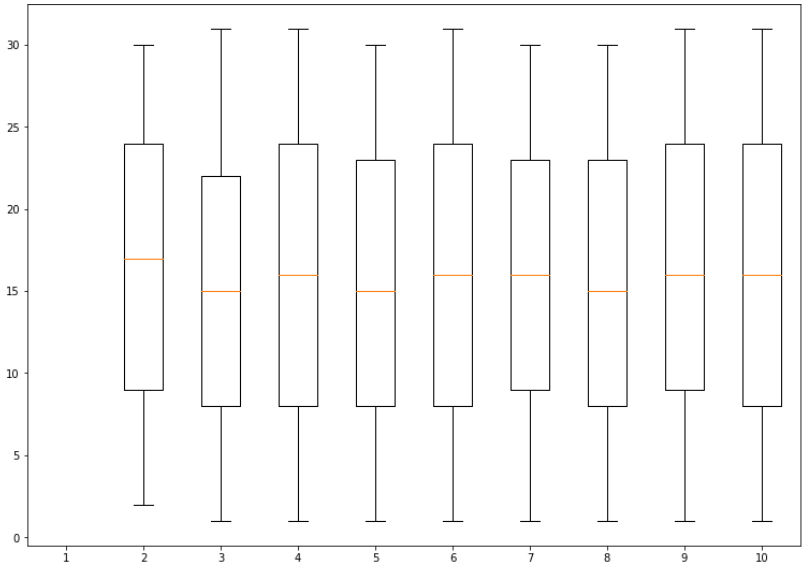

Figure 14: Box plot number of crimes by month:2017

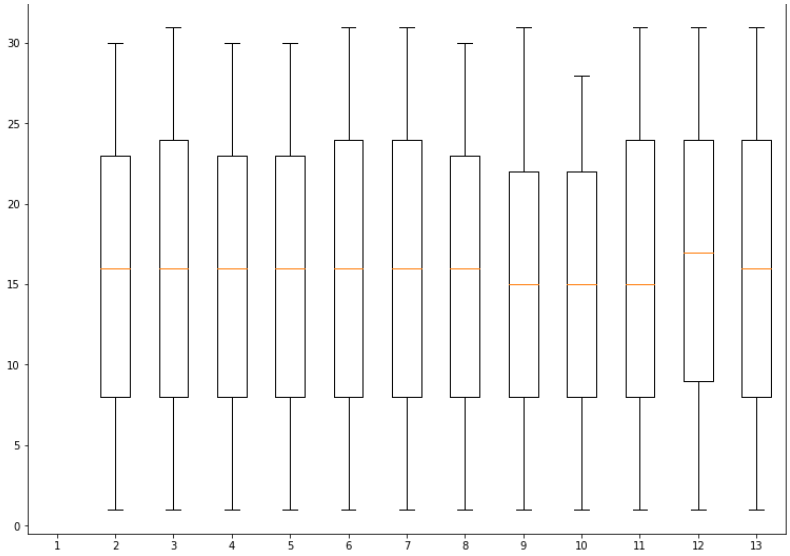

Figure 15: Box plot number of crimes by month:2018

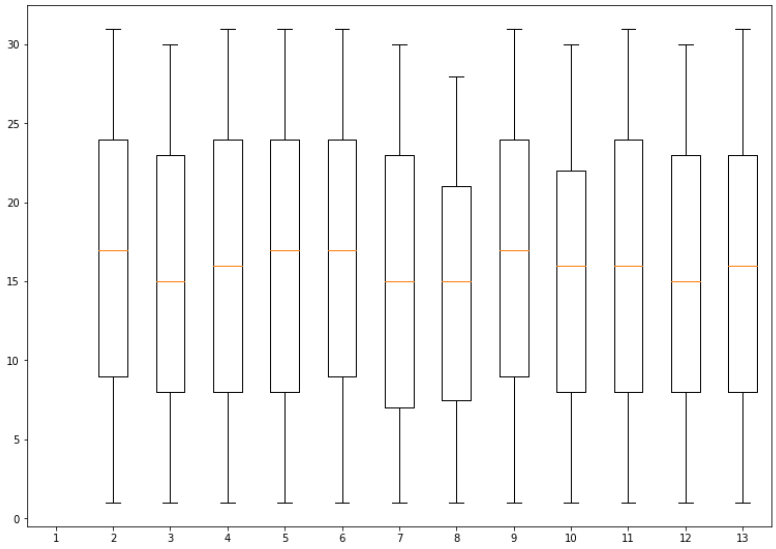

Figure 16: Box plot number of crimes by month:2019

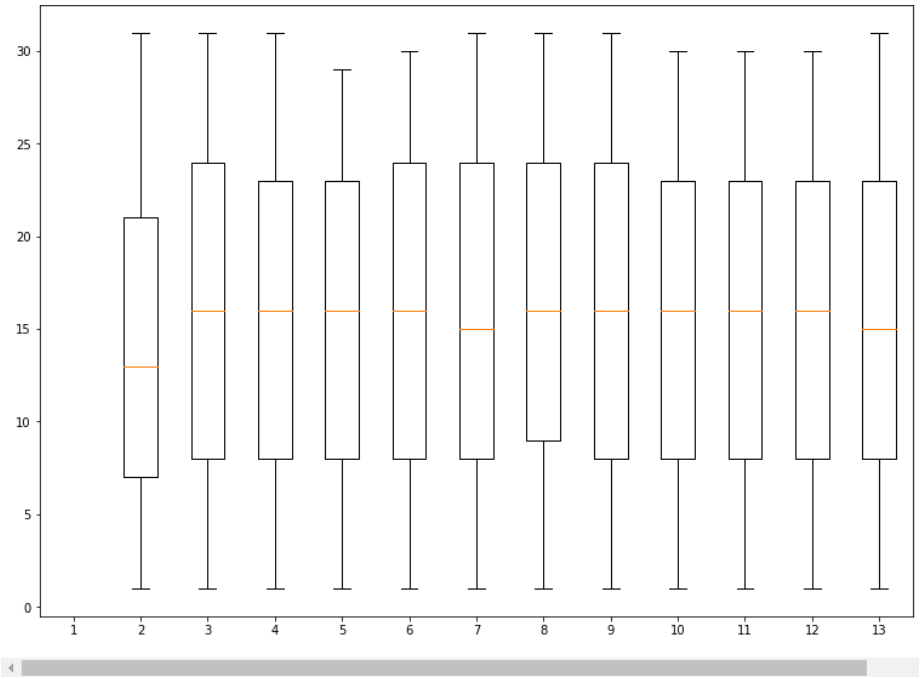

Figure 17: Box plot number of crimes by month:2020

Yue Xi, and Qiaoyuan Liu. Big data analytics and mining for effective visualization and trends forecasting of crime data. *IEEE Access*, 7:106111–106123, 2019.

[8] Daisuke Murakami, Mami Kajita, and Seiji Kajita. Scalable model selection for spatial additive mixed modeling: application to crime analysis. *ISPRS International Journal of Geo-Information*, 9(10):577, 2020.

[9] Quang Pham, Marija Stanojevic, and Zoran Obradovic. Extracting entities and topics from news and connecting criminal records. *arXiv preprint arXiv:2005.00950*, 2020.

[10] Azhar Ali Shah Sehrish Nizamani Saad Nizamani Imdad Ali Ismaili Sarwat Nizamani, Nasrullah Memon. Crime analysis using open source information.

[11] (Member IEEE) CHENWEI WANG1 LIJUN DONG HONG YAO (Member IEEE) XINCHUAN LI, XIAOJUN KANG and SHIXIANG LI. A neural-network-based model ofcharge prediction via the judicialinterpretation of crimes.

[12] Xu Zhang, Lin Liu, Luzi Xiao, and Jiakai Ji. Comparison of machine

learning algorithms for predicting crime hotspots. *IEEE Access*, 8:181302–181310, 2020.

[13] Jianming Zhou, Zheng Li, Jack J Ma, and Feifeng Jiang. Exploration of the hidden influential factors on crime activities: A big data approach. *IEEE Access*, 8:141033–141045, 2020.