

LONDON CRIME

Maddalena Amendola

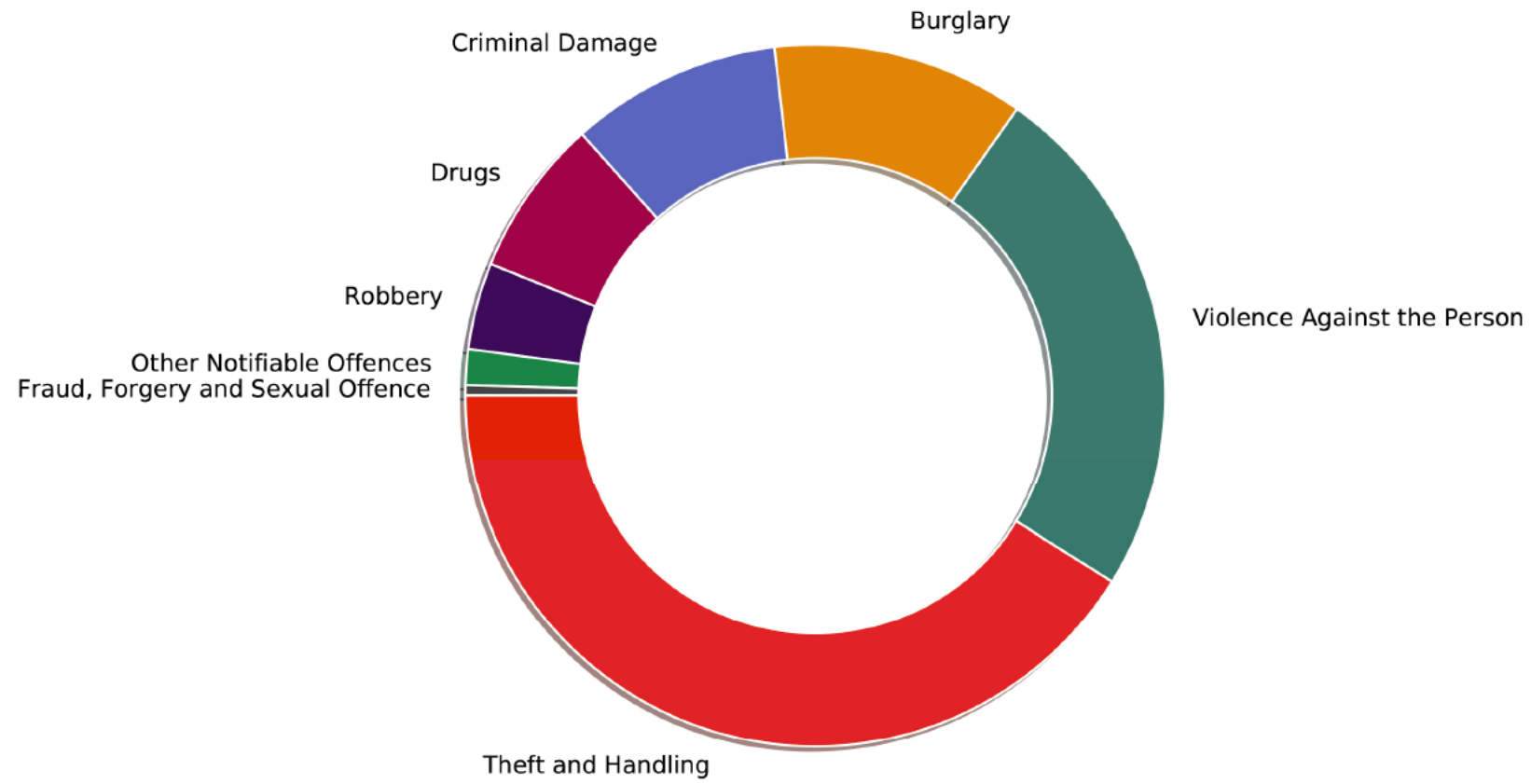
Giuliano Cornacchia

Mario Leonardo Salinas

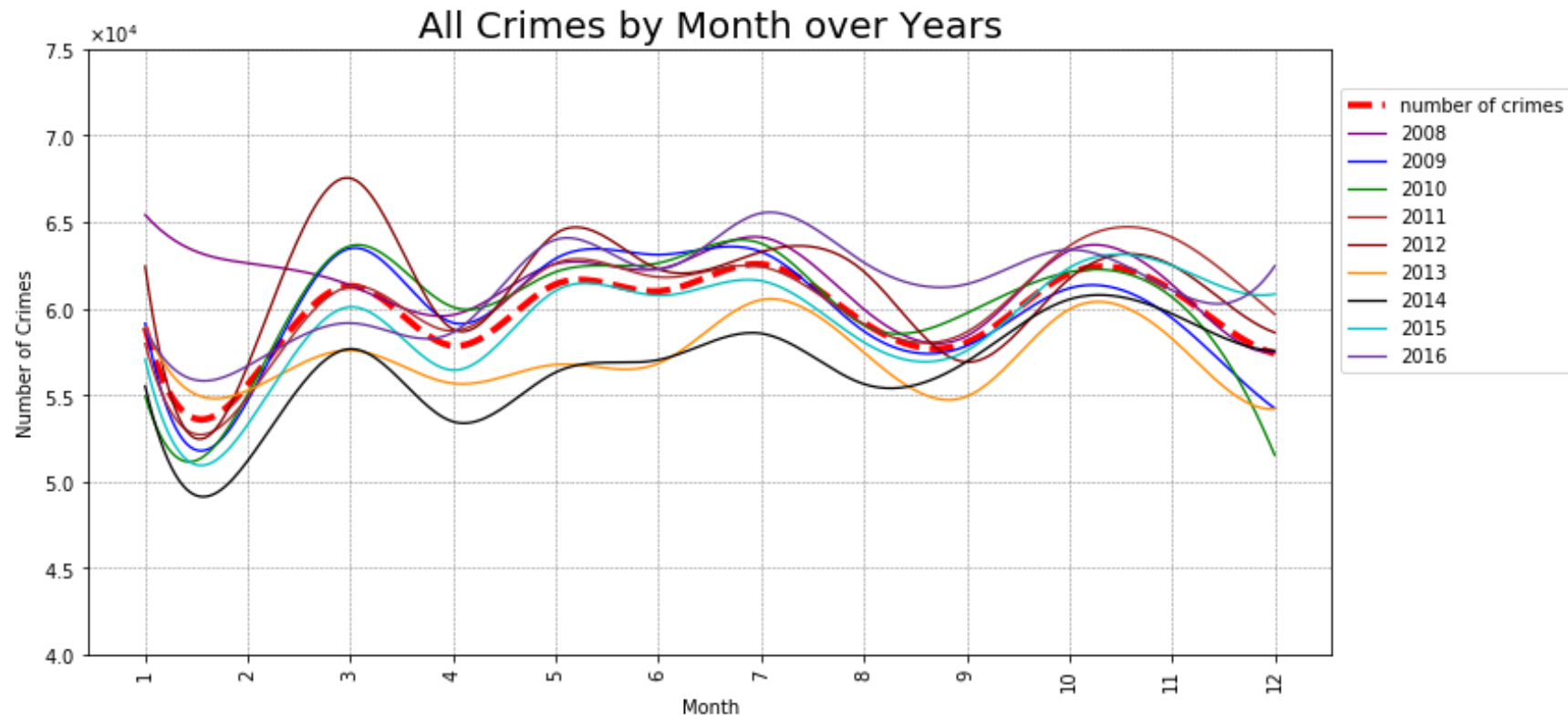
Pisa, 14/12/2018

University of Pisa

Crimes Distribution

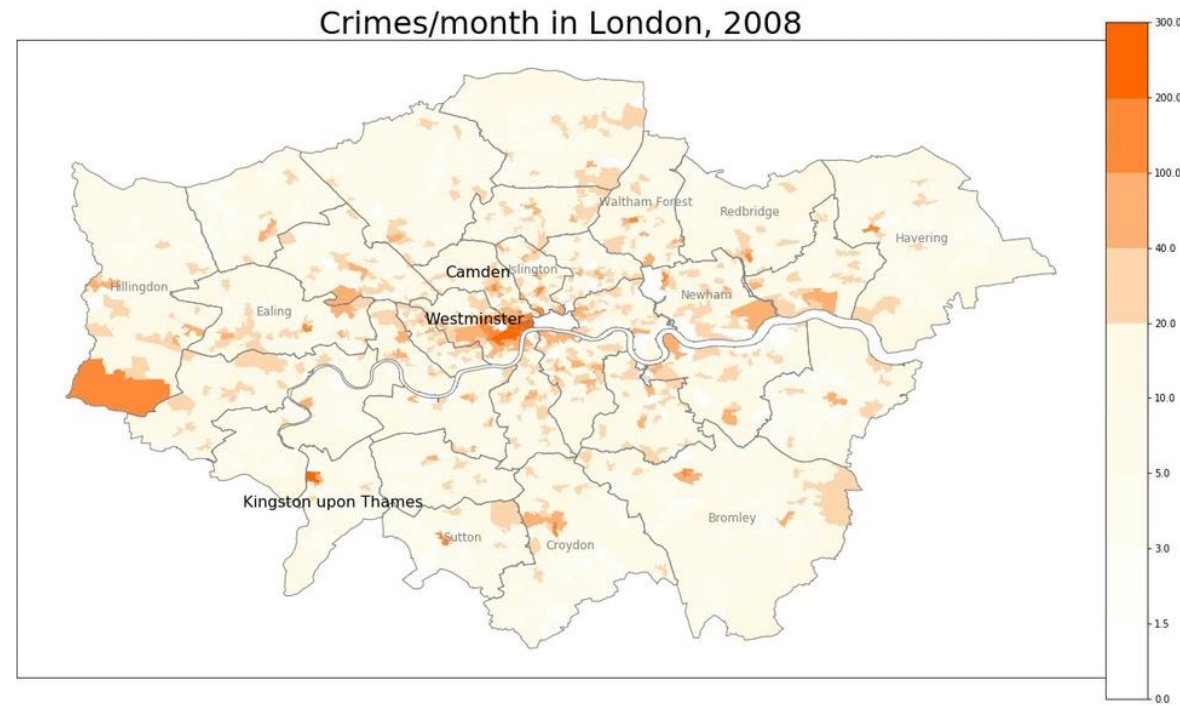


Crime Trend over years



Distribution of crimes

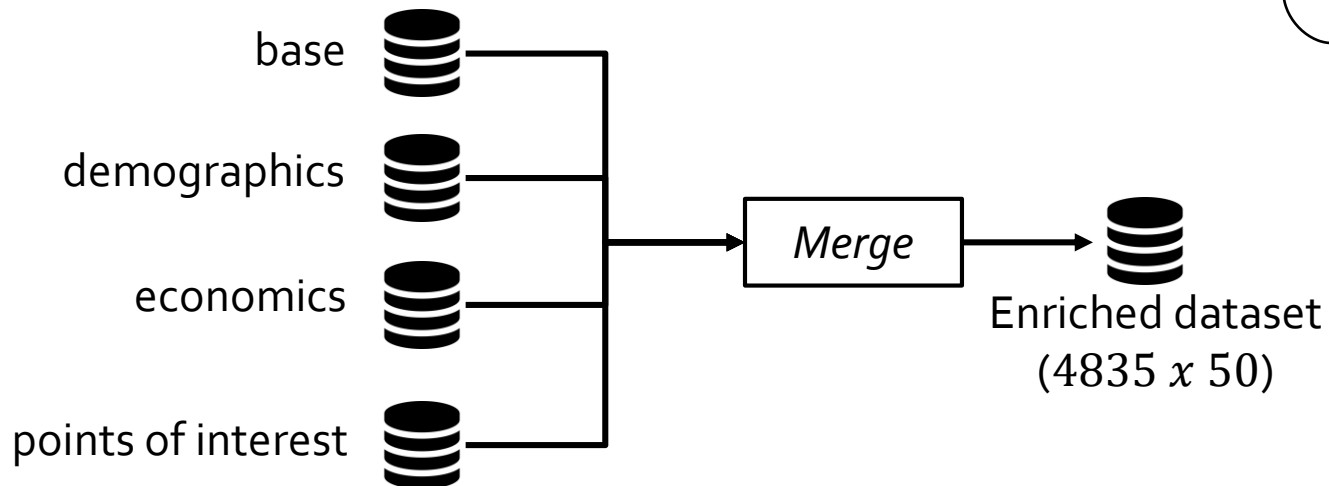
- Criminality concentrated in some areas and constant in time



Enriched DataSet

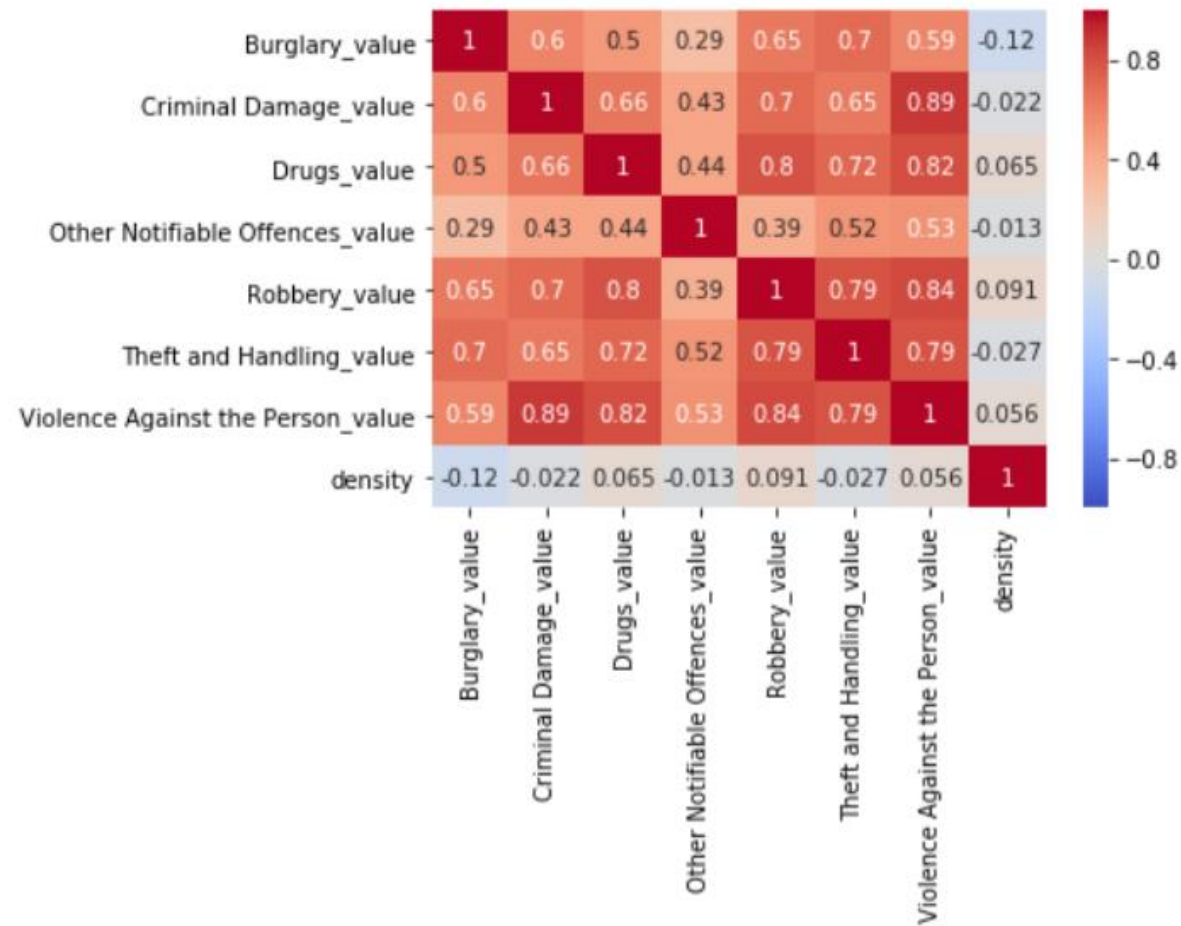
We create a new dataset for each year with one row for **LSOA** and a new attribute for major Crime.

	Isoa_code	borough	value	Burglary_value	Criminal Damage_value	Drugs_value	Robbery_value	Sexual Offences_value	Theft and Handling_value	Violence Against the Person_value	Fraud or Forgery_value	Other Notifiable Offences_value
0	E01000006	Barking and Dagenham	105	12	7	8	6	1	37	34	0	0



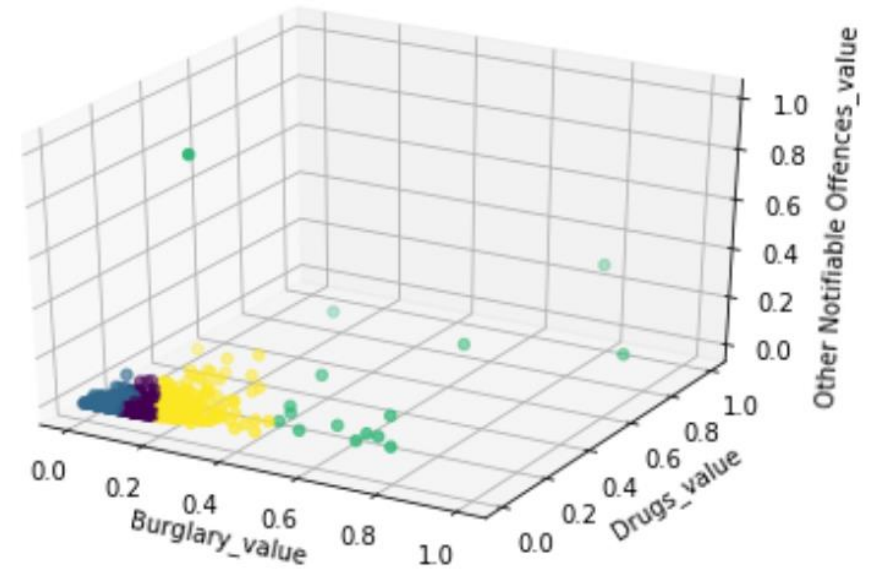
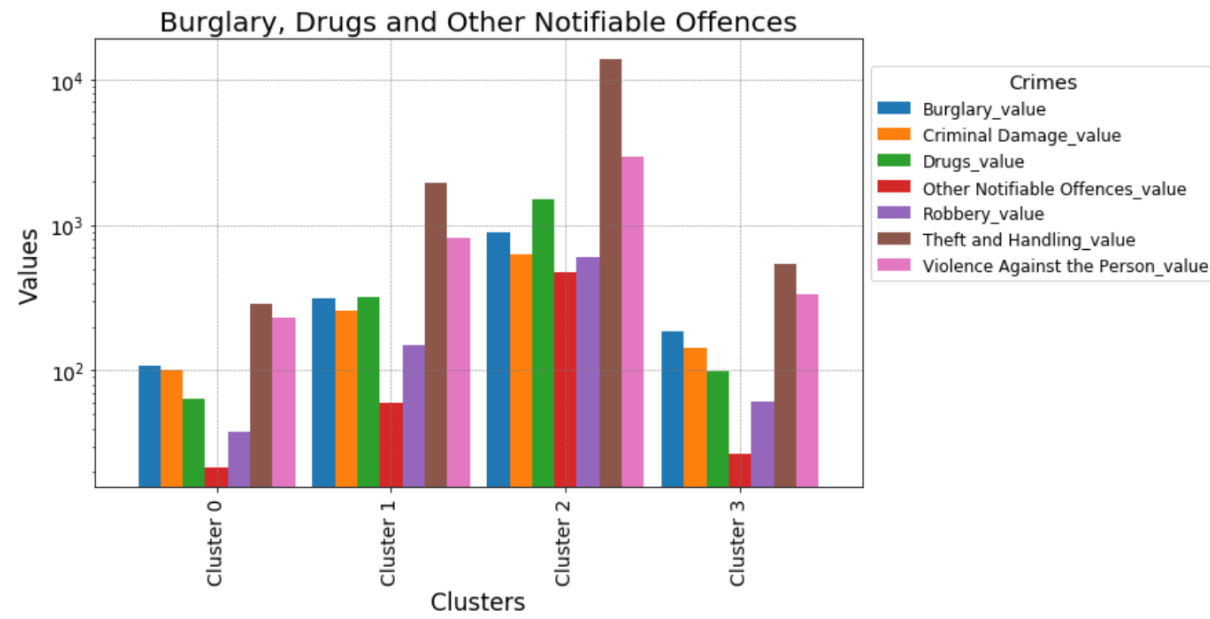
Sum of all «Theft and Handling» crimes occurred in the row-relative **LSOA**

Correlation Matrix



Clustering

- Lsoa clustered according to their dangerousness



Regression

- Time series data
- Regression at Borough level
- Arima:
 - Better than dummy regressors
 - No good results

	MAE	MAPE (%)
ARIMA (1,0,0)	53.86	2.4
Dummy AVG	87	4.1
Dummy Last	148	6.3
Dummy Lasts	234	11.9



New Approaches

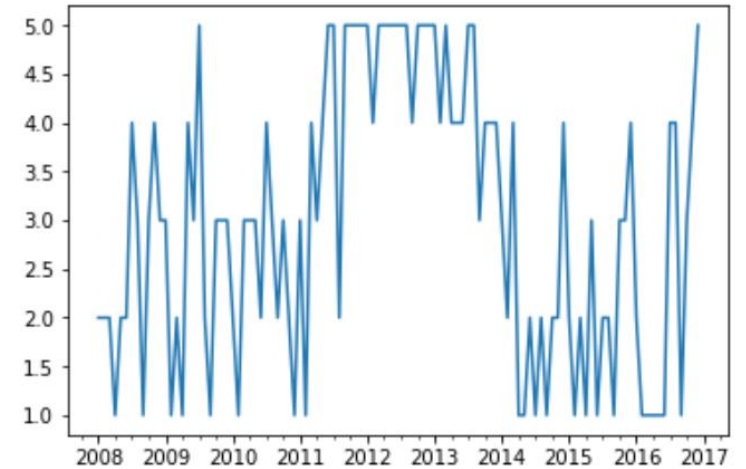
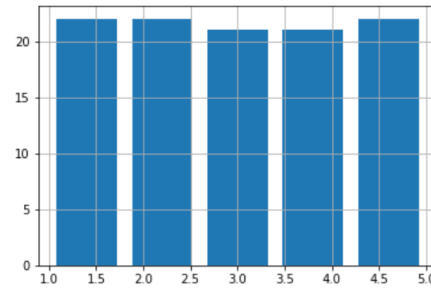
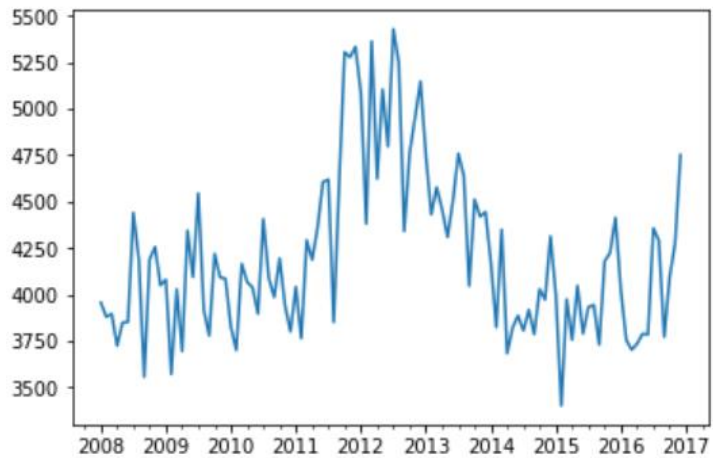
Classification:

- **Idea:** predict the class of danger of a Borough
- **Procedure:** discretization of the value of crimes in 5 classes that represent the dangerousness of a Borough and multi-classification approach

Regression:

- **Idea:** predict the number of crimes at LSOA level
- **Procedure:** construction of a model on a dangerous LSOA using the clustering information
- **Hypothesis:** the model can be used also to predict the number of crimes of LSOAs belonging in the same cluster

Binning - Westminster



Class	Crimes/day
Class 1	120.42
Class 2	130.86
Class 3	137.67
Class 4	146.05
Class 5	166.00

Random Forest Classifier

Parameter	Value
prev_obs	2
n_estimators	16
bootstrap	<i>False</i>
max_depth	<i>None</i>
min_samples_leaf	5

Time-series **CrossValidation**

avg(MAE)= 1.132

std dev= 0.564

prev_obs is a parameter that controls the learning granularity

[1, 2, 3, 4, 5, 6] with *prev_obs* = 3:

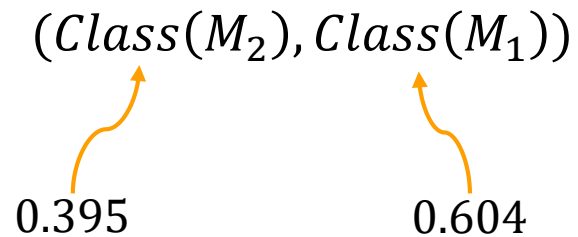


Sample	Target
[1,2,3]	[4]
[2,3,4]	[5]
[3,4,5]	[6]

Random Forest - Interpretation

Feature importance

$(Class(M_2), Class(M_1))$



0.395 0.604

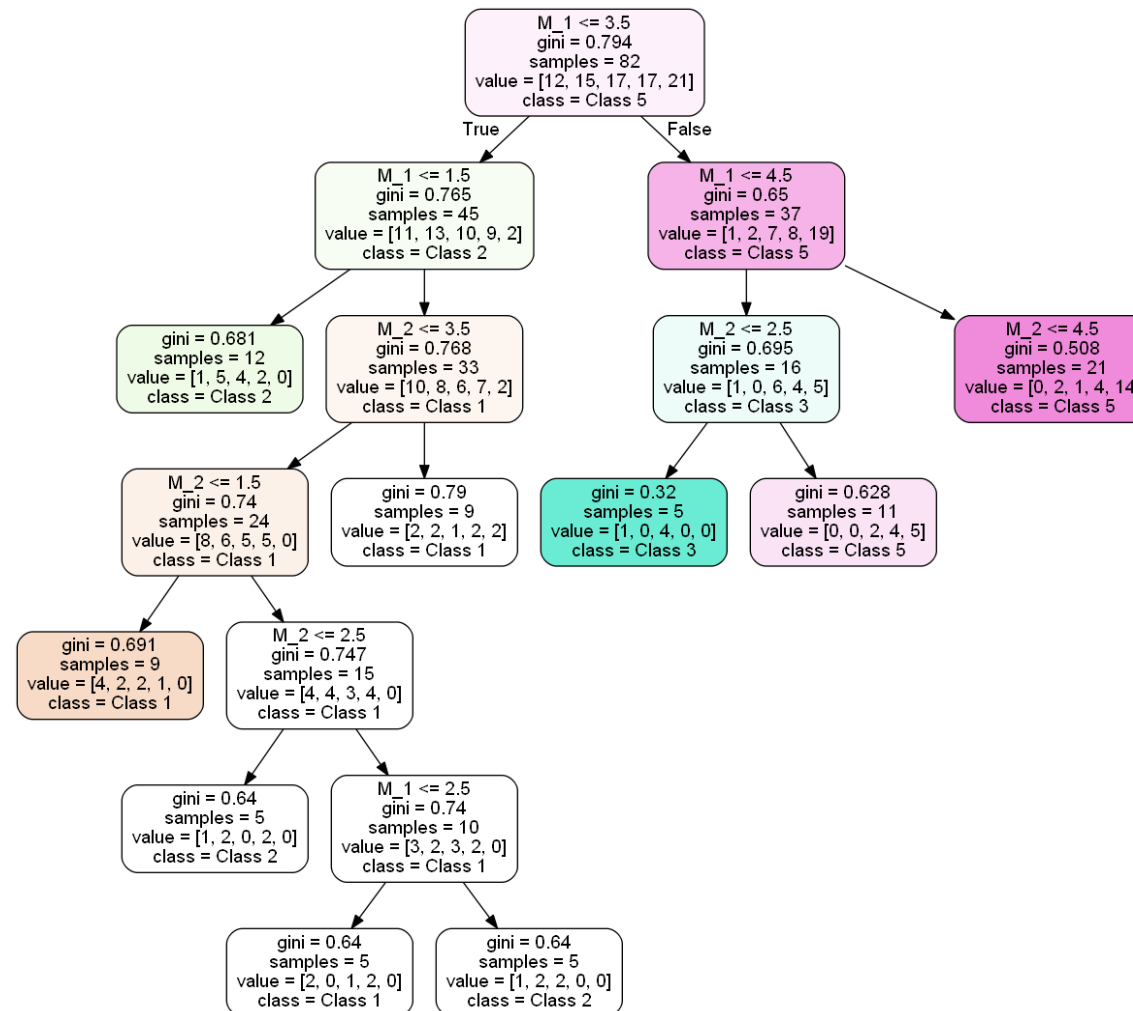
The generic **input** is in the form:

$(Class(M_{prev_obs}), \dots, Class(M_1))$

where:

- M_i is the month at distance i from the month to predict.
- $Class(M_x)$ returns the danger class of the month at distance x .

Random Forest - Interpretation



Gradient Boosting Classifier

Parameters	Value
prev_obs	2
n_estimators	10
learning_rate	0.17
max_depth	4
min_samples_leaf	2

Feature Importance

$(Class(M_2), Class(M_1))$

0.466 0.533

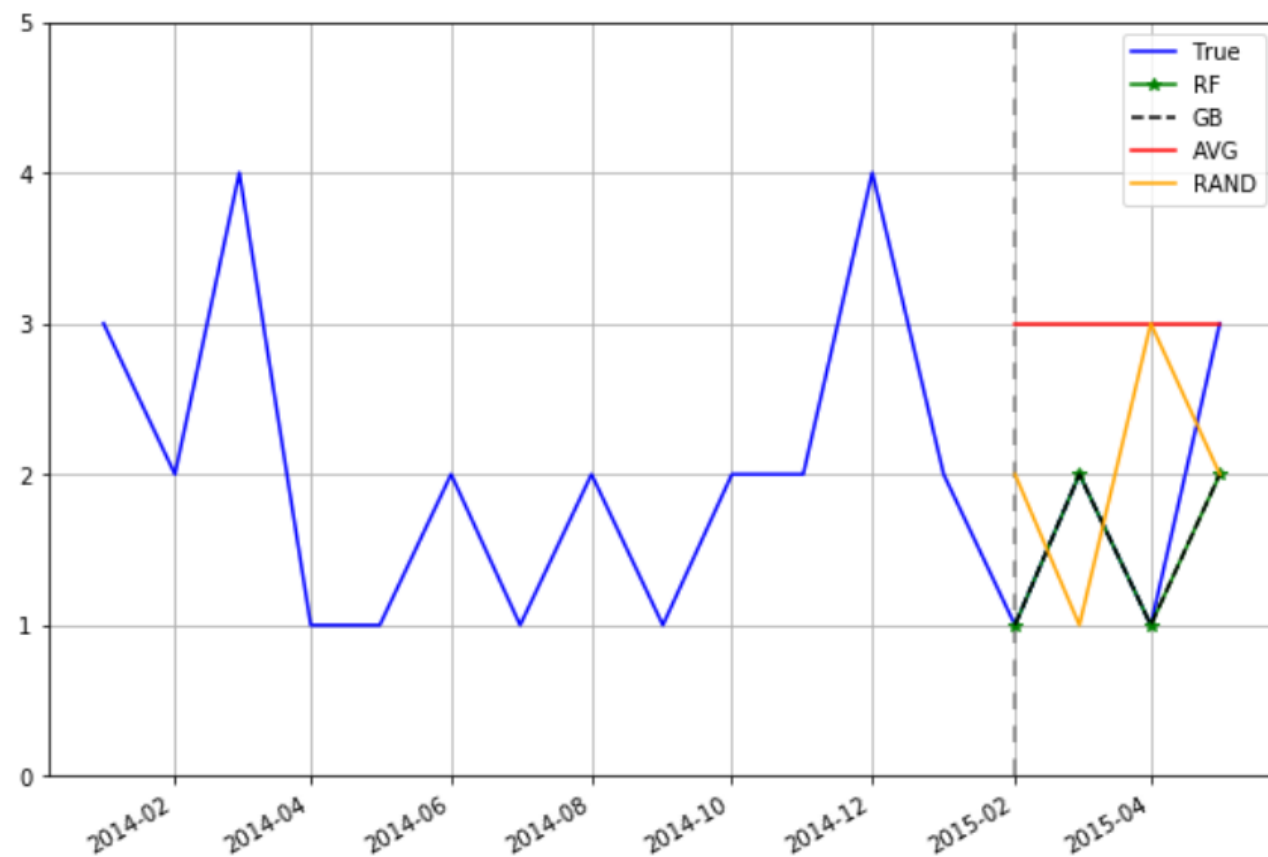
Time-series **CrossValidation**

avg(MAE)= 0.926

std dev= 0.384

Results comparison

Model	avg(MAE)	dev_std	predicted(1,2,1,3)	MAE(test)
Dummy AVG	1.373	0.416	(3, 3, 3, 3)	1.25
Dummy RAND	1.808	0.428	(2, 1, 3, 2)	1.25
Random Forest	1.088	0.487	(1,2,1,2)	0.25
Gradient Boost	0.926	0.384	(1,2,1,2)	0.25



Regression Results

Time Series
Cross Validation
scores



Model	EV	MAE	R2	Corr
Random Forest	0.13	0.113	-0.255	0.47
Extra Trees	0.16	0.119	-0.338	0.4
Gradient Boost	0.17	0.134	-0.47	0.46
Dummy_avg	0	0.14	-0.87	

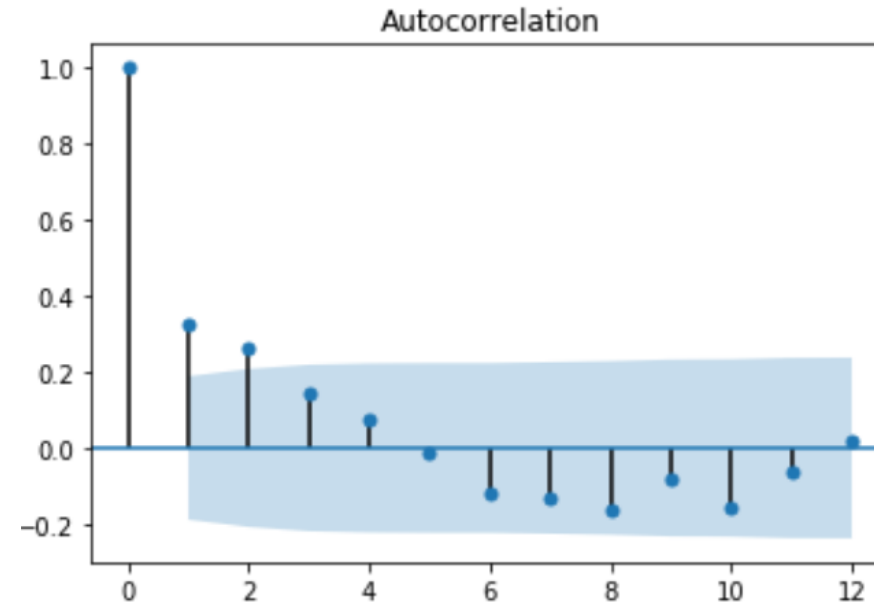
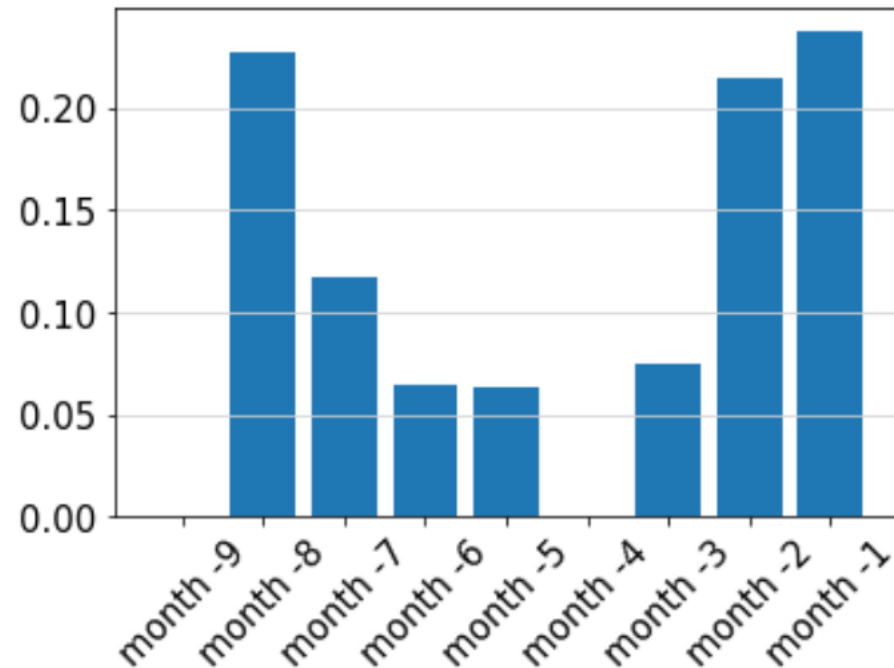
Best Regressor:
Random Forest



Parameter	value
prev_obs	9
max_depth	None
n_estimators	5

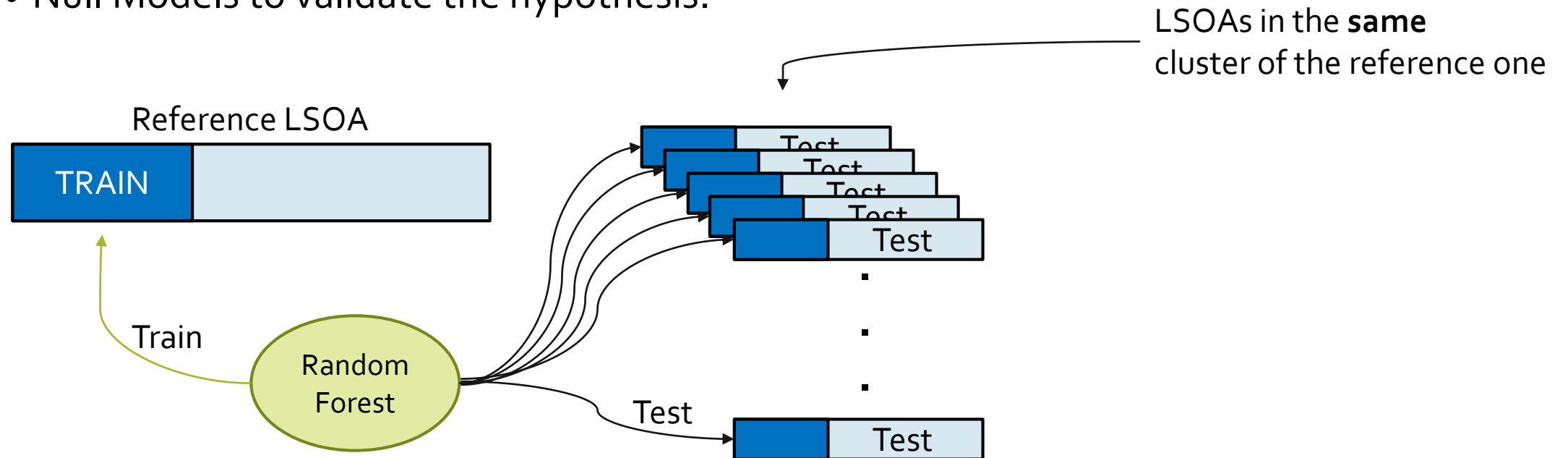
Features Importance

- The most important features are the ones that refer at the previous 2 months and the one collected 8 months before

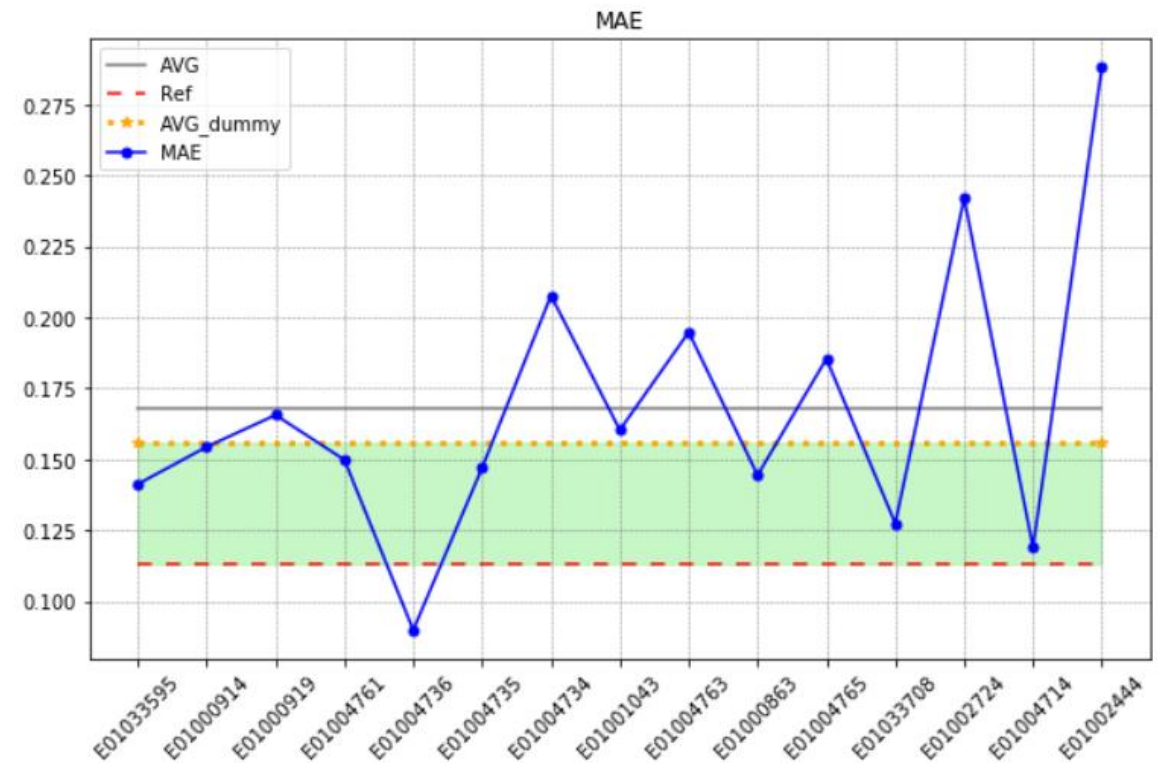
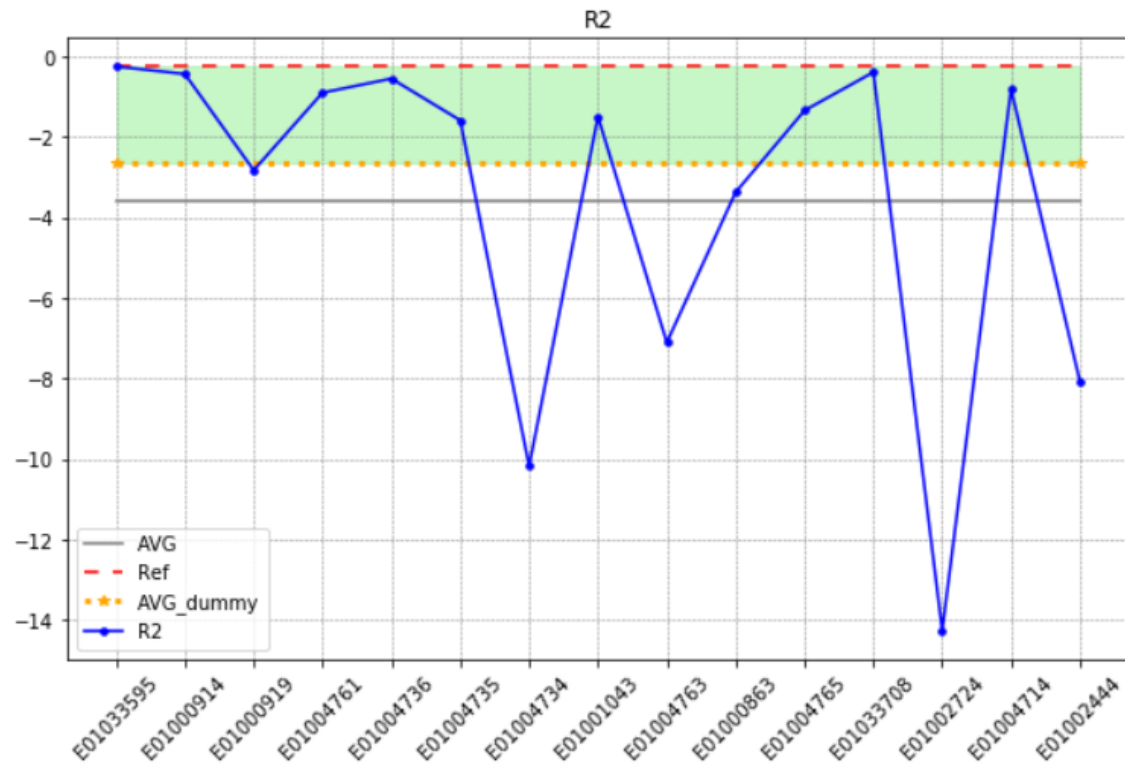


Assumption Evaluation

- Time series CV over a Reference LSOA and the elements of the same cluster
- Null Models to validate the hypothesis.

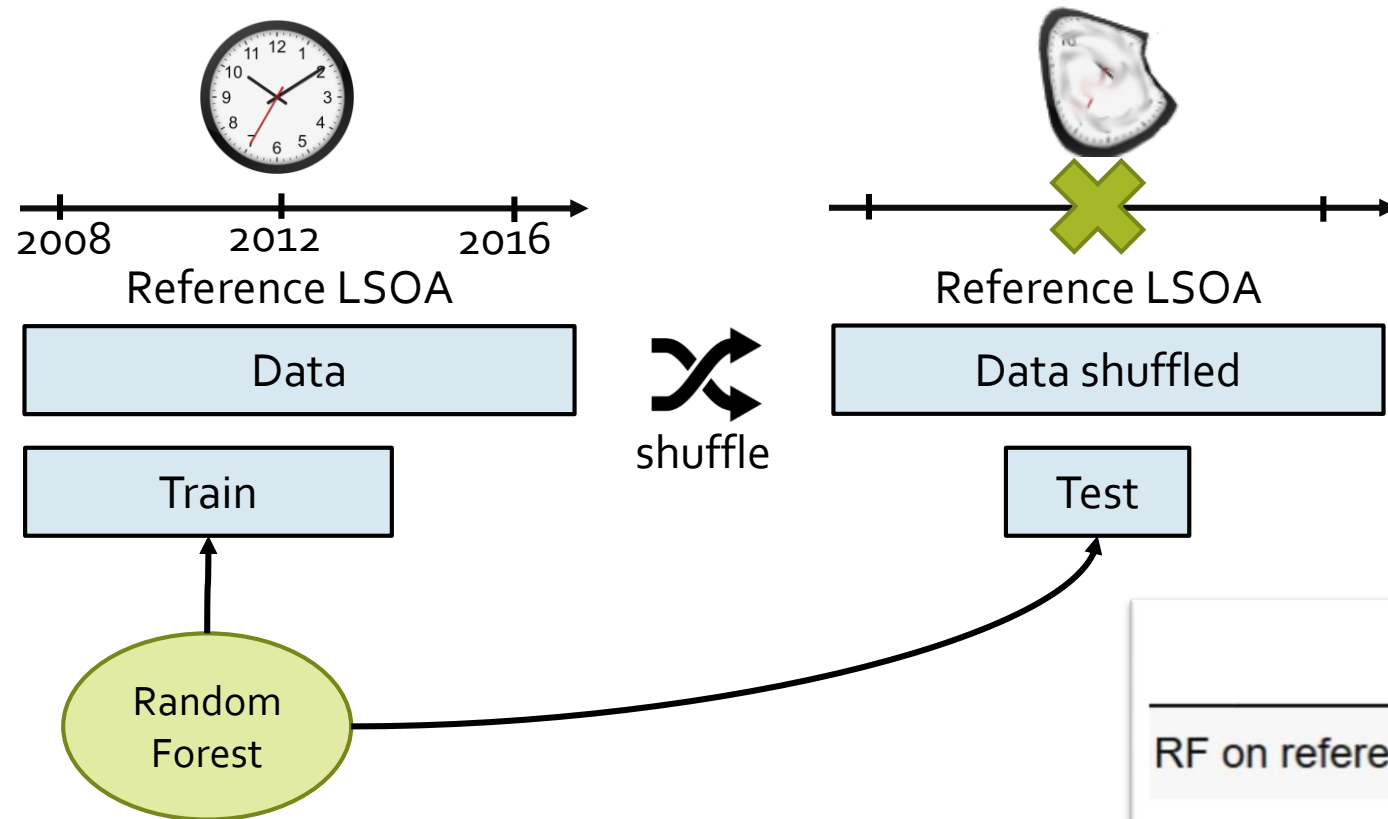


Assumption Evaluation pt1



Assumption Evaluation- Null Model 1

Data Randomization

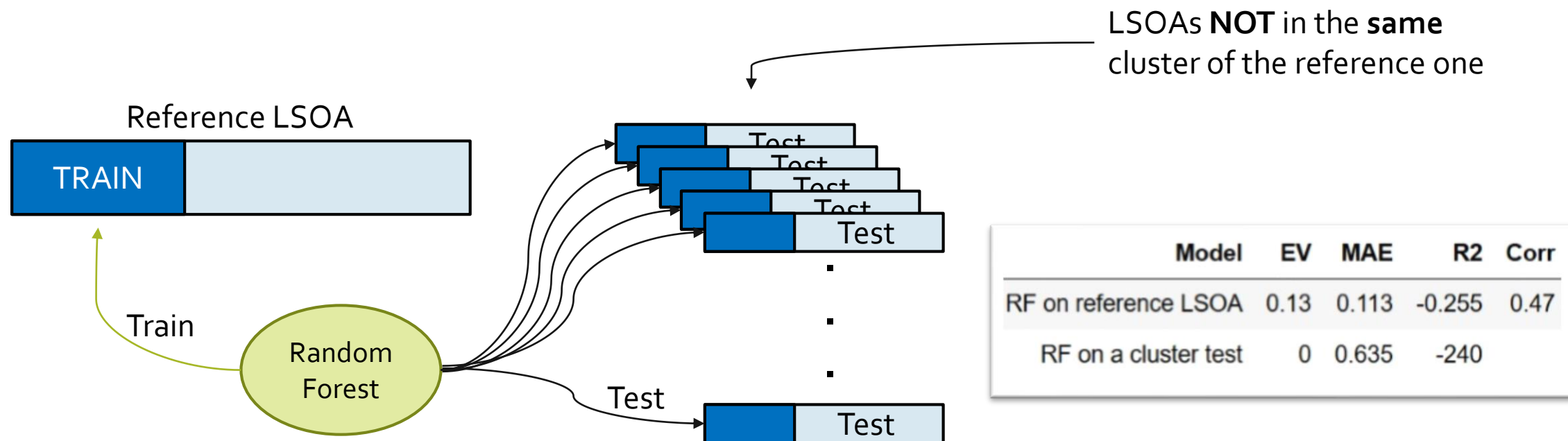


Model	EV	MAE	R2	Corr
RF on reference LSOA	0.13	0.113	-0.255	0.47
RF on shuffled data	-0.66	0.14	-1.004	-0.25

Assumption Evaluation- Null Model 2

Different Cluster Model

Different Cluster Model



	Model	EV	MAE	R2	Corr
RF on reference LSOA		0.13	0.113	-0.255	0.47
RF on a cluster test		0	0.635	-240	

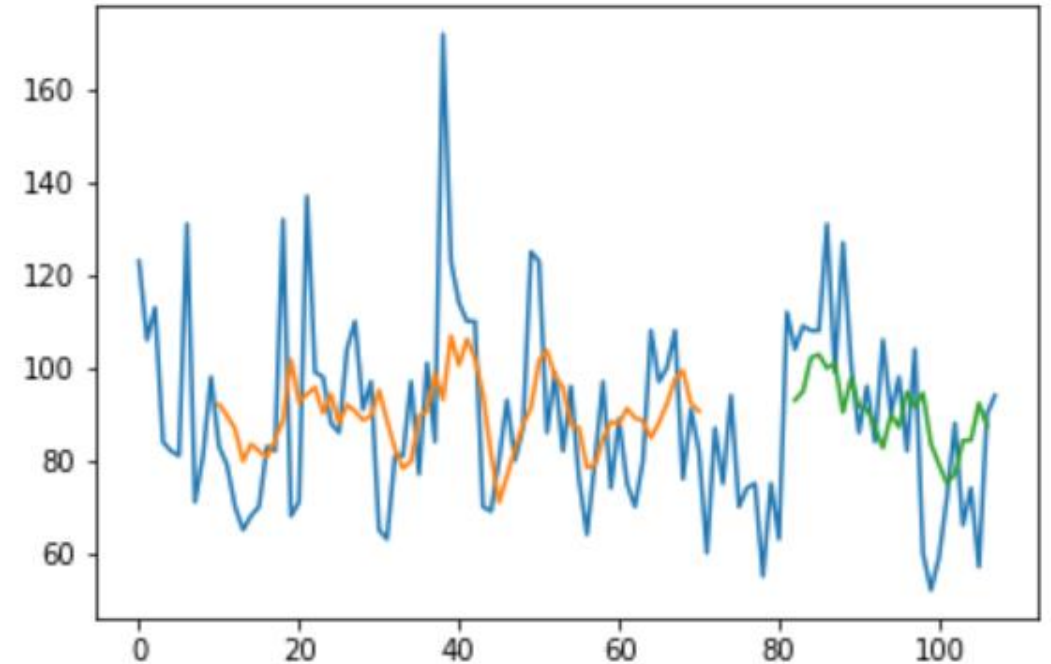
Long Short Term Memory

Prev_obs comparison:

	2	3	4	5	6	7	8	9	10	11	12
EV	0.226	0.218	0.242	0.21	0.228	0.143	0.168	0.177	0.23	0.201	0.136
MAE	0.127	0.129	0.127	0.13	0.128	0.126	0.126	0.123	0.119	0.122	0.127
R2	0.209	0.204	0.227	0.2	0.226	0.139	0.167	0.174	0.229	0.201	0.115
CORR	0.609	0.55	0.563	0.512	0.529	0.379	0.412	0.425	0.518	0.47	0.373

Model comparison:

	EV	MAE	R2	CORR
Random Forest	0.134	0.113	-0.225	0.475
LSTM	0.23	0.119	0.229	0.518

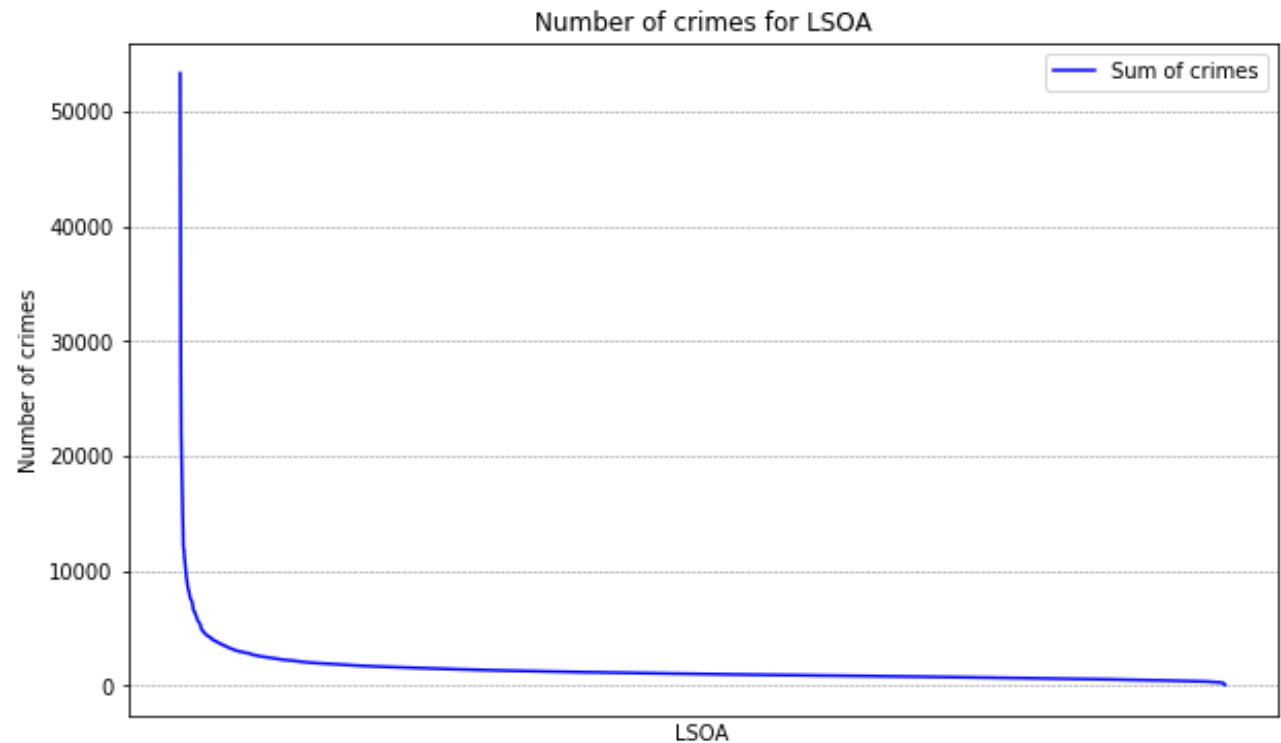


Conclusions

- Peak in 2012 (XXX Olympic Games), such big events can change considerably the number of crimes in London.
- London maintains a certain seasonality during the year (e.g. february the safest and march the most dangerous).
- Crimes distribution remains constant over the years.
- Violence is increasing (x2 in the last 3 years).
- Drug Crimes linearly decreasing.

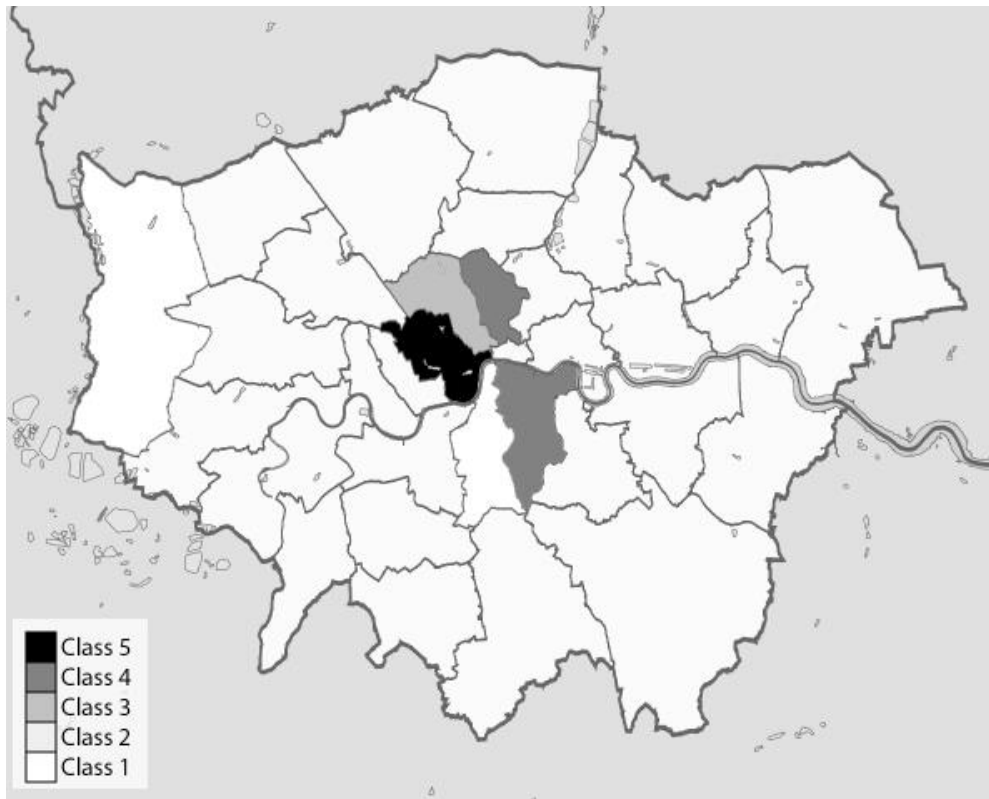
Conclusions

- The installation of CCTV cameras (500 000 in London) represent a good deterrent for criminals.
- The theft crime occurs more frequently in the touristic center (correlation with the number of monuments).
- Crime Locality.



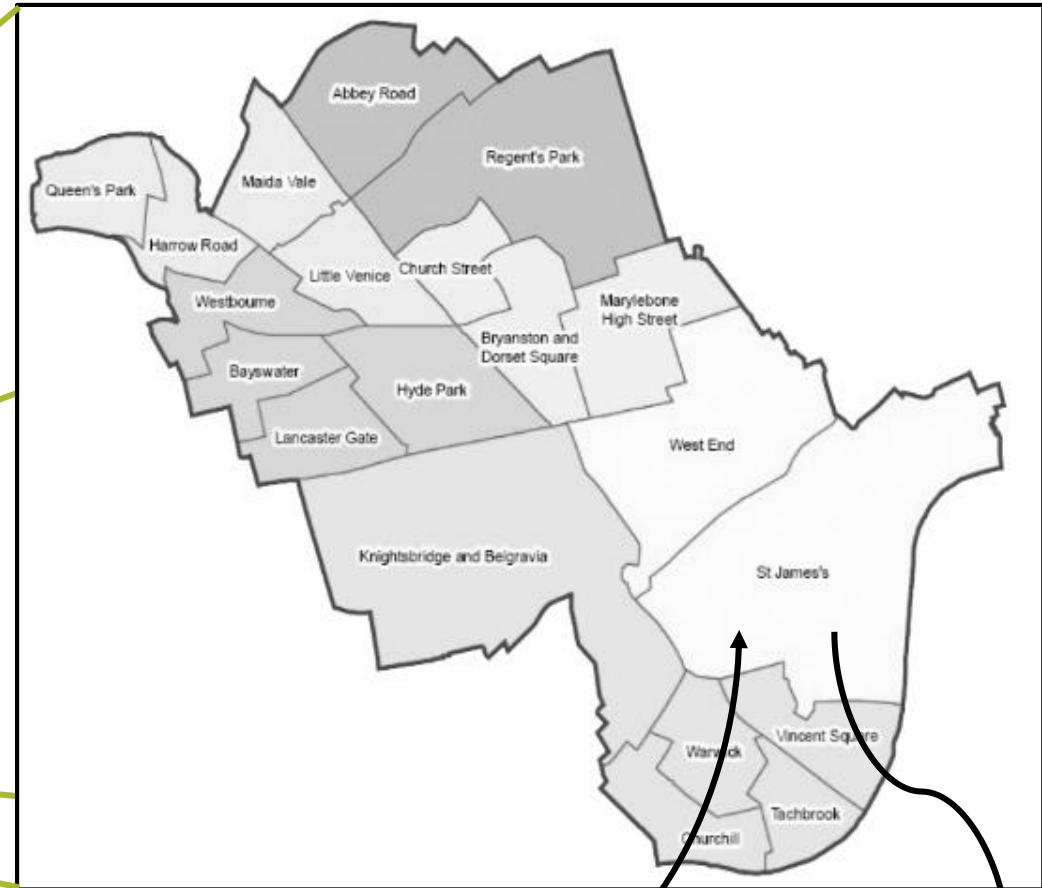
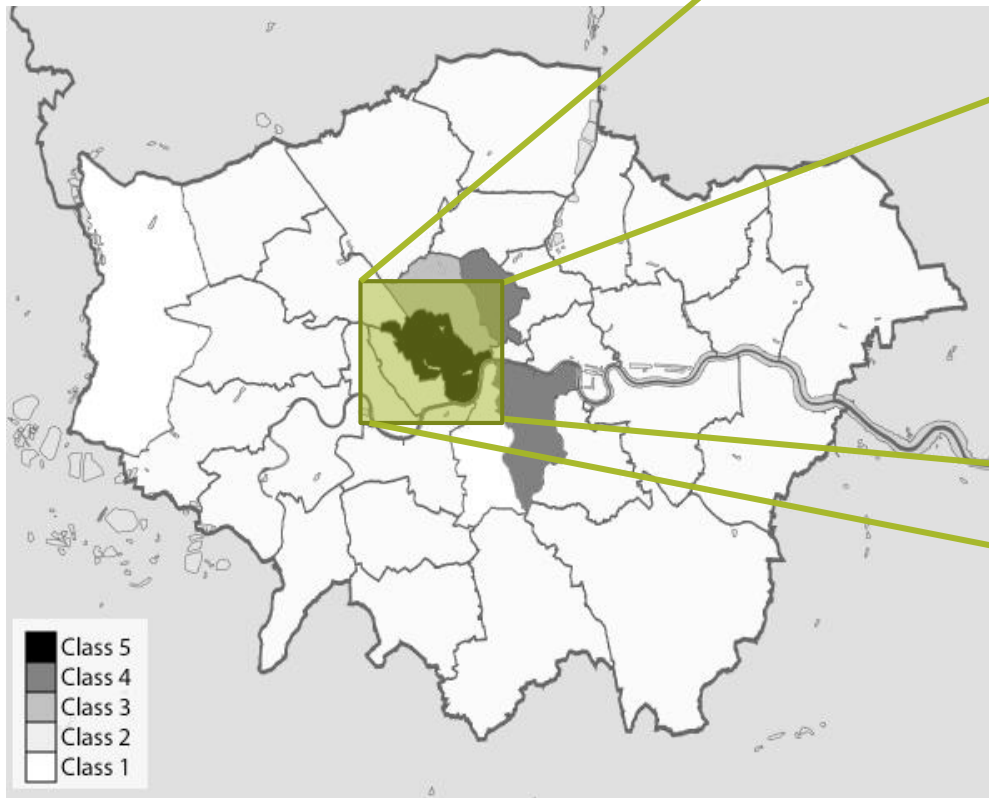
Use Scenario

RANGE MODEL



Use Scenario

RANGE MODEL



VALUE MODEL

140 crimes