Report of the Project regarding the C.U. Time Series

Exploratory Analysis of the All Fire/Emergency Incidents in New York City, from 2009 to 2017

Master in data science

Project made by Inês Tavares

Introduction

This work consists of the analysis of a database regarding fires/incidents reported by the Fire Department of New York City (FDNY), in a space of 8 years, containing 4368 total occurrences as well the average response time for each, by the FDNY. This information was obtained using the Data World website, with data provided by the FDNY [1].

The database contains information citywide and its 5 boroughs (Bronx, Brooklyn, Manhattan, Queens and Staten Island), and categorizes the fire/incidents: "All Fire/Emergency Incidents", "False Alarm", "Medical Emergencies", "NonMedical Emergencies", "NonStructural Fires" and "Structural Fire".

It was chosen to disregard the divisions of the database, in favour of an overall view of the city's needs. Considering this, this project focuses on the "All Fire/Emergency Incidents" in the "citywide" context.

As such we were left with 96 entries of 2 variables, "YEARMONTH" and "INCIDENTCOUNT", the first ranging from July 2009 to June 2017, and the second being the number of occurrences in each month of a given year, as stated by the first variable.

The main goal is to evaluate and study the evolution of All Fire/Emergency Incidents in NY city, throughout the years, and to forecast future happenings.

This project was developed in R Studio.

Exploratory Analyses

Before going into the analysis, the database included information about each month of the Year and for each specific year. Since the latter is redundant, the rows containing only the year were excluded from the analysis.

Therefore, the final database has 96 entries and 2 variables, with no missing values or wrong values.

To have a better understanding of the time series, a plot (Figure 1) was made. The time series has its highest values after 2015, and the lowest before 2014. It shows a tendency, that seems not to be linear and a non-constant variance, making the series non-stationary.

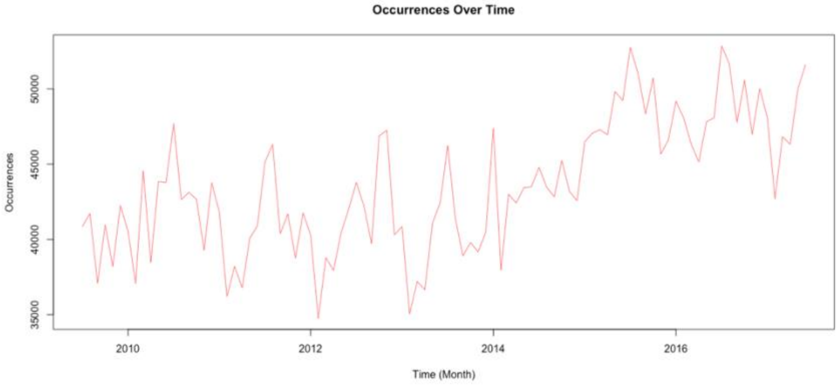


Figure 1- Plot of the Occurrences over Time (from July 2009 to June 2017)

When plotting the lags of the time series, it accentuates the existence of a trend, since all values of the slope are around 0,5.

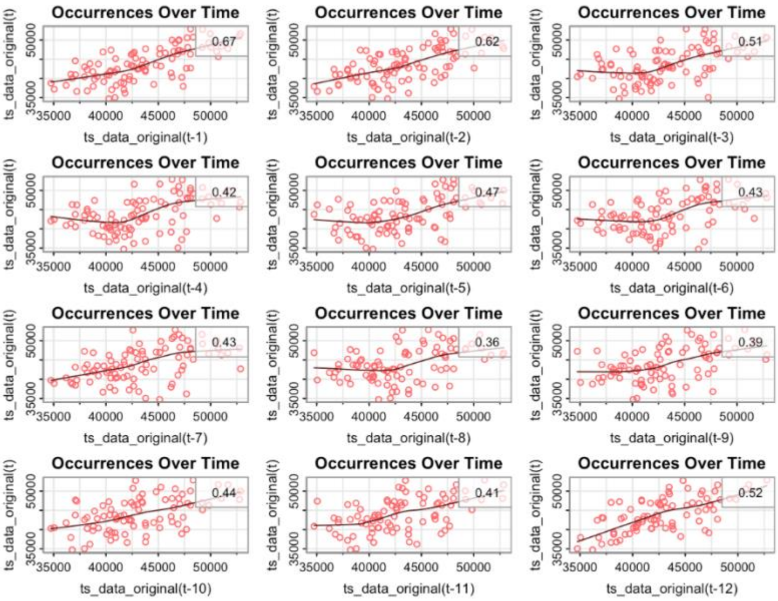


Figure 2- plot of the Lags

To analyze the seasonality, two seasonal plots were made (Fig.3). In both graphics the variance between the months is noticeable, and again, a major factor in making a series nonstationary.

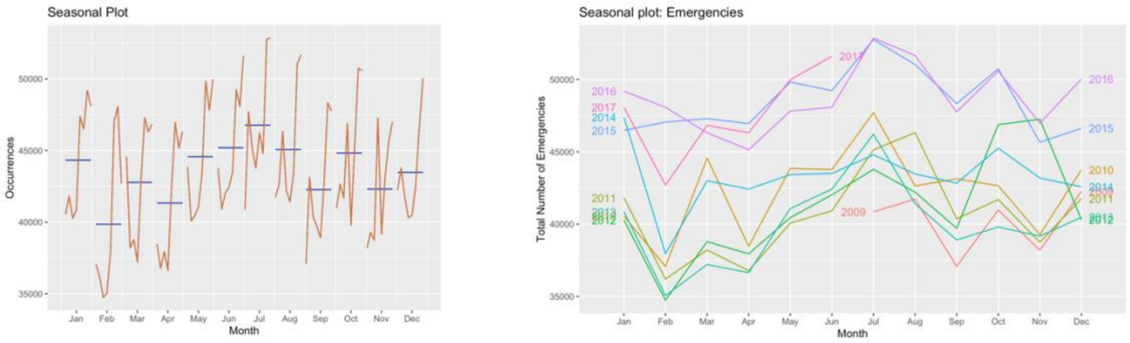


Figure 3- Seasonal Plots

As it was said before, visually, the variance is not constant. To overcome this problem, the time series was transformed, through a BoxCox transformation and a Days of the Month Adjustment (Fig. 4).

Comparing these two transformations, the Days of the Month Adjustment seems more constant, but it loses the period component, and the variance is higher, so the Box-Cox was chosen.

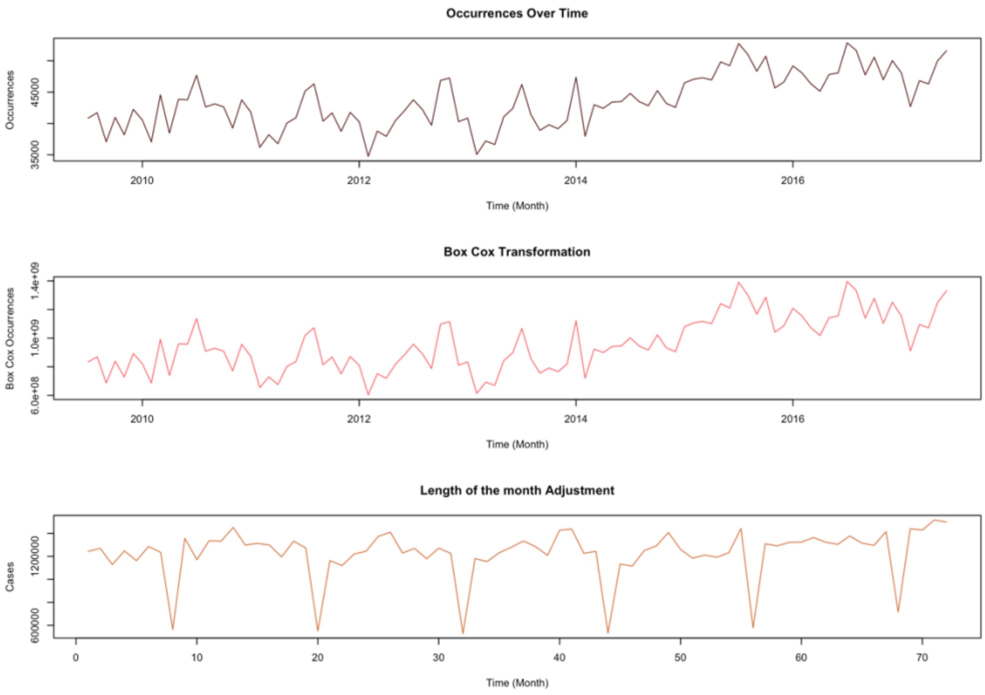


Figure 4- Top Panel: Time series; Middle Panel: Box-Cox Transformation; Botton Panel: Days of the Month Adjustment

Decomposition

As was said before, the series is not stationary, and seasonality seems to play a big role, so it is important to handle this situation first, by having a seasonal differentiation with order 1 of the time series.

After differentiating the time series, it is possible to compare the series before and after this procedure (Fig. 5 Top and Middle Panel). On the series deseasoned, the seasonality appears to be stabilized (Fig. 5 Bottom Panel), the series is not trend stationary.

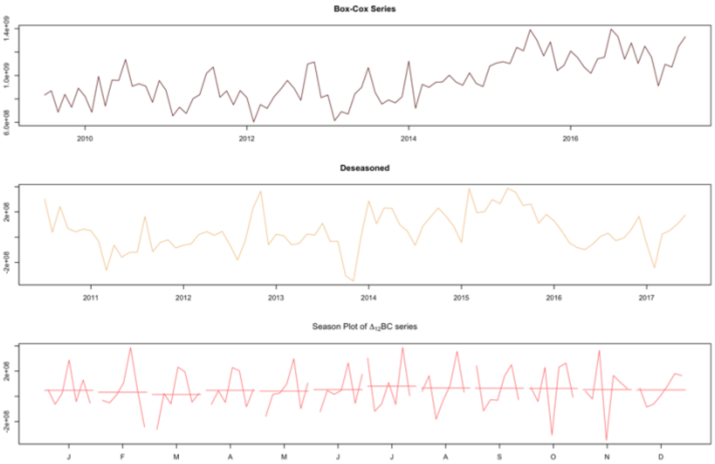


Figure 5- Top Panel: Box-Cox of the Time Series; Middle Panel: Series Deseasoned; Bottom Panel: Season Plot of the Seasonal Differentiated Series

To detrend the series, another differentiation is needed, but not seasonal.

Fig.6 shows the series deseasoned (Top Panel) and the series deseasoned and detrended (Bottom Panel). The last series appears to be stationary.

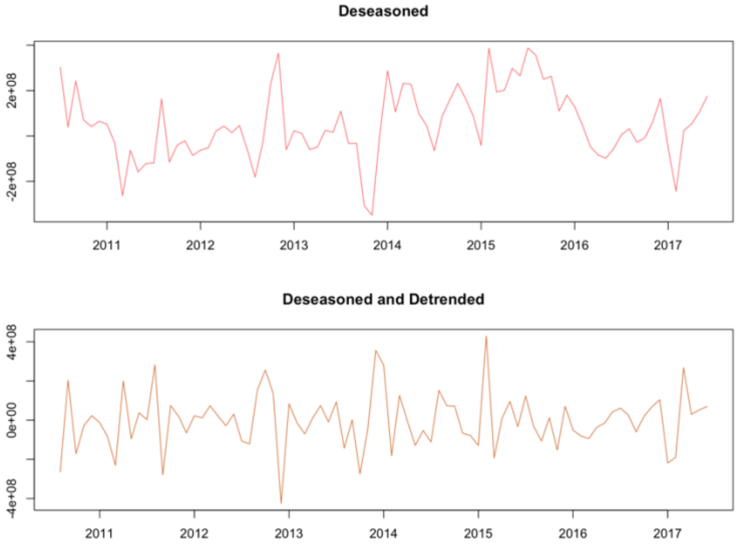


Figure 6-

In Fig. 7 (Appendix 1), there is a multiplot of different lags of the deseasoned and detrended series. In these lags, the trend is almost null, which verifies what was said previously.

To test if the last series is stationary, two tests for unit root were performed: the KPSS Test and the Augmented Dickey-Fuller Test. In both tests, the result was that this series is stationary. (Appendix 2).

Before we go to try to find the best model, Plots for 4 time series, and their respective ACF and PACF were made. This series are the box-cox transformation (bc), the first differentiation of bc, the first seasonal differentiation of bc and two differentiations of bc (one being seasonal). These plots allow us to have a better understanding of which SARIMA (p,d,q)x(P,D,Q)12 model would fit better.

Looking at the ACF and PACF plots, the Δ_{12}bc and the $\Delta\Delta_{12}$bc have the best results regarding the non-correlation between residuals, however, the ACF of Δ_{12}bc show that residuals are in some way in groups. The ACF and PACF plots of $\Delta\Delta_{12}$bc lead to the assumption that a seasonal auto-regressive model should be applied.

Face this, the starting point is to have d=1, D=1, and P=1.

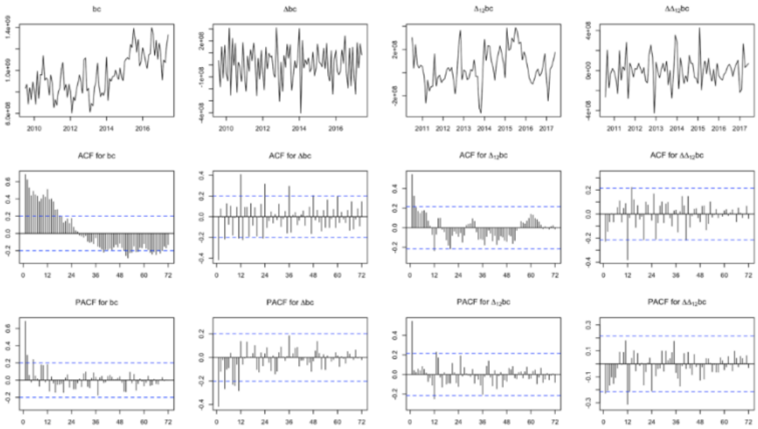


Figure 7- Plots for 4 time series , and their respective ACF and PACF

Modelling

Before beginning to search for the SARIMA that would fit best, the dataset was separated into a test and a training set, the first having the last year cycle and other the remaining cases. This will help evaluate the model.

With the training set, we began to find the parameters for the seasonal component of the serial model. For this task, a few models were tested. The first criterion to accept a model is to have significant parameters. After eliminating those that didn't have, the best model would be the one to have all values pass the Ljung-Box test, and the residuals to be non-correlated and in the q-q norm.

The model that covers the requisites and has the best AIC was with the following parameters: P = 2, D = 1 and Q = 0 (Appendix 3). With this finding, two plots were generated, to have a better understanding of the remaining parameters- ACF and PACF for the residuals (Fig.8). These plots suggest that the non-seasonal component should have a moving average.

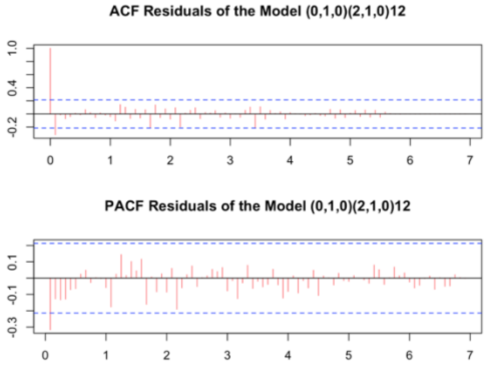


Figure 8-

Following this, we proceed to find the remaining parameters. Three models passed the criteria, so the model that had the lowest AICc would be chosen (Appendix 4). We opted to choose the AICc as a decision criterion since the sample is small [2].

The model that had the lowest is: SARIMA (0,1,1)x(2,1,0)$_{12}$

This final model is shown in Fig. 9, and it has the residuals non correlated that fit in the q-q norm, have a mean equal to 0 and pass the Ljung-Box test.

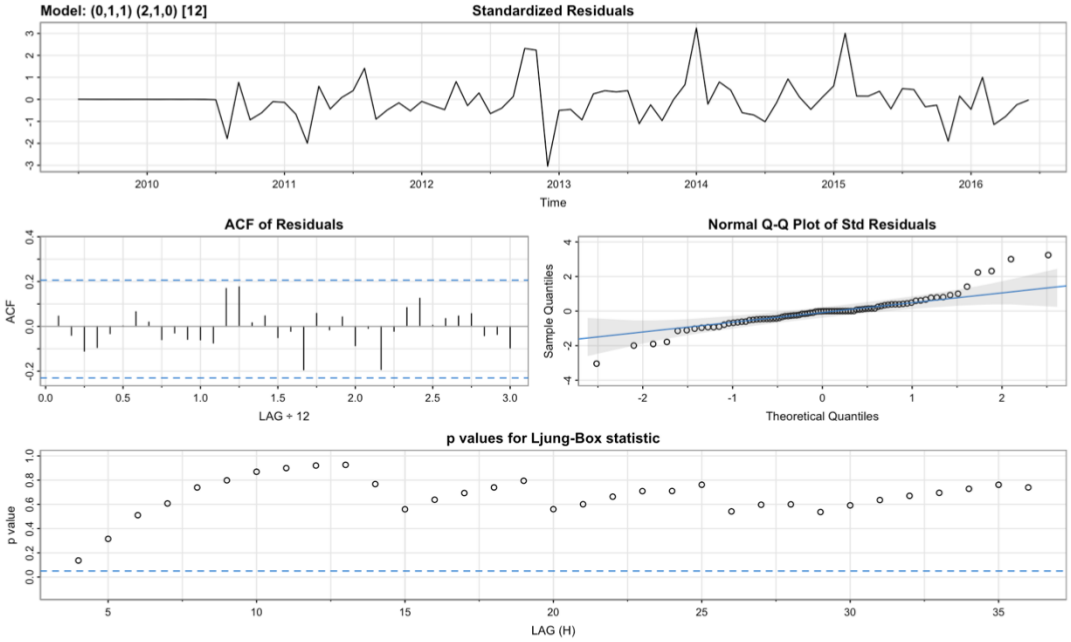


Figure 9-

Testing the Model

Since we performed a Box-Cox transformation, we had to invert the series.

To evaluate the predictions of the model, we compared the predictions of the model of the last year, with the training set.

Fig. 10 shows the fitted values of the model, the prediction for the last year, and the observed values of the last year.

The predicted values don´t align completely with the observations, and this may be because the trend doesn't follow a known function, being harder to manipulate.

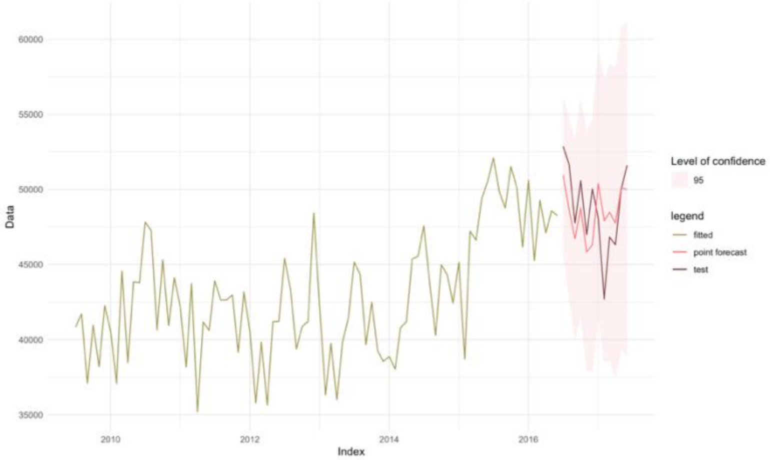


Figure 10- Plot of the model fitted values, model forecast of the last year cycle and the test set

Forecast

In Fig.11 there is the forecast out of sample of the SARIMA model, and it shows that the incidents will increase in the next 2 years.

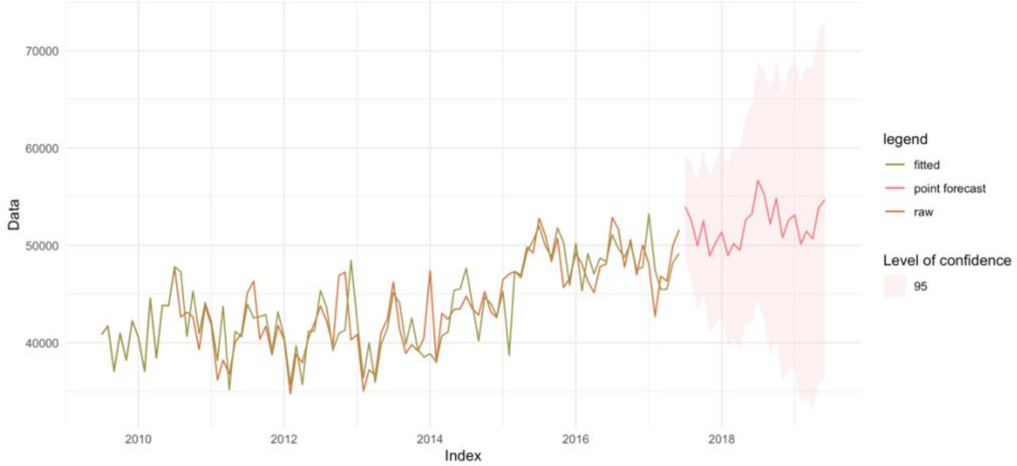


Figure 11- Forecast out of sample using the SARIMA model

Bagged ETS forecast

Bagging, short for Bootstrap Aggregating, is a technique that involves training multiple models on different subsets of the training data and combining their predictions to improve overall accuracy and robustness. When combined with exponential smoothing methods can produce highly accurate forecasts and improve the forecast accuracy relative to traditional methods [3].

As it was expected, these methods performed better than the SARIMA model, but still failed to "align" completely (Fig.12).

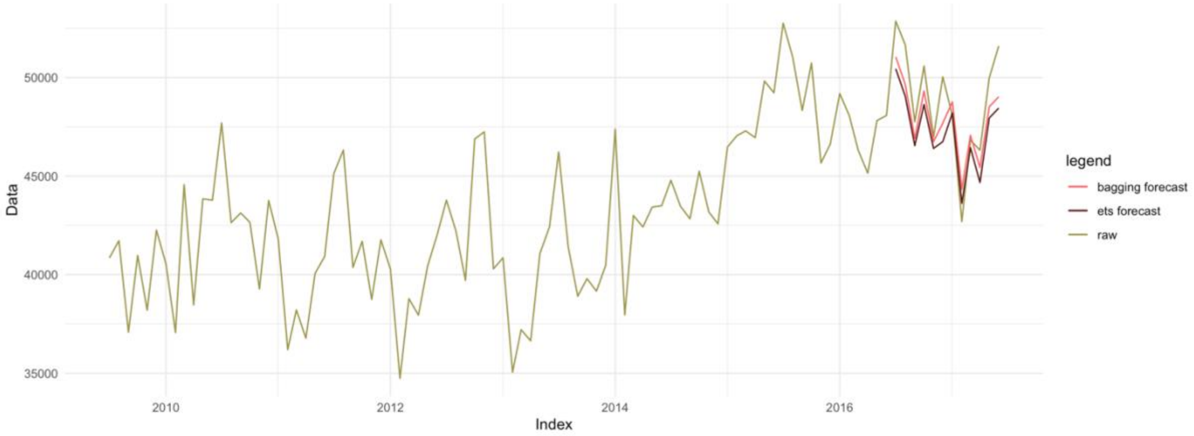


Figure 12- Comparing bagged ETS and ETS applied directly to the data.

In Fig.13, there is the forecast of the next 2-year cycles of these models. Despite the prediction of the SARIMA model, these methods predict that the trend is going to stabilize.

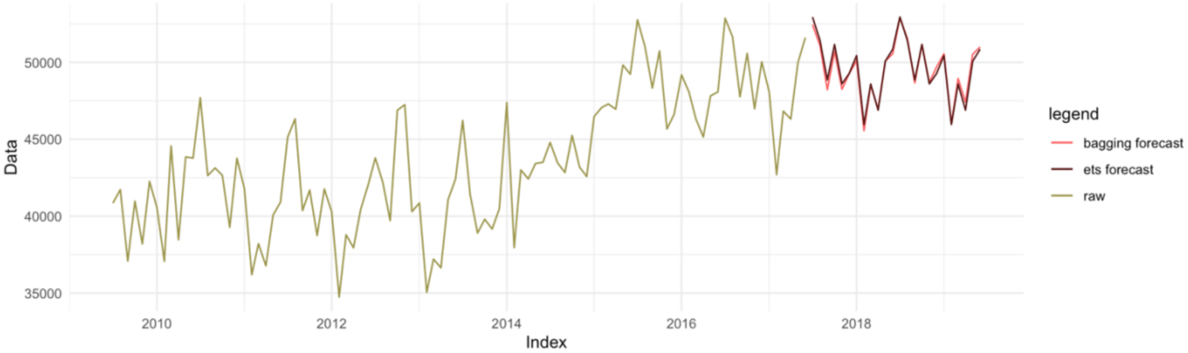


Figure 2- Forecast out of sample of bagged ETS and ETS

Discussion

In this report, we obtained a SARIMA model and a bagging ETS for all incidents that the fire department of NY had to face from July 2009 to June 2017.

The major limitation of this study was that the trend time series didn't follow a known function, which made it harder for the models to predict future events.

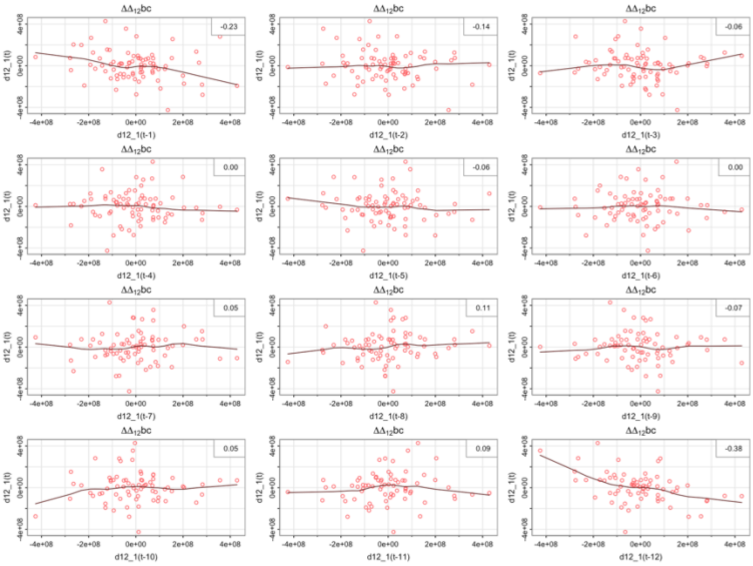
However, despite this limitation, the SARIMA model predicted that the total of incidents would increase in the next two-year cycles, which ended up being true.

Bibliography

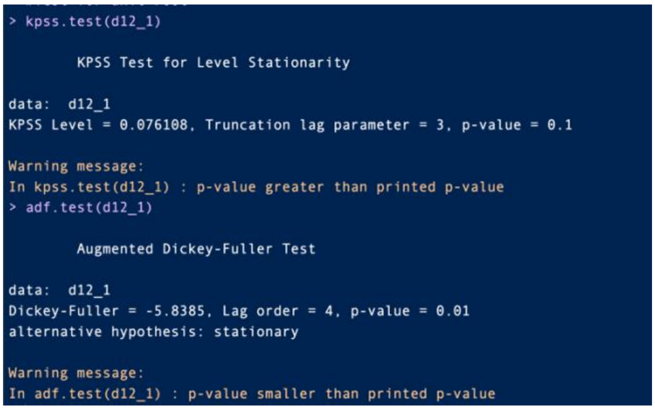
1. https://data.world/city-of-ny/j34j-vqvt (October 10th 2023)
2. https://github.com/angela-xu/aic-aicc-performance-comparison-in-model-selection (January 3rd 2024)
3. https://www.sciencedirect.com/science/article/abs/pii/S0169207018300888 (January 13rd 2024)
4. https://otexts.com/fpp2/bootstrap.html (January 13rd 2024)

Appendix

1.

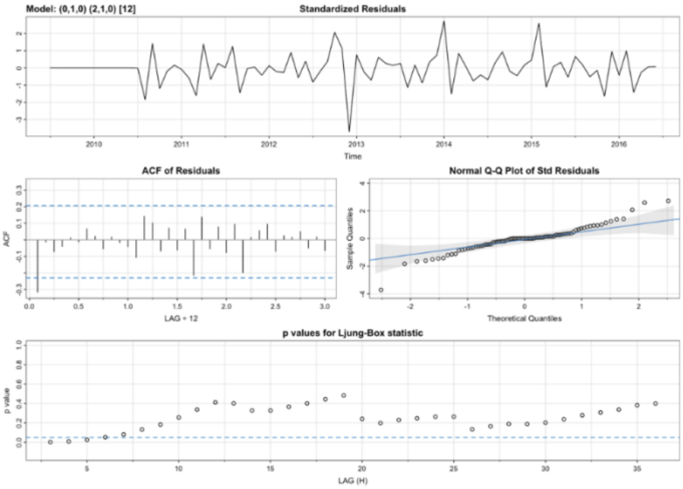


2.



```
> kpss.test(d12_1)

        KPSS Test for Level Stationarity

data:  d12_1
KPSS Level = 0.076108, Truncation lag parameter = 3, p-value = 0.1

Warning message:
In kpss.test(d12_1) : p-value greater than printed p-value
> adf.test(d12_1)

        Augmented Dickey-Fuller Test

data:  d12_1
Dickey-Fuller = -5.8385, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(d12_1) : p-value smaller than printed p-value
```

3.



4.

Model	(0,1,0)x(2,1,0)12	(1,1,0)x(2,1,0)12	(0,1,1)x(2,1,0)12
AICc	40.24844	40.16883	40.13796