Spatio-Temporal analysis of crime in London

Vithurshan Vijayachandran

Abstract— An effictive visiual analytics approach to identify how frequently crime occurs during certain time of the year, how the boroughs are affected by the temporal varitons of crime and how youth influence the occurance of crime. The purpose of the study is to allow the Mentropiltan Police Service to anticipate crimes and allocate the right amount of forces to given location at the given time to prevent crime and save lives. In this report author explores the spatiotemporal dynamics of crime in London. Used over 25,000 data to identify trends across time and space.

1 Problem Statement

Crime is a surging social problem in major metropolitan cities around the world, and London is no exception. With increasing rates of victims, London has reached its peak, thus far, in this dangerous domain of social concerns; reaching a record high knife crime with 15,023 offences in a year.[1] Warnings were raised by the Mayor of London with increase of evidence showing violence has been normalized by youth of London[7] and the consequences of this violence are often drastic, leading to severe injuries and even loss of lives.

Crime is not randomly located,[3] identifying locations more vulnerable to crime during certain time of the year and identifying boroughs vulnerable for youth crimes will allow the Metropolitan Police Service (MPS) to produce counter plans to prevent crimes.

In order to meet the aim, visual analytics techniques will be applied to answer the following questions:

1. Do crimes occur more frequently during certain time of the year?
2. How are boroughs affected by the temporal variations of crime?
3. What is the influence of youth in London crimes?

The following dataset will be used:

1. Recorded Crime: Geographic Breakdown (2020) [2]
2. London Census Data (2011) [8]

The datasets used are authentic as they are retrieved from official sources. The crime dataset published by MPS records crime by geographic breakdown, on a monthly basis, according to crime type. The census dataset contains population information by age group and other socio-demographic factors.

2 State of the Art

The spatial-temporal mapping is a topic often discussed when studying visual analytics. When concerning the study of crime rates, various methods have been developed and applied in visualising patterns over space and time of these crimes. This provides a better understanding of criminality as well as predictability in anticipating crime.

The spatial-temporal study of crime by Cheng and Williams (2012) involved the clustering method, together with visual inquiry tool, Geovisualization (GeoViz) when studying the real-time police dataset of London Borough of Camden. Exploratory Spatio-Temporal Data Analysis (ESTDA) of the study focused on cluster detection and analysis, where Kernel Density Estimation (KDE) was implemented, translating a specific geographical point into a density surface. KDE is mainly affected by the cell size and bandwidth parameters; a higher bandwidth is more efficient in analysing wider spaces (by providing a smoother density surface), whereas a lower bandwidth is apt in identifying hotspots. The authors realised that clustering was a good approach as it depended less on input arguments. Using SatScan for spatial analysis, results determined 7 main and 13 supplementary hotspots in a particular neighbourhood. Clusters were also found to spread out over a time of 14 days. The methodologies by Cheng and Williams are applicable in different datasets by fine-tuning the spatial and temporal parameters. It is understood that spatial scans identify spatial-temporal clusters aptly.

Focusing on temporal distribution, Corman and Mocan (2000) carried out a high-frequency time-series analysis of crimes, deterrence and drug use in New York City. The dataset concentrated around five types of crimes and drug-use and police force were fundamental factors in their study. The authors applied unit root tests for break points of segmented trends and investigated changes in crime between years 1970 to 1996. An optimal lag length was chosen for each variable and natural logarithm were taken out before differencing. Examining the trends, the authors had successfully distinguished the peak of the crime epidemic, that is in 1990, with a 24% increase from its predecessor peak. Overall, results from this study show that there is a positive correlation between drug use and robberies and burglaries throughout the duration studied. Also, with time, as the police force increased, crimes tended to take a downturn. The time-series method was beneficial to the authors in solving the simultaneity issues of their study when looking at multiple's factors of crimes. Comparable to their study, this project investigates factors such as the role of boroughs, seasonal changes and poverty in knife crimes. Therefore, time-series analysis will be implemented to dive deeper into the properties of this criminal behaviour. As this project is intended to study trends of crimes in London, similar to the study by Corman and Mocan, a time-series analysis will be adapted to observe changes in it the rate of crimes.

The approaches taken by the reviewed literatures are very useful for the data to be used in our study and will help towards the aim of this project.

3 Properties of the Data

MPS publishes recorded crime data by geographic breakdown, on a monthly basis, according to crime type which can be retrieved from London Datastore.[2] The data is provided at borough level (33 boroughs) to be used to investigate the spatial distribution of the crimes. The data covers the period from March 2018 to February 2020, it will be broken down and inspected by months to distribute the time series of crime as it didn't contain information regarding the time crime took place and day. The dataset doesn't include crime data for City of London as it is managed separately by the City of London Police.

The London census dataset is collected by ONS every 10 years with the dataset used for this study being taken in 2011.[9] The dataset contains population by age group under different categories. For the purpose of our studies the following age categories will be used: *15-16*, *17-18*, *19-21* and *22-24* as 15 to 24 is defined as youth according to United Nations. [10]

The central subject matter of this study surrounds the topic of crimes in London. Therefore, the following 11 major type of crimes were investigated: *Arson and criminal damage*, *Burglary*, *Drug offences*, *Miscellaneous crimes against society*, *Possession of weapons*, *Public order offences*, *Robbery*, *Sexual offences*, *Theft*, *Vehicle offences* and *Violence against the person*. The 54 minor crimes investigated entails a lower-level crime type, specific to a major crime listed above. For example, for the main crime of drug offence, the minor crimes include drug trafficking and possession of drugs which is typically used to describe the crime type by the officers. For this reason, minor crimes were removed from our study.

The data to be analysed in this study represents the crimes that occurred in the boroughs of London. For representing the spatial distribution of crimes, the shapefile format was used as it's a "geospatial vector data format for geographic information system (GIS) software".[11] This method was implemented to gain geographical information on the longitude and latitude of each location studied as MPS crime dataset didn't contain the information.

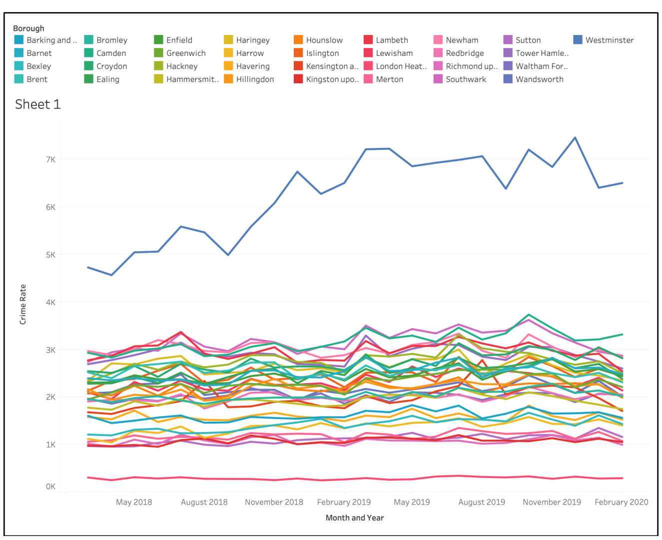

Fig. 1. Spatial variation of variables

The MPS crime data and census data did not contain any null values. The datasets were merged using the borough names but as the census data did not contain space between each word in the name, e.g. BarkingandDagenham instead of *Barking and Dagenham,* manual filtering was required to overcome the issues.

It is expected that some boroughs will have higher crime number compared to other but in order to understand the spatial distribution of values, Fig. 1. was used. It is clear that the crime in *Westminster* is nearly double that of crimes in other boroughs of London. Due to the nature of the dataset and our study, this cannot be considered as outlier or simply ignored.

4 Analysis

4.1 Approach

Python is the software tool that will be mainly used for data pre-processing, feature engineering, data analysis and visualization for this study. Tableau will also be used for visualization as drag and drop features will allow the visuals to be implemented easily and quickly while resulting in interactive and easy to interpret graphs.

Aggregate level results will be used for temporal and spatial analysis as number of crimes plays a major role in finding answer to the questions this paper aim to answer.

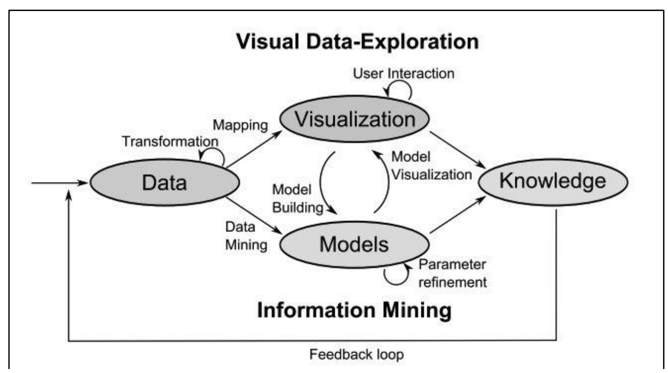

Fig. 2. Visual Analytics Process by Keim et al.

Visual analysis process proposed by Keim et al in Fig. 2. will be applied which shows the effective process of using computational power and visual analytics to extract and interpret insights. This process also involves reiterating each milestone if necessary. Data will be cleaned after each visualization if required, and previsualized data would be applied to the models and models will be refined at key stages if necessary, as shown by Keim's diagram.

For temporal analysis, time series decomposition will be used to search for trends, seasonality and patterns. Initial expectation is that there will be increase in crime over the last 24 months and crime numbers are expected to be at their peak during the holiday period (summer and Christmas) compared to rest of the months as identified by Statista. [6] Due to data being recorded on monthly basis and crime values being available for the last 24 months, temporal and time series analysis will be investigated per month rather than yearly.

For spatial analysis, longitude and latitude retrieved from the shapefile will be used to visualize the crime rate over the boroughs. Temporal values of crimes will be used to view the sequence of spatial distributions over boroughs at different time of the year. Space (borough's) will then be selected and temporal variation of that space will be investigated.

Partition-based clustering will be applied to the set of the spatial distribution of crime values which will then be used to investigate the temporal distribution of the cluster membership. K-means clustering will be used cluster local spatial variations (values associated with different locations at the same time step) by similarities of multiple attributes which will not be visually easy to identify otherwise.

In order analyse the influence of youth in crime in London, percentage of youth in each borough will be calculate using the following formula:

$$youthRatePerBorough = \frac{\text{sumOfYouthAtBorugh}}{\text{sumOfTotalBoroughPopulation}} * 100$$

sumOfYouthAtBorough will get the total number of individuals in the age range of 15 to 24. Youth rate will then be used to compare borough by the percentage of youth. Values of crimes will be used to view the sequence of spatial distributions over boroughs based on type of crime to view the influence of youth in London crimes.

4.2 Process

Fig. 1. Shows increase in crime rate borough over the last 24 months so the process was initiated with time series decomposition in search of initial seasonal patterns. It was created by looking at the total value of crimes per month from March 2018 to February 2020 as shown in Fig. 3.

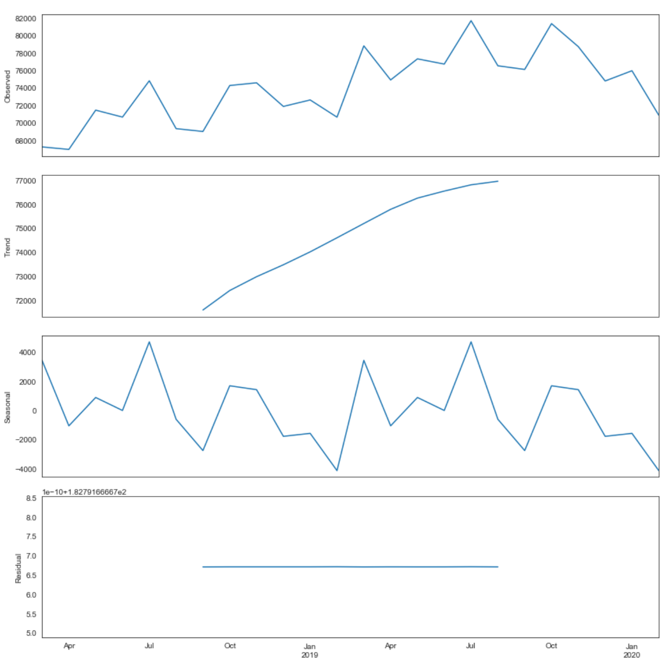

Fig. 3. Additive timeseries decomposition

Time series decomposition was done using two methods, additive and multiplicative. Residuals in multiplicative however were not centred on zero and as the variation around the trend-cycle does not vary with the level of the time series it was deemed that additive is more suitable for our analysis.[5]

As expected, additive model shows the increase in crime with number of crimes in March increased by 12,000 from 2018 to 2019. This pattern is repeated through the series as each month saw crime rate rise compared to the year before. However, number of crimes in February in 2019 and 2020 remained in the region of 65,000 which may indicate crime rate starting to fall but it is not shown in Trend so it can only be proved by retaining data for following month which is not possible for our study.

Over the last 24 months, the high number of crimes have been in July followed by October and November. The high number of crimes in July is understandable as it's during the middle of summer and may attract a lot of tourists. but there is also a huge drop in the following month (August) which is usually the peak of summer holidays for locals due to school holidays, but this could also suggest many locals travel out which leads to the lower crimes. Seasonal trend also shows the volume of crime is high during the month of March. Looking into how type of crimes varies over time will allow us to better interpret the seasonal patterns.

Therefore, correlation matrix was created to identify what type of crime contributed to towards the temporal variation of crime as shown if Fig. 4.

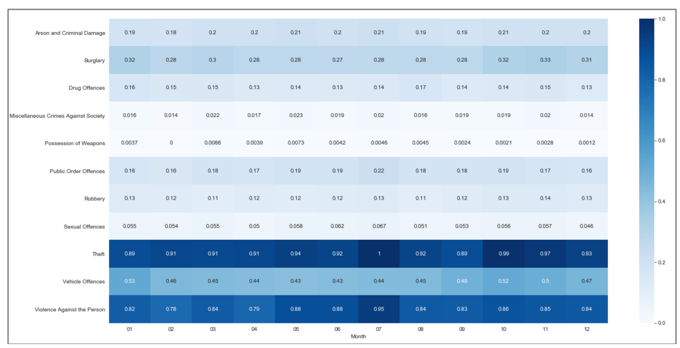

Fig. 4. Correlation matrix for crime types

Early indication is that the London crimes are dominated by two type of crimes: *Theft* and *Violence Against the Person*. Most crimes occur more frequently in the month of July followed by October and November as shown earlier by Fig. 3. Two dominant crimes are also at their peaks in July which further proves pick pocketing and street robbery is a huge concern for London tourism. [12] However, some type of crimes does differ as *Vehicle offences* at its peak in January and *Burglary* is at its peak in November. The standout one is *Drug offences* at its peak in August when all the other crimes see a drop in that month.

Following on, author-initiated the analysis of spatial distribution of temporal variations by displaying the number of crimes by borough over 12 months. Author decided to combine the data by month due to earlier analysis showing monthly patterns were identical. The results can be seen in Fig. 5.

Results show crimes are high in central part of London and low in outer part of London for all months. *Westminster* is the dominant borough with high number of crimes while *Sutton* and *Bexley* are two of the boroughs which saw consistently low amount of crimes every month. This is a clear indication that even though number of crimes varies per month, overall temporal variation of crimes has no effect on spatial distribution as when the crime rate does drop, it tends to drop throughout the city. However, Fig. 3. showed that the number of crimes increased over the last 24 months but Fig. 5. was not able to show that which indicates combining the data by month doesn't show the full picture.

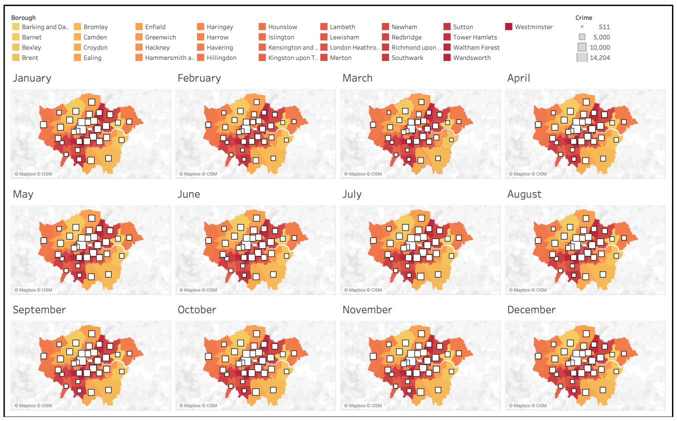

Fig. 5. Spatial distribution of temporal values

In order to have a better understanding of the spatial distribution of crimes, Fig. 1. can be used. It further proves how the crime in *Westminster* is nearly double that of crimes in other boroughs of London and also shows the number of crimes saw a 40% increase within 3 months (from September 2018 to December 2018) in *Westminster*. Even though few other boroughs saw an increase in crime for the same period, it wasn't to the same extent. Further analysis with space as a whole is required to understand and interpret the rise in crime for that period.

Overall, it is clear that temporal distribution is identical with most of the boroughs. *London Heathrow and London City Airports has* a very little crime rate with barely any temporal distribution which is understandable as it covers a very small area and with minimal residency and high security. *Kensington and Chelsea* is the borough with a unique temporal distribution compared to other boroughs with the number of crimes increasing in the month of August in 2018 and 2019 whereas all other boroughs saw the number of crime drop, therefore further analysis is performed below to understand the current results.

In order to analyse the spatial distriubtion of crime, we cannot rely on human interpretation to identify location that share a lot similarity with number of crimes. Therefore, K-means was applied to cluster local spatial variations by similarities of multiple attributes (type of crimes). In order to apply K-mean clustering, sillhoutte plot was created to identify the optimal number of clusters but failed to do so as results showed there are negative values in all clusters which suggests there are points that don't belong in that cluster. Therefore, Elbow method was implemented to determine the optimal number of clusters for k-means clustering and it showed clear elbow at 4 so it was selected as the cluster number to apply to K-means clustering. The results of this is shown in Fig. 10.

Westminster was the only borough in cluster 1, Boroughs north, east and south east of *Westminster* were also clustered together in *cluster3* which further indicates crime in central part of London is high as mentioned in our findings earlier. *cluster2* dominated outer London as boroughs from north, west and south east were clustered together. In order to interpret the clusters Fig. 6. was created.

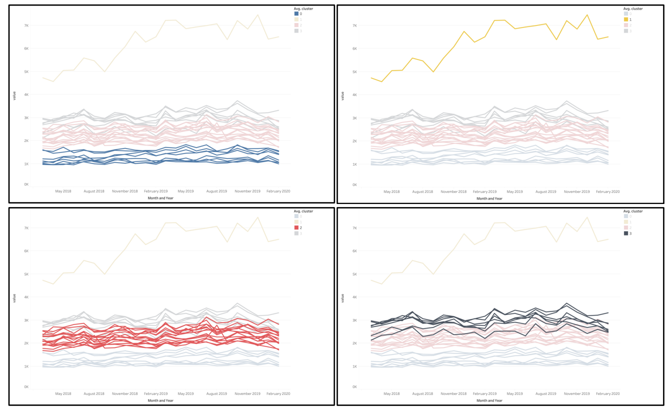

Fig. 6. Temporal variations of clusters

This shows boroughs in *cluster0* generally have lower number of crimes, boroughs in *cluster2* have identical temporal variation and crime number is in the middle range and boroughs in *cluster3* have higher number of crimes and the temporal variation varies a little more between the boroughs compared to the other clusters. In order to analyse each cluster further, one borough per cluster was selected to analyse the temporal variation of type of crimes to further interpret the results which is shown in Fig. 7.

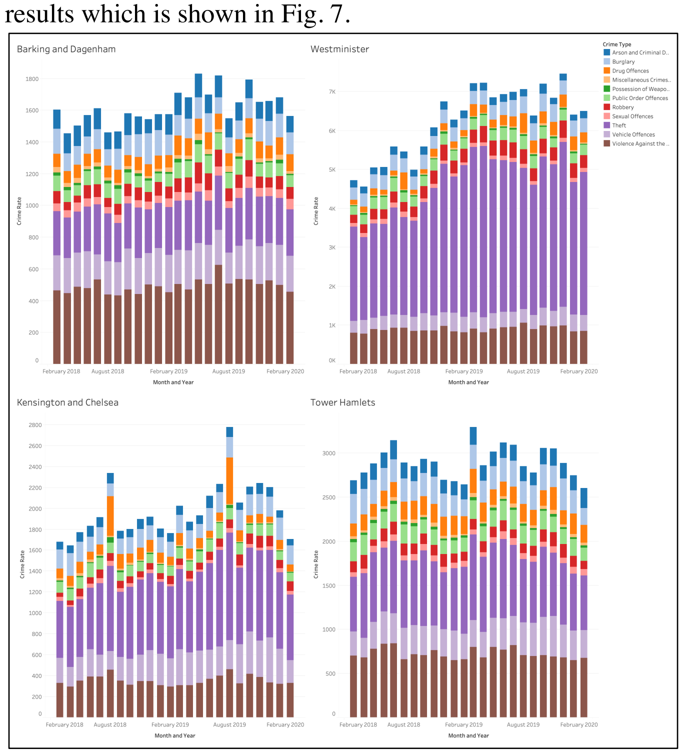

Fig. 7. Temporal variations by type of crime

Tower Hamlets was chosen from *cluster3* as it is the borough with highest youth population which could lead to finding some initial answers to one of the aims of the study. *Kensington and Chelsea* was chosen from *cluster2* to further analyse earlier finding of unique temporal variations in the month of *August*. *Barking and Dagenham* was a random choice from *cluster0,* and *Westminster* was the only pick from *cluster1*.

It is clear how the type of crime plays a major role based on the location of borough as *Wesminster* and *Kensington and Chelsea* (boroughs based in central london) are dominated by *Theft*. It reduces in *Tower Hamlets* but still significantly high as its more centraly located too and it drops further in *Barking and Dagenham* which shows *Theft* crimes are lower as boroughs are further away from central as those boroughs attracg little tourists.

Crimes in *Barking and Dagenham* and *Tower Hamlets* are dominated by *Violence gainst the* Person which shows violence are high in more residential areas.

Kensington and Chelsea shows a high rise in *Drug Offences* in the month of August in 2018 and 2019 along with *Violence gainst the Person* and *Theft* also was it highest in August 2019 which provides some insights into our earlier findings. ‘Notting Hill Carinival’ takes place every year in the August which explains the increase in those crime.[13]

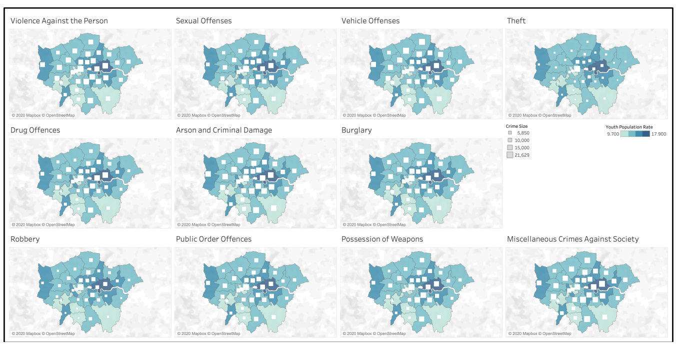

Fig. 8. Influence of youth on crime

Following on, percentage of youth per Borough was calculated and merged with crime data. Percentage of youth was filtered into four different categories with following range; *Range1: 17+, Range2: 14 to 17, Range3: 11 to 14* and *Range4:* < 11. *Tower Hamlets* and *Newham* were the two boroughs with high number of youths with both above 17% while *Richmond upon Thames* was the lowest as 9.7% of its population was youth. In order to view the sequence of spatial distributions over boroughs based on type of crime to view the influence of youth Fig. 8. was used.

Drug Offences and *Miscellaneous Crimes Against Society* are the two crime types significantly influenced by youth as areas with higher youth population are more vulnerable to those crime types. This could further explain why *Drug Offences* high in *Kensington and Chelsea* in the month of *August* as students are on their summer break and Notting Hill Carnival also takes place which attracts the youth.[13]

It also shows *Theft* and *Robbery* have very little correlation with youth population which further proves it is cause by tourism.

4.3 Results

Crime is clearly widespread in particular boroughs of London as shown in Fig. 10. The hotspots are based around the City of London with *Westminster* affected the most by crime. The cold spots are located East, South West and North West of London.

Crime do occur more frequently during the month of *July* for majority of London boroughs as shown in Fig. 9. followed by the months of *October* and *March* as shown in seasonal trend of timeseries decomposition.

Fig. 9. Temporal variation by spatial distribution

Overall, temporal values of crime do not affect the spatial distribution as when the crime number increases at given time, it increases throughout the city of London. However, this is not the case for *Westminster* and *Kensington and Chelsea* as they see high number of crimes during the month of August due to external factors such as tourism and events which was clear when looking at temporal variations of the type crimes for the chosen boroughs.

Drug Offences and *Miscellaneous Crimes Against Society* are crimes with high influence of youth. *Robbery* and *Theft* have no correlation to youth which confirms violence is normalized by the youth of London.[7]

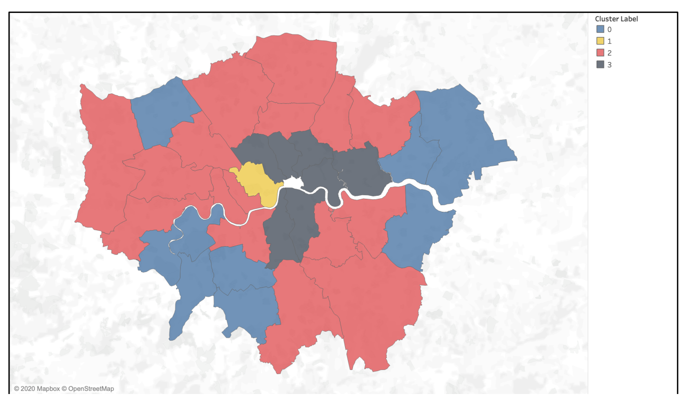

Fig. 10. Cluster of boroughs sharing similar relationships

5 Critical Reflection

Crime in London has always been a major concern in recent time due to the negative image it brings on one of the most successful financial cities in the world. The study aimed to find when and where the crime occurs more frequently and the type of crime to help the officials take measures in preventing the crime from taking place.

The finding of the study shows an effective approach was chosen for our investigation. Effective use of visualization along with the spatio-temporal analysis help identify time and location crimes occur more frequently in London. The process of reiterating computational and exploratory visual exploration contributed heavily towards effectiveness of the investigation.

Even though the study managed find the months more vulnerable for the frequents of crime, it may contribute significantly less towards the prevention of crime due to the nature of temporal information. If the time and date of the crime was made available (not made available for security reasons), further study could be done to look at time and day crimes frequently occur which could help the officials take much better measures as less resources will be required to prepare for time period in a day compared to a whole month.

To further investigate the spatio-temporal variations of London crime, extended analysis of over the period of 5 years will provide better opportunities to find seasonal pattern as the data used only covered the period of 24 months from 3 different years. Could also look at the spatio-temporal distribution by ward for each borough to have a better understanding of the behaviour of each borough. This could also help identify focal points of crime which could allow the official to be better prepared as smaller areas to cover.

Even though the study looked at some of the external factors that could have contributed towards odd pattern in our findings, further study is required back our reasonings. For the partition-based clustering, K-means was used which resulted in good finding. However, comparisons of results with other models such as DBScan and EM would allow the study to judge to what extent clustering was successful.

The census data used to retrieve the youth population was collected in 2011 which further questions the credibility of the study as youth rate in Tower Hamlets was 17.2% whereas 26% in 2019 as confirmed by private a research [14] but the data was unavailable to use. Even though the findings reasonable and were backed up by the visualization, UpToDate information about the percentage of youth is required to really understand the influence of youth in London crime.

Table of word counts

Problem statement	239
State of the art	490
Properties of the data	462
Analysis: Approach	451
Analysis: Process	1493
Analysis: Results	187
Critical reflection	431

References

[1] Martin Bentham. O 2019, “London knife crime at record high with 15,023 offences in a year.” Evening Standard. [

[2] Metropolitan Police Services D 2019, “Recorded Crime: Geographic Breakdown.” London Datastore.

[3] Cheng, T & Williams, D 2012, “Space-time Analysis of Crime Patterns in Central London”, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 39, pp 47-52

[4] Corman, H & Mocan, H.N 2000, “A Time-Series Analysis of Crime, Deterrence, and Drug Abuse in New York City”, The American Economic Review, vol. 90, no. 3, pp. 584-604

[5] Nikolaos Kourentzes, “Addotove and multiplicative seasonality – can you identify them correctly”, Nov. 2014.

[6] Daniel Clark, “Violent crime in London 2019 Statistic”, Statista, 2020.

[7] “‘Poverty links’ to youth violence – London Mayer”, BBC News, July 2019

[8] “2011 Census – London Datastore”, Data.london.gov.uk, 2019

[9] “Home - Office for National Statistics”, Ons.gov.uk, 2019

[10] “Youth”, Wikipedia, 2020

[11] “Shapefile”, Wikipedia, 2020

[12] Alex Bomberg, “How safe is London?”, Intelligent-Protection, Aug. 2019.

[13] Joe Gamp, “Notting Hill Knife Carnival: One stabbed, 49 knives seized, 385 held – crime 33% up”, Express, Aug. 2018.

[14] “Young Peoples Capital of the World”, London Youth , 2019.