Boston Crime Data Analysis

Monthly Prediction with Autoregressive Integrated Moving Average and Gaussian Process Regression

Benjamin Osafo Agyare , Ilaria Vinci, Francesco Zuniga

Abstract

We compare predictions from an Autoregressive (**AR**) Model and a Gaussian Process Regression (**GPR**) Model based on Exploratory Data Analysis (**EDA**) to establish which approach gives the best result in terms of minimum error. The data set analyzed is about the daily crimes in Boston, it contains records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and where it occurred. Records in the new system begin in June of 2015 to September 2018: 319,073 observations are recorded of 17 variables.

Background

Recent studies on this topic focused on the comparison of different model to make prediction for time series data based on visualization. These methods are Autoregressive Integrated Moving Average (**ARIMA**) and Seasonal Autoregressive Integrated Moving Average (**SARIMA**).

The Data and Design

Table1: Total Crime per Year

Year	Tot. Crime
2017	93006
2016	91459
2018	60712
2015	51115

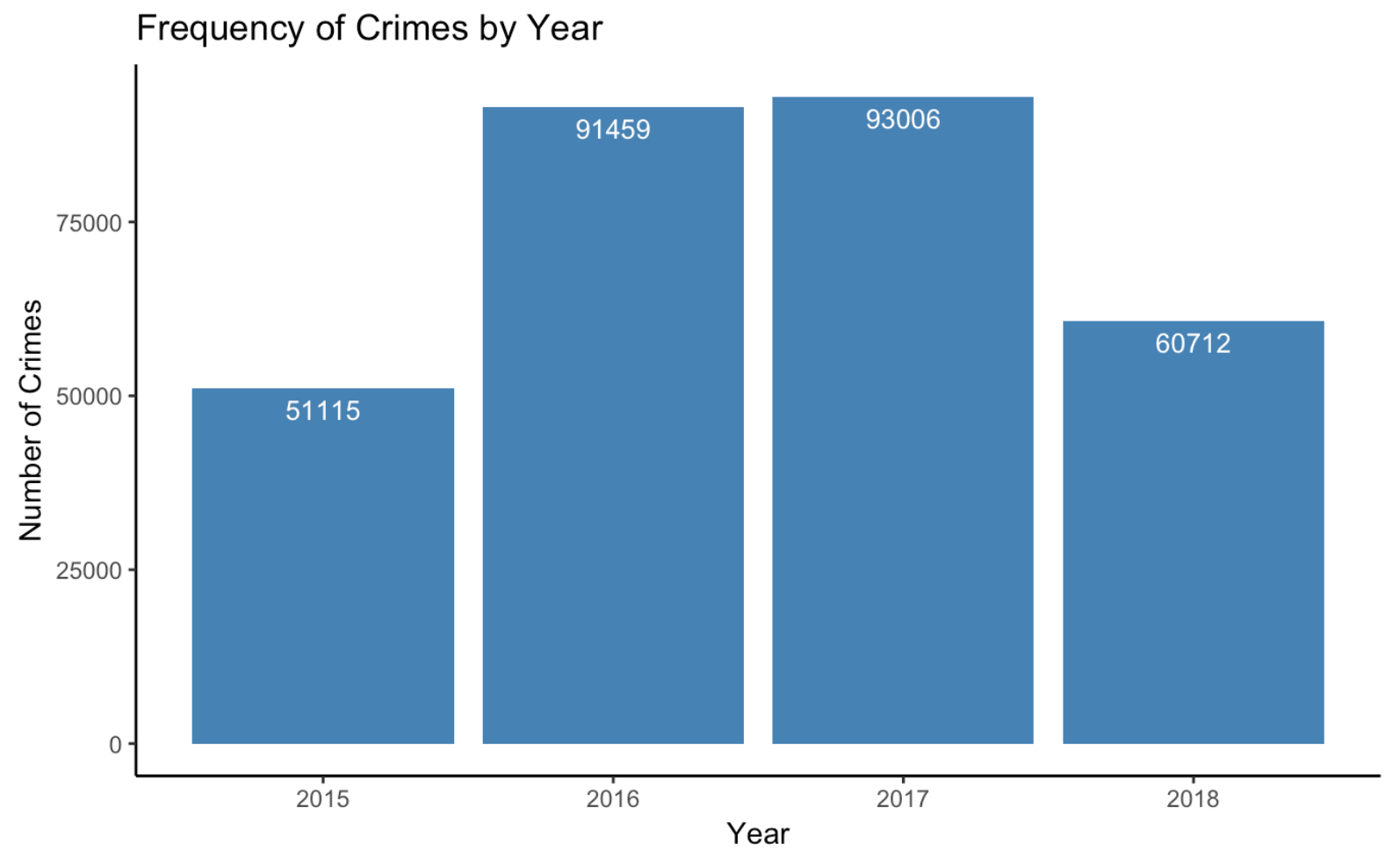

Fig. 1: Histogram for total number of crimes per year

The project aims to analyze a data set with the Boston crime rates, using techniques from **EDA**. Observing the datasets, out of the 17 variables only few are fundamental for the visual analysis:

- DISTRICT: which provide the Code of the district,
- YEAR: which provide the year when the crime was reported,
- MONTH: which provide the month when the crime was reported,
- DAY_OF_WEEK: which provide the day of the week for the crime,
- HOUR: which provide the hour when the crime was reported.

Frequency of Crimes per District for each year:

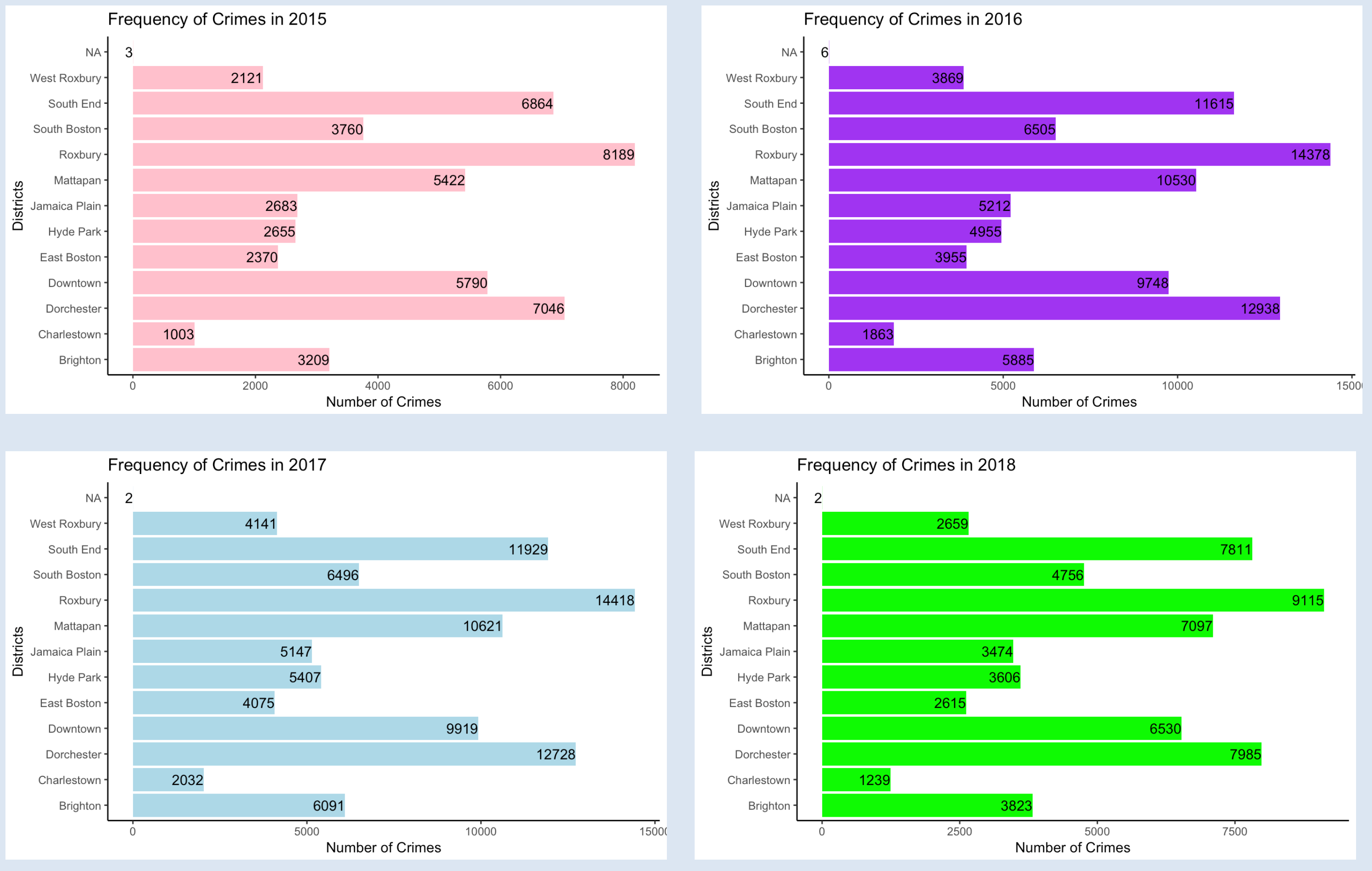

Fig. 2: Histograms of total number of crimes per district for each year (from June 2015 to August 2018

The three districts with the higher amount of crime reports are considered for the model and prediction: Roxbury (B2), South End (C11) and East Boston (D4). These represent more than 42.3% of the total number of crimes in Boston. Plots and analyses are displayed only for district B2. During the fitting process it was noticed that the months 06/2015 and 09/2019 are not complete with all the reports. Because of these two months, the fit of the model was altered in both cases giving also a large error in the forecast process. The two months are therefore cut-out from the model fitting process.

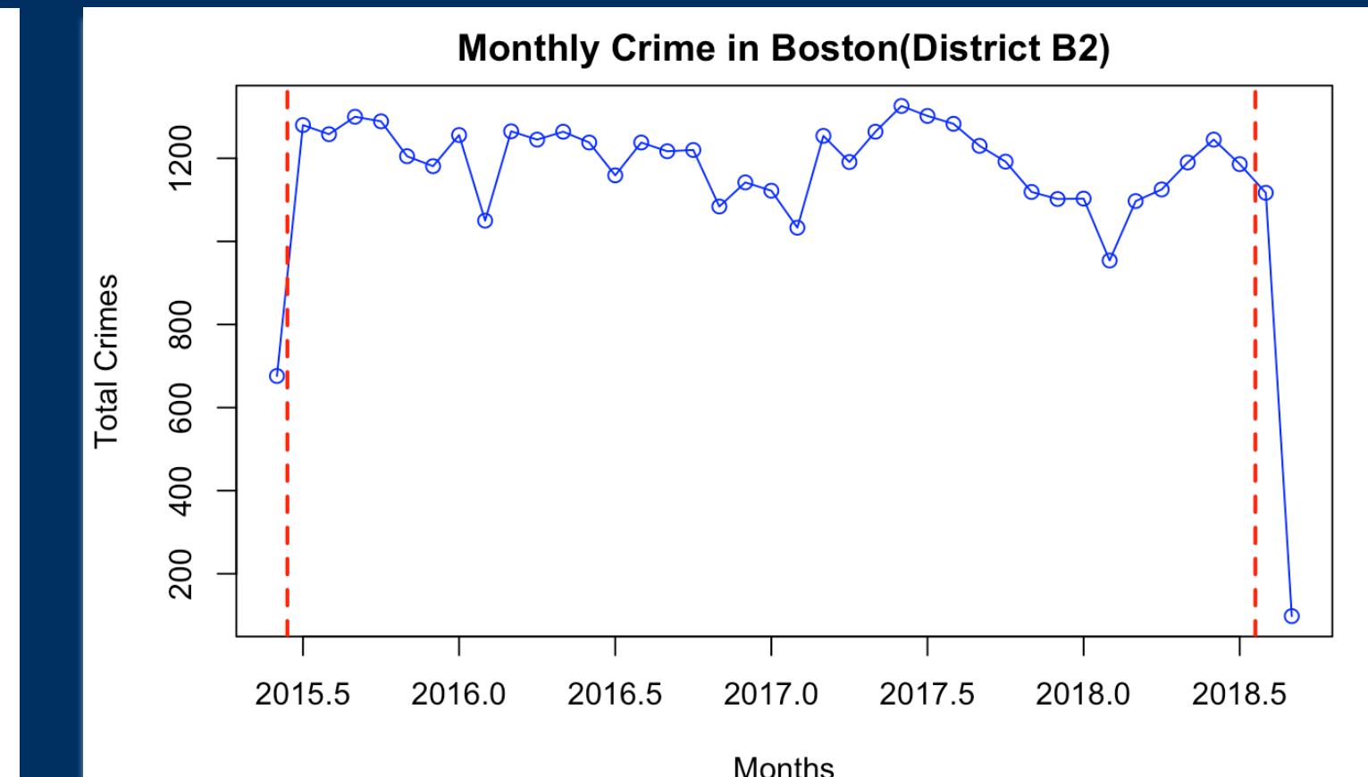

Fig. 3: Plot of the total data for district B2.

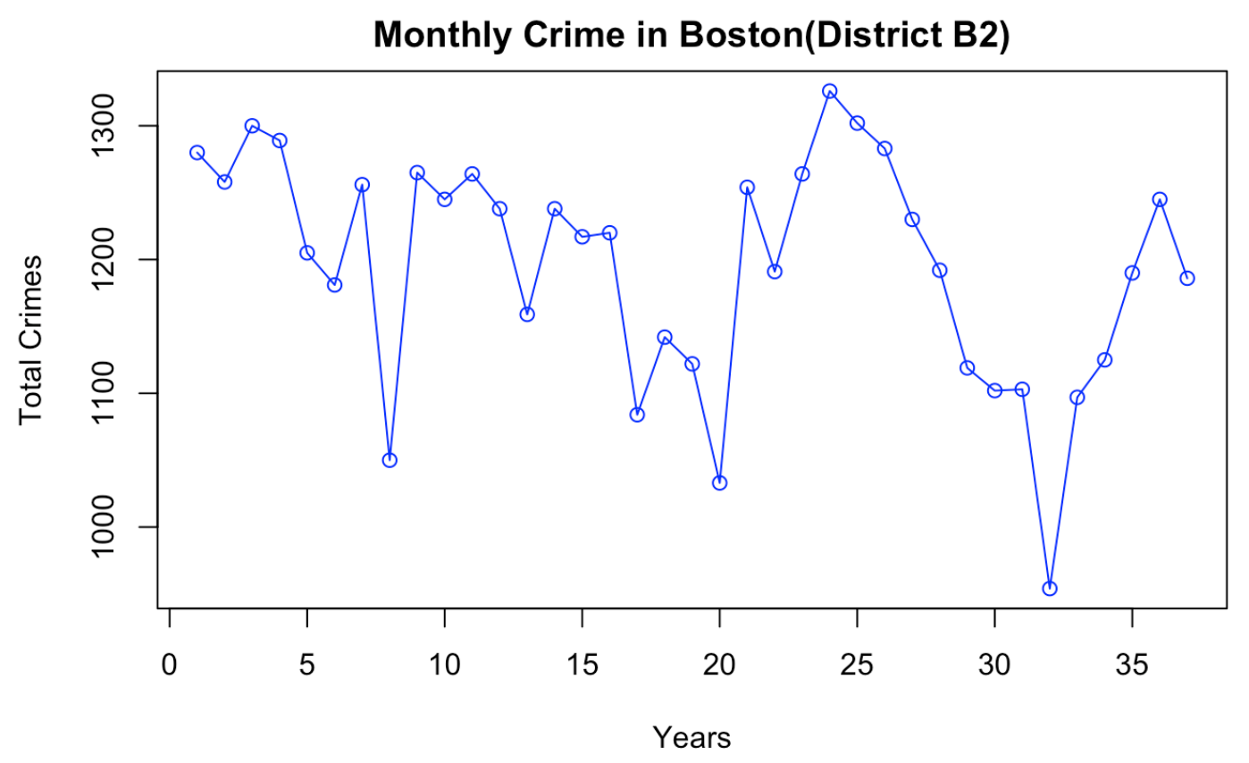

Fig. 4: Plot of the data for district B2 for fitting process.

AR(p)

We first consider the AR(p) model, which is given by:

$$Y_t = \mu + \sum_{i=1}^{p} \phi_i \left(Y_{t-1} - \mu\right) + Z_t,\ Z_t \sim N(\mu, \sigma^2)$$

where p is the order of the process. August 2018 is also removed from the analysis since it would be used to assess the predictive accuracy of the model (test data).

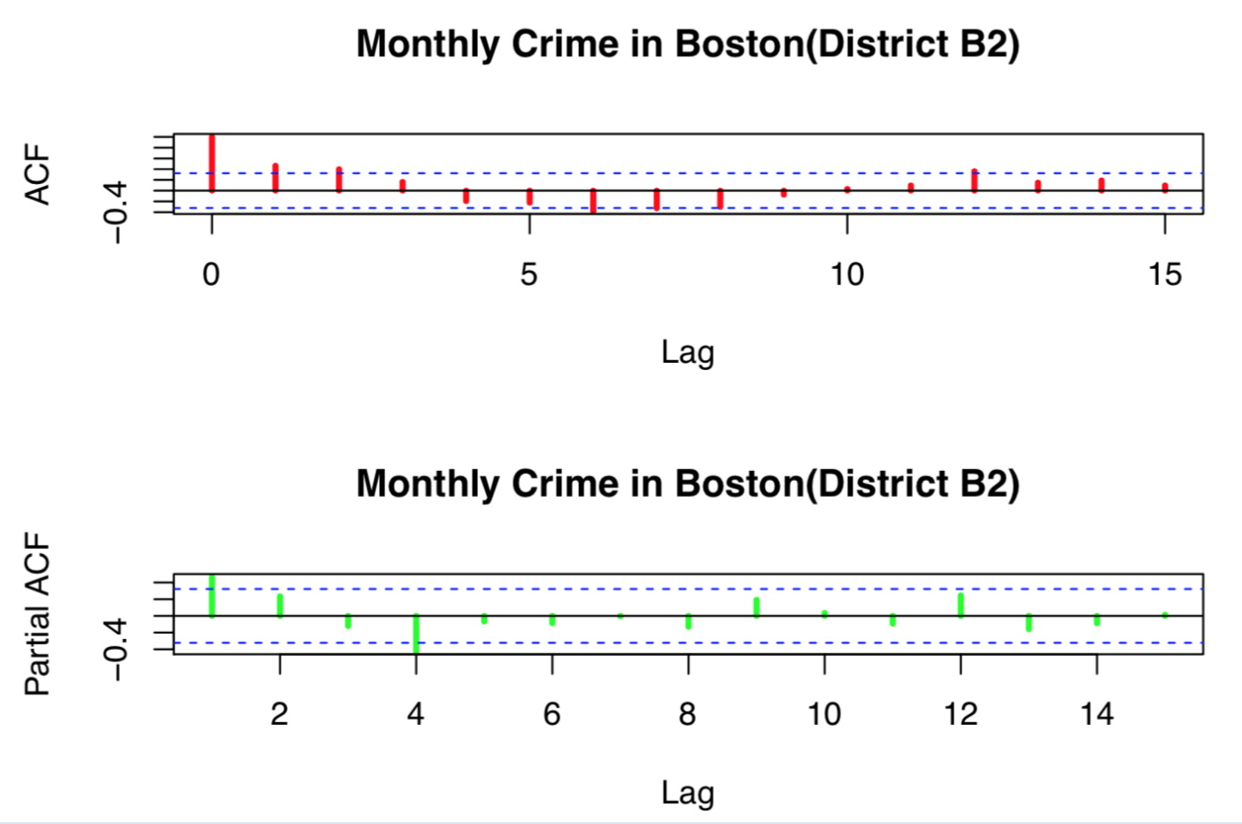

Fig. 5: Plot of ACF and PACF for district B2.

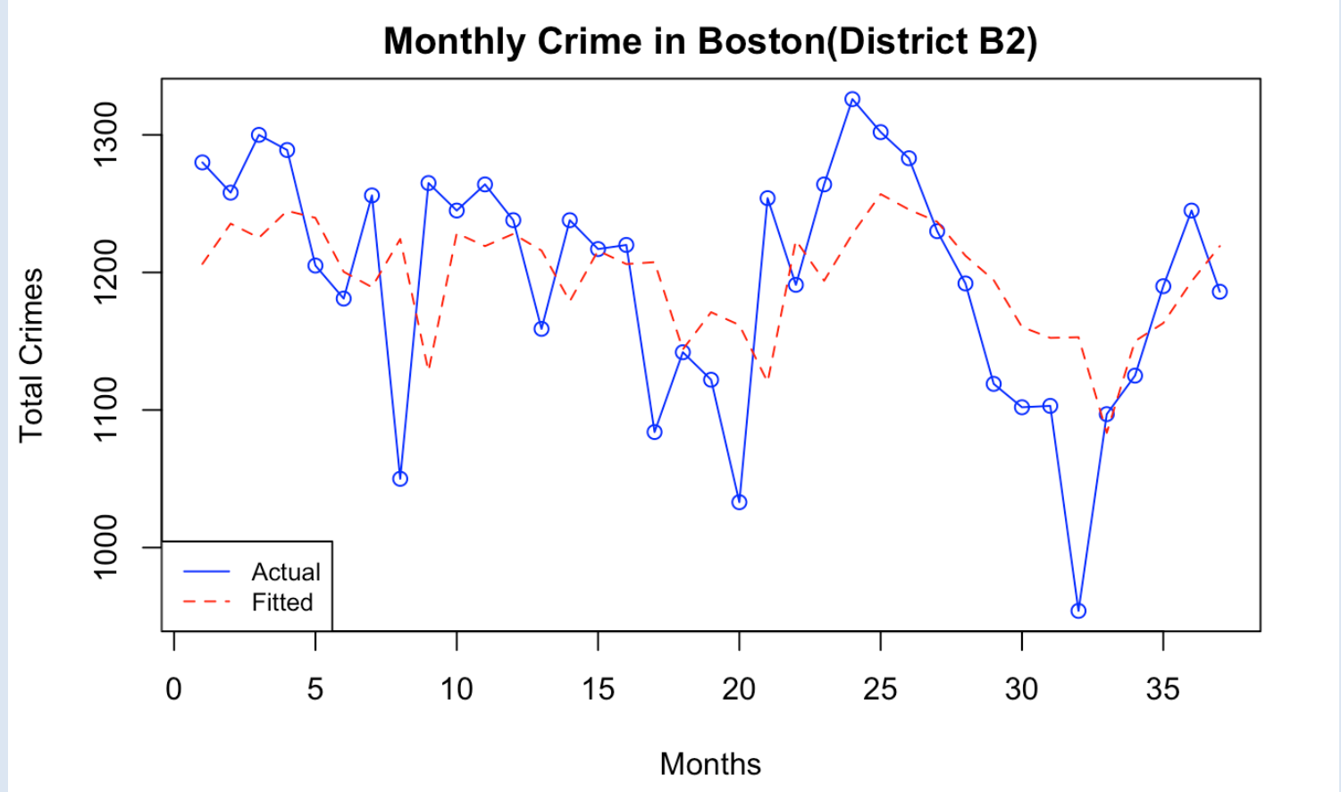

Fig. 6: Plot of the fitted model for district B2.

```
arima(x = ts_data_DB2, order = c(1, 0, 0))

Coefficients:
                 ar1    intercept
              0.4664    1196.5569
      s.e.    0.1439      22.4781

σ^2 estimated as 5564:  log likelihood = -212.17,  AIC = 430.34
```

GPR

GP models give distributions for the predictions. Realizations from these distributions give an idea of what the true function may look like.

Kernel Selection: To do prediction for a GPR model the kernel must be specified. For this research, the decision is to try three different kernels and to observe the differences:

- Gaussian (RBF) Kernel,
- Matern5/2 Kernel (nu = 2.5, twice differentiable functions),
- Exponential Kernel.

The best predictions only are displayed, these are provided by the GPR with the exponential kernel (stationary kernel) parameterized by a length-scale parameter greater than zero.

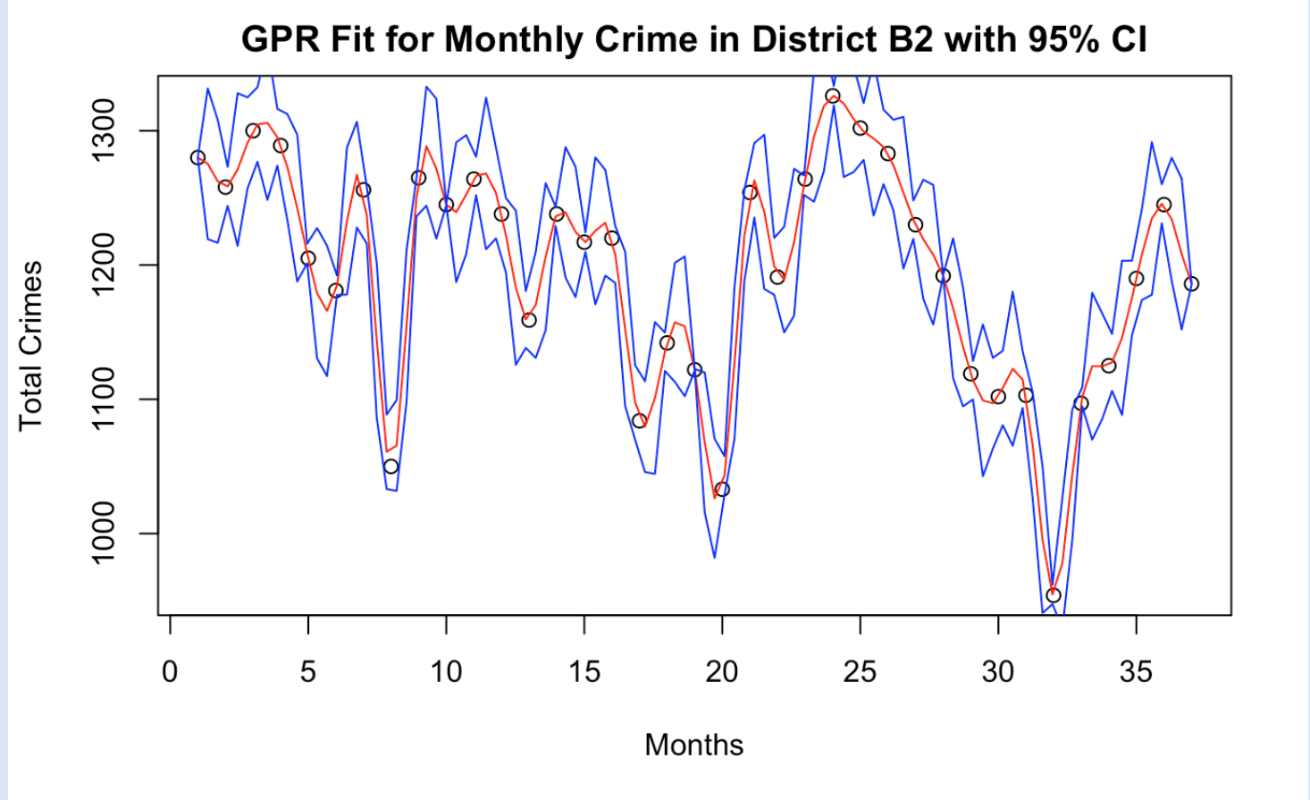

Fig. 7: Plot with the fitted GPR model and the PI for district B2.

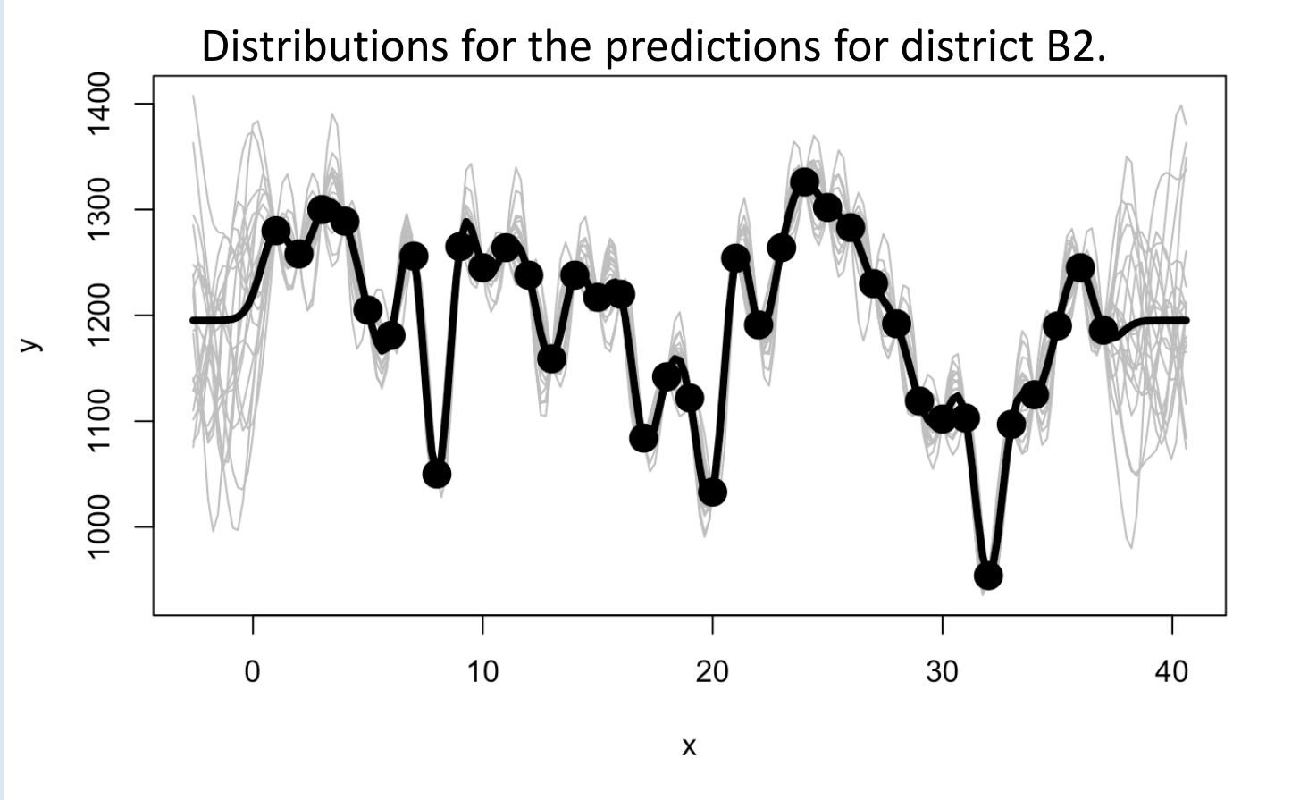

Fig. 8: Plot with the distributions for the predictions for district B2.

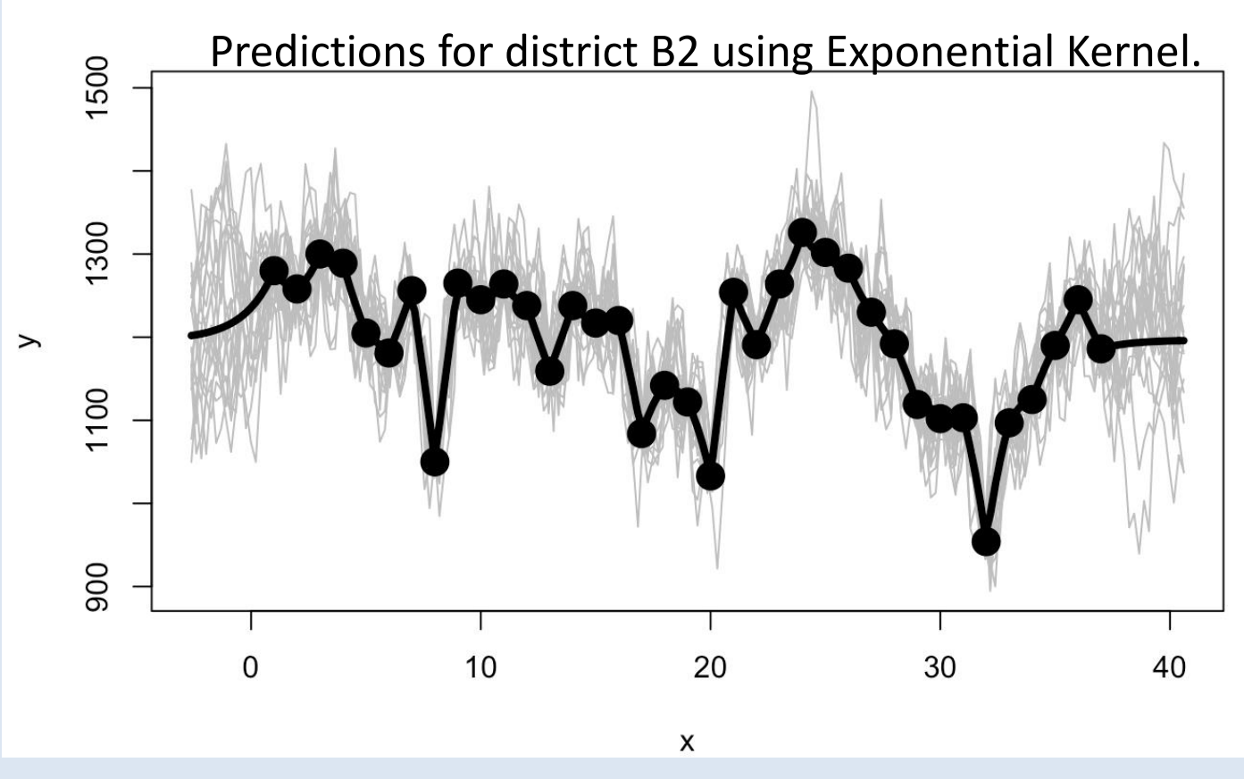

Fig. 9: Predictions for district B2 generated by GPR using Exp. kern.

Kernel	Prediction	S. Error
Gaussian	1198.53	6.80%
Matern 5/2	1198.85	6.82%
Exponential	1196.50	6.60%

Results

The total number of crimes for the month August 2018 was forecasted using the AR(1) and the GPR model (with exponential kernel) applied to the districts B2, C11 and D4. Then, the error with the actual number of crimes for the same month was calculated.

The comparison between the prediction obtained from the two model shows that with AR(1), there is a standard error generally less than 7%.

Using GPR the error obtained is about 1% smaller that the previous model.

For all the districts with the GPR model, it is observed that the analysis using the gaussian and the exponential kernel perform better. The exponential kernel has a huge effect on the model fit. This assumes the correlation between points dies off very quickly. Therefore, there is much more uncertainty and variation in the predictions and sample paths, and this allows to achieve a more accurate result for the prediction.

District	Actual Report	GPR pred.	GPR Error	AR(1) pred.	AR(1) Error
B2	1117	1196.5	6.6%	1196.6	6.7%
C11	970	1037.6	6.4%	1060.1	8.5%
D4	982	991.9	1.1%	996.6	1.5%

Discussion

Honestly, a different result was expected from this research. That is, if the predictions are “really close” with a minimum difference, and the standard errors are “really low” the goal was to obtain a significantly better prediction from the GPR model, compared with the AR(1).

One reason for this unexpected result may be attributed to the dataset —this is a time series with only two years of full records. In fact, the year 2015 and 2018 are recorded for half of less of the total length. This lack of information may lead to a wrong pattern identified by the GPR model that translate as "poor" prediction.

Another reason may be related to the lack of seasonality in the data, in fact GPR prediction can be performed with stationary kernel (as performed in the project) or with non-stationary, this last model takes care of the seasonality and it is reasonable to think that it would provide a better fit and more accurate predictions.

Future interest research in this area can involve a similar analysis with a more completed data set and introducing seasonality to the GPR model and compare with a SARIMA model .

References

- https://www.kaggle.com/AnalyzeBoston/crimes-in-boston
- https://www.kaggle.com/ankkur13/boston-crime-data
- https://towardsdatascience.com/a-visual-comparison-of-gaussian-process-regression-kernels-8d47f2c9f63c
- https://www.kaggle.com/gudipallyp/boston-crime-data-analysis
- https://en.wikipedia.org/wiki/Exploratory_data_analysis
- https://mail-attachment.googleusercontent.com/attachment
- https://cran.r-project.org/web/packages/GauPro/vignettes/GauPro.html
- https://rdrr.io/cran/kernlab/man/gausspr.html
- https://scikit-learn.org/stable/modules/gaussian_process.html#gaussian-process-kernel-api
- https://github.com/lightning-viz
- https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
- https://www.stat.berkeley.edu/~arturof/Teaching/STAT248/lab07_part1.html
- https://scikit-learn.org/stable/modules/gaussian_process.html
- Carl Eduard Rasmussen and Christopher K.I. Williams, “Gaussian Processes for Machine Learning”, MIT Press 2006