Big Data Analytics: London Crime Data Analysis

Gianmarco Ricciarelli[1]

[1]University of Pisa,
gianmarcoricciarelli@gmail.com ,
GitHub: https://github.com/germz01

Overview

1 Introduction

2 Data Understanding

3 Cluster Analysis

4 Forecasting

The analysis' purpose

To **discover** the clusters among the criminal activities in the London metropolitan area in a distinct window of time and to **forecast** a possible development for future crimes.

The Dataset(1)

London Crime Data, 2008-2016: this dataset, hosted by **Kaggle**, is composed by 13 millions rows describing the London metropolitan area's criminal activities by *Borough*, *Category*, *Month* and *Year* in a window of time that ranges from January 2008 to December 2016.

The Dataset(2)

The dataset is composed by 7 variables:

- **lsoa_code**: code for Lower Super Output Area in Greater London;
- **borough**: common name for London borough;
- **major_category**: high level categorization of crime;
- **minor_category**: low level categorization of crime within major category;
- **year**: year of reported counts, 2008 – 2016;
- **month**: month of reported counts, 1 – 12;
- **value**: monthly reported count of categorical crime in given borough;

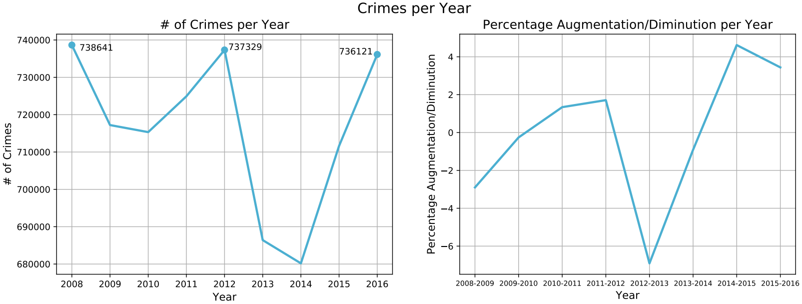
The Dataset(3)

The variables *lsoa_code*, *borough*, *major_category*, *minor_category*, *year* and *month* are **categorical** variables, while *value* is a **discrete numerical** variable.

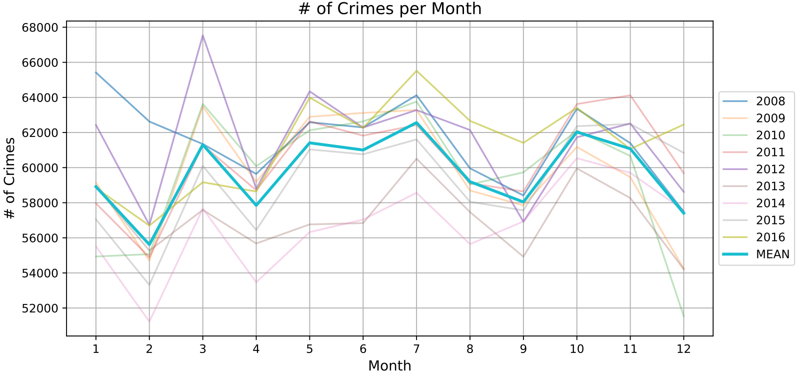
Numeric Variables' Analysis

value is the only numeric variables in the dataset, its **mode** is 0, which appears in the 74.56% of the dataset's samples (**10,071,505 records**). We can conclude that, on a superficial level, the window of time from 2008 to 2016 wasn't too dense of criminal activities.

Crimes per Year



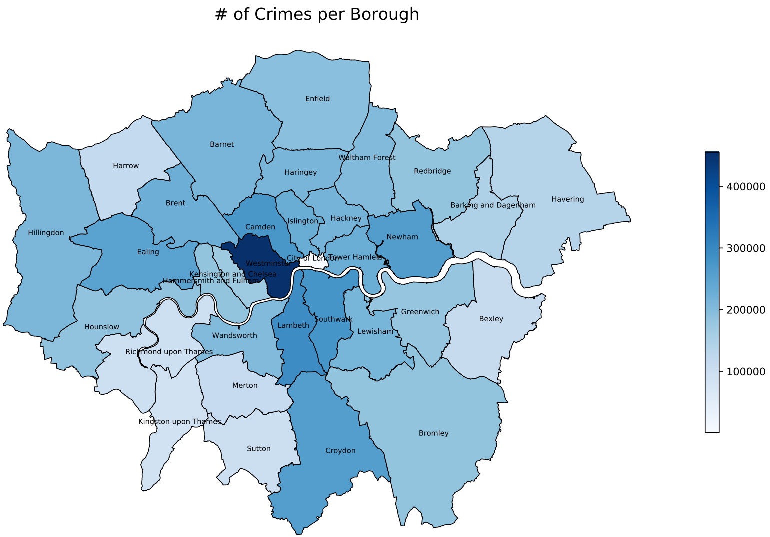
Crimes per Month



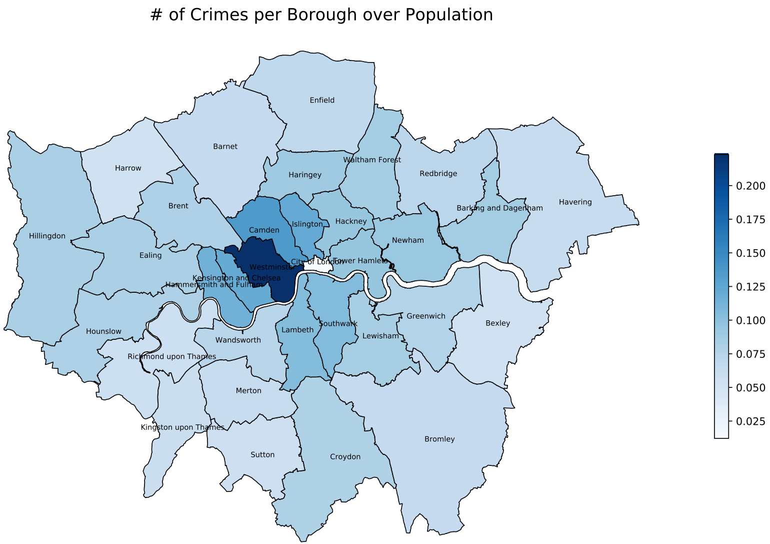
Categorical Variables' Analysis

- **borough** has 33 unique values, of which Lambeth is the most frequent, appearing in the 4.47% of the cropped dataset's records;
- **major_category** has 9 unique values, of which Theft and Handling is the most frequent, appearing in the 33.25% of the cropped dataset's records;
- **year** has 9 unique values, of which 2016 is the most frequent, appearing in the 11.45% of the cropped dataset's records;
- **month** has 12 unique values, of which 7 is the most frequent, appearing in the 8.66% of the cropped dataset's records;

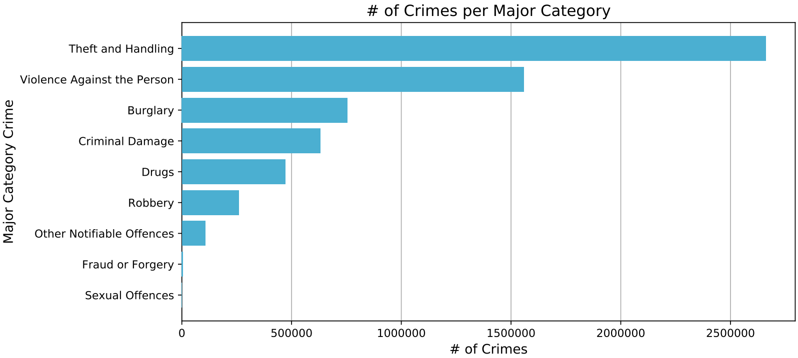
Crimes per Borough



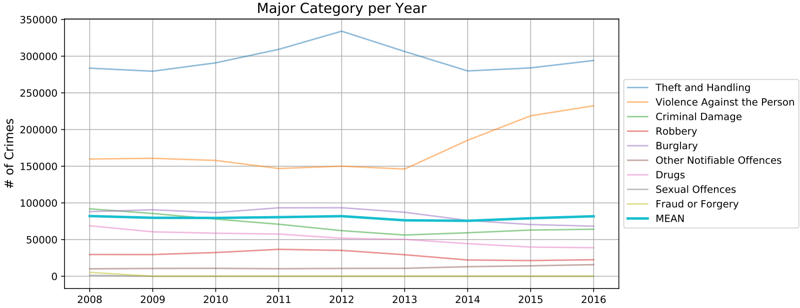
Crimes per Borough over Population Density



Crimes per Major Category



Major Category Crimes per Year



Correlation Analysis

	lsoa_code	borough	major_category	minor_category	value	year	month
lsoa_code	D	D	D	D	D	D	I
borough	D	D	D	D	D	D	D
major_category	D	D	D	D	D	D	D
minor_category	D	D	D	D	D	D	D
value	D	D	D	D	D	D	D
year	D	D	D	D	D	D	D
month	I	D	D	D	D	D	D

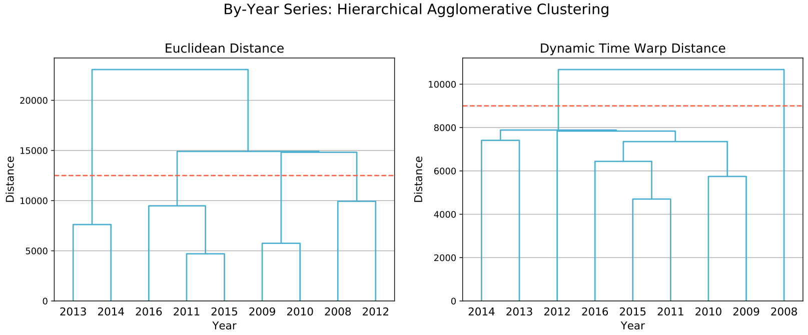
Introduction

I have decided to enrich the informations provided with the data understaning by searching for possible **cluster-like structures** in the **time series** extracted from the main dataset, that is, hoping to discover the similarities (or dissimilarities) among the series describing the criminal activities from 2008 to 2016.

Choice of the Algorithms

I have used three popular clustering algorithms, that is, the **KMeans algorithm**, the **Hierarchical Agglomerative algorithm** and the **DBSCAN**. The three algorithms were adapted depending on the different series they were applied on.

By-Year Series: Hierarchical Agglomerative Clustering



By-Year Series: KMeans Algorithm and DBSCAN Clustering(1)

	Cluster
Year	
2009	0
2010	0
2011	1
2015	1
2016	1
2008	2
2012	2
2013	3
2014	3

(a) KMeans

	Cluster
Year	
2008	-1
2012	-1
2009	0
2010	0
2011	1
2015	1
2016	1
2013	2
2014	2

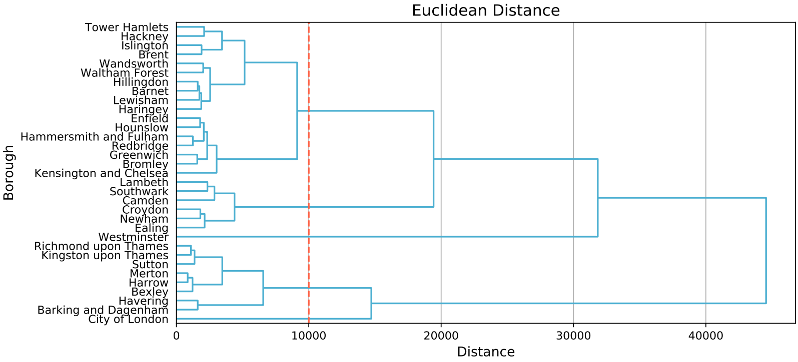
(b) DBSCAN

By-Year Series: KMeans Algorithm and DBSCAN Clustering(2)

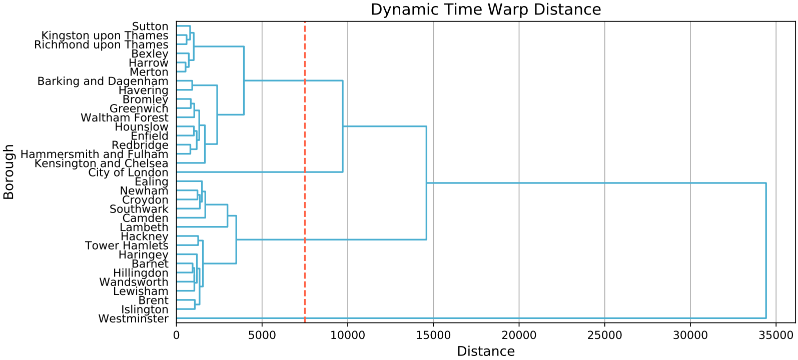
By-Year Series: Conclusions

- the series representing years 2013 and 2014 are the least dense of criminal activities, hence are clustered together;
- the series representing years 2008 and 2012 are the most dense of criminal activities, hence are clustered together;
- the remaining series are splitted into two distinct clusters;

By-Borough Series: Hierarchical Agglomerative Clustering(1)



By-Borough Series: Hierarchical Agglomerative Clustering(2)



By-Borough Series: KMeans Algorithm and DBSCAN Clustering

Borough	Cluster
Brent	0
Kensington and Chelsea	0
Barnet	0
Haringey	0
Tower Hamlets	0
Hillingdon	0
Lewisham	0
Waltham Forest	0
Hounslow	0
Islington	0
Wandsworth	0
Bromley	0
Hackney	0
Hammersmith and Fulham	0
Enfield	0
Redbridge	0
Greenwich	0

(a) KMeans 1

City of London	1
Westminster	2
Havering	3
Richmond upon Thames	3
Kingston upon Thames	3
Bexley	3
Harrow	3
Merton	3
Barking and Dagenham	3
Sutton	3
Camden	4
Lambeth	4
Southwark	4
Croydon	4
Newham	4
Ealing	4

(b) KMeans 2

Borough	Cluster
City of London	-1
Westminster	-1
Merton	0
Harrow	0
Richmond upon Thames	0
Havering	0
Kingston upon Thames	0
Sutton	0
Barking and Dagenham	0
Bexley	0
Hounslow	1
Enfield	1
Hammersmith and Fulham	1
Wandsworth	1
Ealing	1
Islington	1
Redbridge	1
Lewisham	1

(c) DBSCAN 1

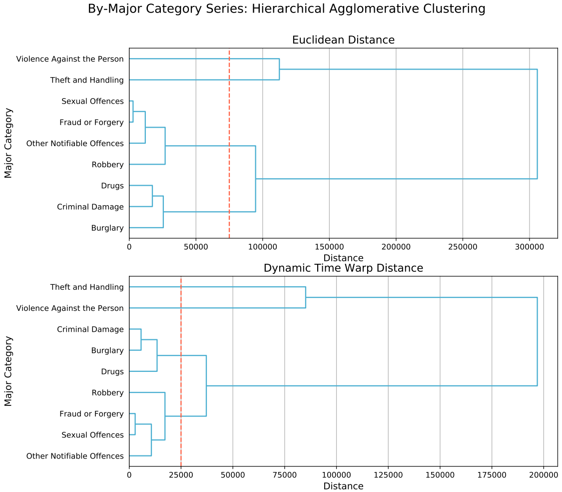
Hillingdon	1
Hackney	1
Tower Hamlets	1
Camden	1
Bromley	1
Greenwich	1
Waltham Forest	1
Newham	1
Barnet	1
Haringey	1
Lambeth	1
Southwark	1
Croydon	1
Kensington and Chelsea	1
Brent	1

(d) DBSCAN 2

By-Borough Series: Conclusions

- The series representing Westminster and City of London are the two extremes among the boroughs, hence they are clustered by the themself in dinstinct clusters;

By-Major Category Series: Hierarchical Agglomerative Clustering



By-Major Category Series: KMeans Algorithm and DBSCAN Clustering(1)

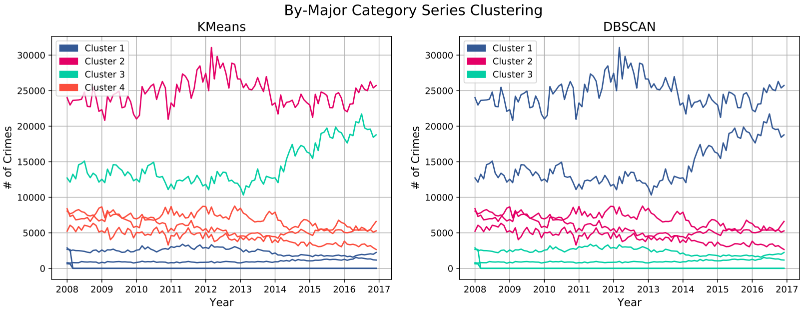
Major Category Crime	Cluster
Robbery	0
Fraud or Forgery	0
Sexual Offences	0
Other Notifiable Offences	0
Theft and Handling	1
Violence Against the Person	2
Criminal Damage	3
Burglary	3
Drugs	3

(a) KMeans

Major Category	Cluster
Theft and Handling	-1
Violence Against the Person	-1
Criminal Damage	0
Burglary	0
Drugs	0
Robbery	1
Fraud or Forgery	1
Sexual Offences	1
Other Notifiable Offences	1

(b) DBSCAN

By-Major Category Series: KMeans Algorithm and DBSCAN Clustering(2)



By-Major Category Series: Conclusions

- Fraud or Forgery, Sexual Offences, Other Notifiable Offences and Robbery are the least popular types of crimes, hence their series are clustered together;
- Theft and Handling and Violence Against the Person are the most popular types of crimes, hence they form distinct clusters for themeself;
- the other categories are clustered together;

The Models - ARIMA Family Models

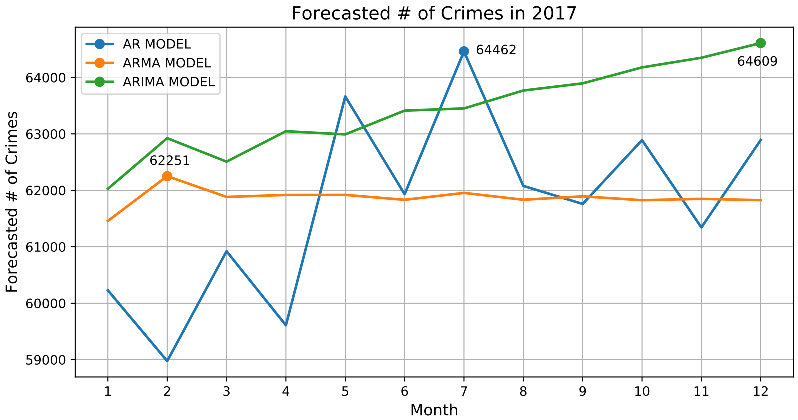
Represent by the formulas:

AR Model: $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$

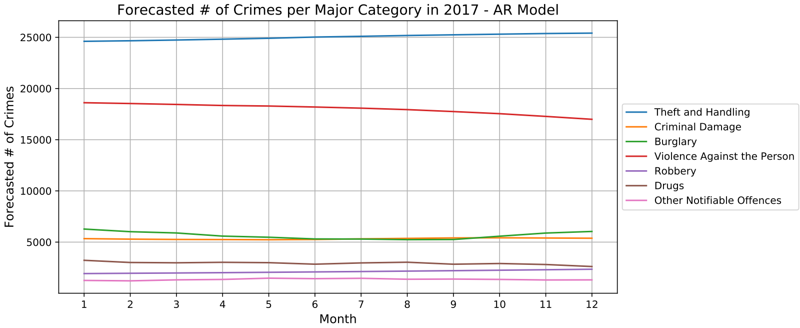
ARMA Model: $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$

ARIMA Model $\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t$

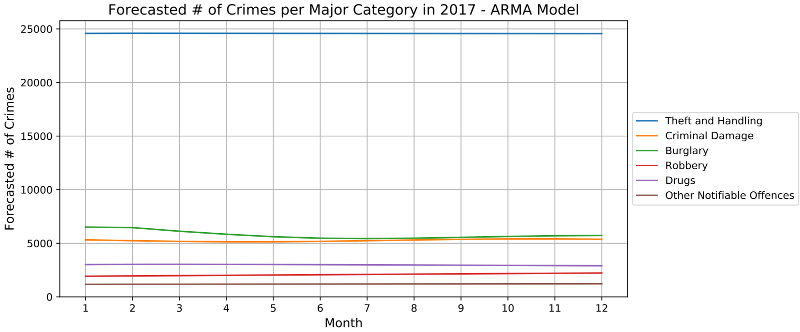
Forecasting of # of Crimes over the city



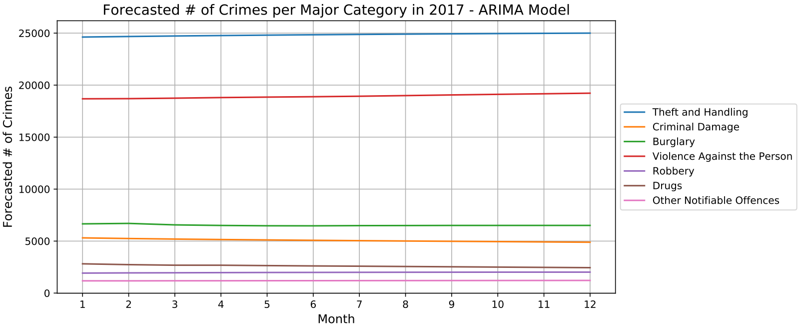
Forecasting of # of Crimes per Major Category - AR model



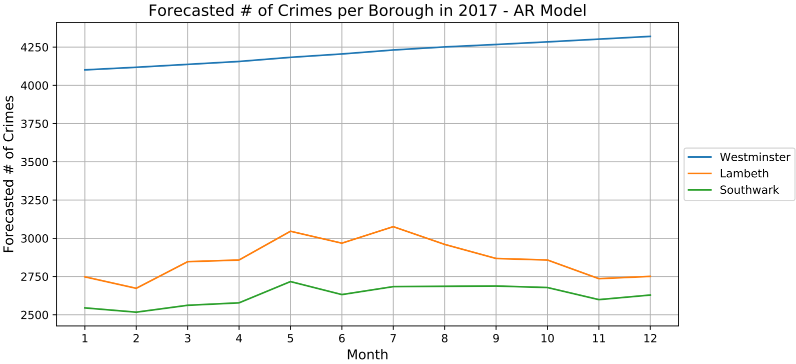
Forecasting of # of Crimes per Major Category - ARMA model



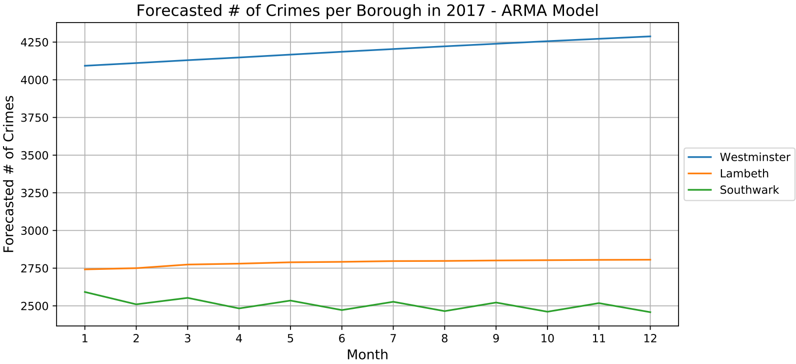
Forecasting of # of Crimes per Major Category - ARIMA model



Forecasting of # of Crimes per Borough - AR model



Forecasting of # of Crimes per Borough - ARMA model



Conclusions

In general, the **Autoregressive model** seems to perform better than the other models in forecasting the progression of criminal activities.