Kenneth Luna

The data selected for this time series exercise is NYC 311 Calls starting from January 2010 until the end of March 2015. [[1]](#footnote-1) The dataset was not filtered for complaint type; therefore the data contains calls ranging from heating, noise, traffic etc. An initial dataset contained only housing complaints to the NYC Housing Preservation and Development Agency but it was decided that exclusion of other complaints would limit some interesting characteristics of call patterns to 311. This will be discussed later in the paper when we adjust for seasonality.

The length of the dataset was 68 observations after grouping by month and year. The date was read as a factor type but was converted to a date value using R’s as.Date() function. The last observation was removed from the data set as it contained data for only April 1st 2015.

**Figure 1** highlights the preliminary time series plot for 311 calls. What first stands out in the plot is that 311 calls from 2010 to present day show seasonal characteristics. The volume of 311 calls appears to increase during the winter and the summer, but then drops off during the spring and fall. Furthermore, there appears to be several shocks in the dataset, which correspond to severe snowstorms in NYC. To handle this volatility and seasonality, the log of 311 calls will also be evaluated.

**Figure 2** shows the time series plot for the log of 311 calls. There is not much difference between the log of 311 calls versus that of the absolute counts of 311 calls. Most of the same seasonal patterns and deviations continue to appear after this adjustment. For the rest of this exercise, the log 311 call (log(calls)) will be used.

The ACF and PACF for log(calls) are shown in **Figure 3** and **Figure 4** respectively. The behavior of the ACF and PACF show slight indications of an AR model; the former dies down and the latter cuts off immediately. Unfortunately, both the ACF and PACF show some structure around lag-8 and later. This structure will need to be removed.

Before solving for the best ARMA model, the time series was seasonally adjusted by subtracting the corresponding monthly averages from each monthly observation. **Figure 5** is a stepwise representation of the monthly averages across the entire dataset. Interestingly, we see spikes in calls during the winter and summer months; the calls in the winter are due to heating issues during the cold and the calls in the summer are due to noise complaints as everyone spends time outdoors.

After adjusting for seasonality by subtracting the monthly average from each monthly observation, the result plot is **Figure-6.** The seasonality and structure realized in the prior plots is now no longer present. We can verify this by observing the ACF and PACF for this seasonally adjusted plot. **Figure 7** and **Figure 8** show characteristics that make it consistent with an ARMA model as both ACF and PACF appear to slowly die down and any structure realized after lag-8 in the prior ACF and PACF plots is no longer present.

The first difference and the second difference of the seasonally adjust log(calls) were taken into consideration for this exercise. Unfortunately, for both difference-1 and difference-2, the lag-1 results were negative, indicating that there was over-differencing. The result for difference-1 can be found in **Figure-9** and **Figure-10** and difference-2 can be found in **Figure-11** and **Figure-12** respectively.

Based on this preliminary analysis, the AICC values were then considered. The AICC values for all eighteen different ARIMA models can be reviewed in **Table-1.** Based on these results, the best model is ARIMA(2,0,2) without a constant with an AICC value of -166.12. The parameter estimates for the selected model can be found in **Table-2.** The complete form for the model is the following:

*x(t) = .4622 \* x(t-1) + .3911 \* x(t-2) + eps(t) + .2932 \* eps(t-1) + -.5413 \* eps(t-2)*

Based on the z-statistics, not all of the parameters appear to be statistically significant. The AR(1) and MA(1) have a z-statistic less than two, indicating there might an issue with these parameters in our ARMA(2,0,2) model. Fortunately, z-statistics are not always reliable therefore this can be overlooked.

The Ljung-Box statistics for the first 12, 24, 36 and 48 lags had p-values above .05, indicating that the current model is stationary and that there is no more dependence. Based on these p-values, one can conclude that this model is adequate for this analysis. The p-values can be found in **Table-3.**

The plot for the residuals in **Figure-13** does show a slightly discernible structure by tending to veer below the mean but then shoot above it entering winter; fortunately the ACF and PACF in **Figure-14** and **Figure-15** respectively reveal that any autocorrelation has been eliminated.

The forecast plot with lead times 1-50 and a 95% confidence interval can be found in **Figure-16.** The forecast does seem reasonable based on this plot. It captures the entire range of monthly log(311) calls after solving for seasonality. One can question the value and validity of the forecasts beyond a certain lead-time; it is difficult to forecast 50 observations in the future with a dataset of only 68 monthly observations. Based on this forecast model, one can predict that the range of 311 calls in the month of April will be the following:

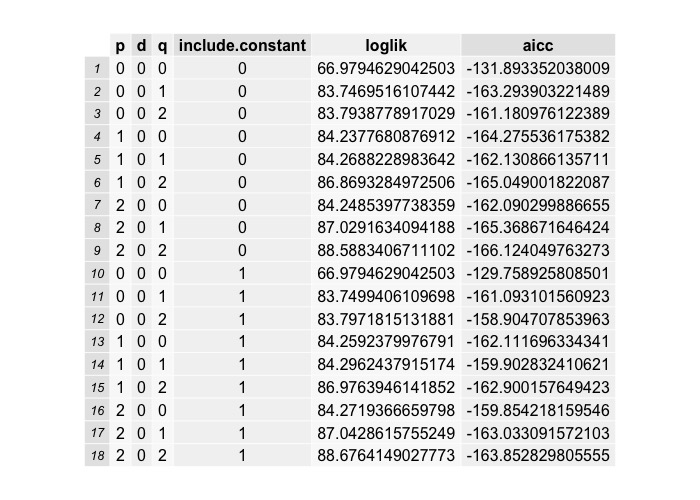
|  |  |
| --- | --- |
| **Confidence Interval** | **Values** |
| Lo 95% | 13,7582 |
| Point\_Forecast | 14,0597 |
| Hi 95% | 17,3109 |

This estimate for April 2015 is reasonable.

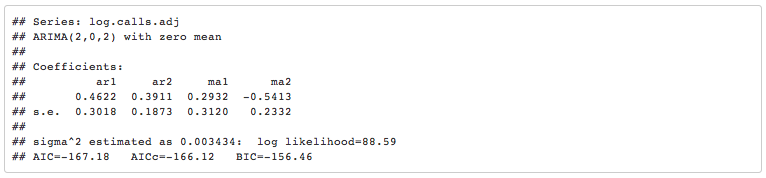
After building this initial forecast model, one could take into consideration an AR(1) model based on the ACF and PACF plots in **Figure-7** and **Figure-8**. After reviewing the results of an ARMA(2,0,2) forecast model and the questionable z-statistics for the coefficients, it would be worthwhile to see if there is a better fit using an AR(1).

Overall, this dataset was not the easiest to use to build a forecast model. 311 calls in NYC are highly seasonal. The original dataset, which was daily, required adjusting for both weekly and monthly seasonality. Furthermore, the dataset does not go far back enough in time to capture more historical trends. 311 call centers have been around for over a decade but only recently has the city mandated collecting and publishing this data.

**Table 1**

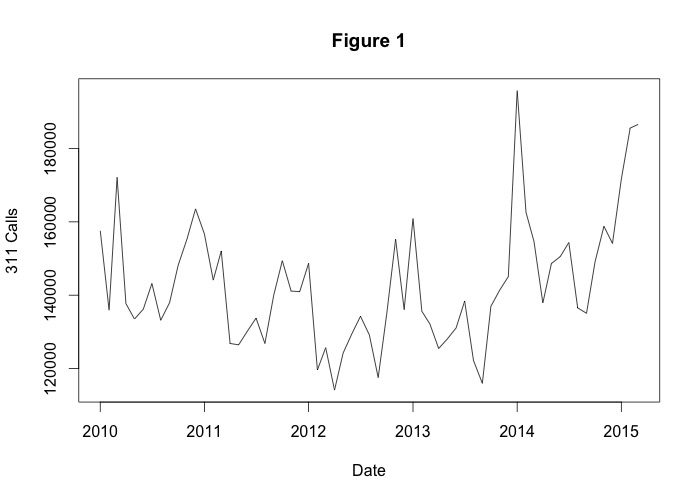


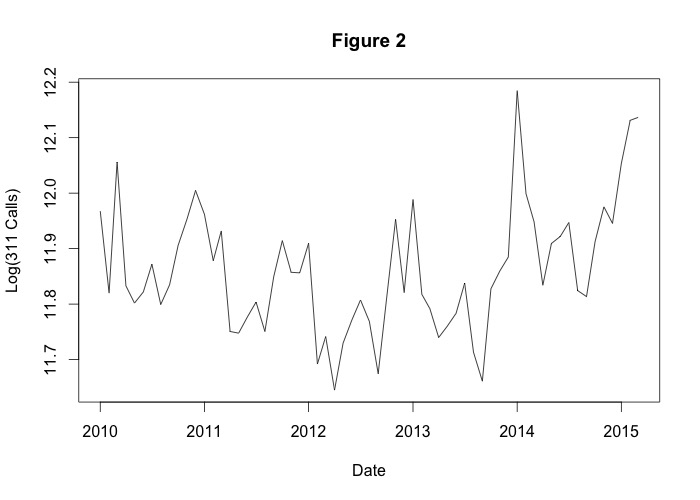
**Table 2**

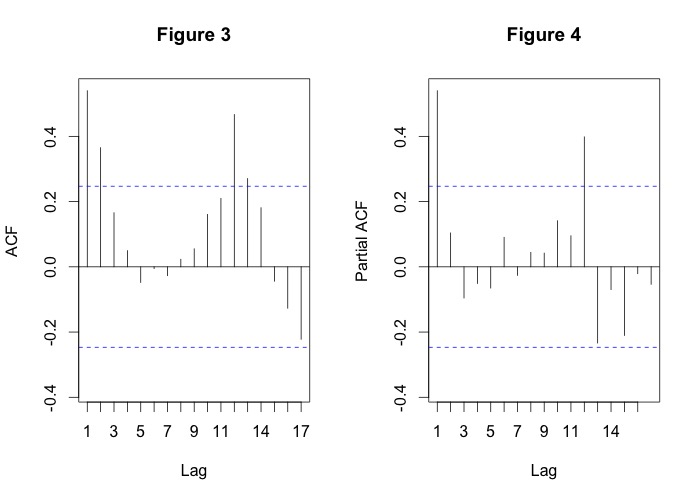
****

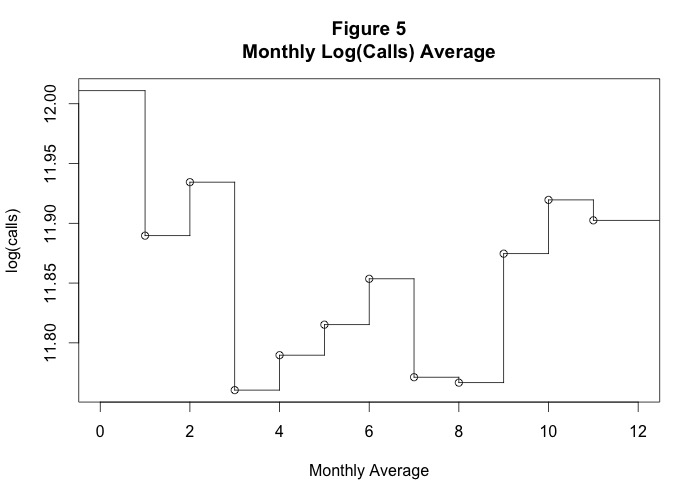
**Table 3**

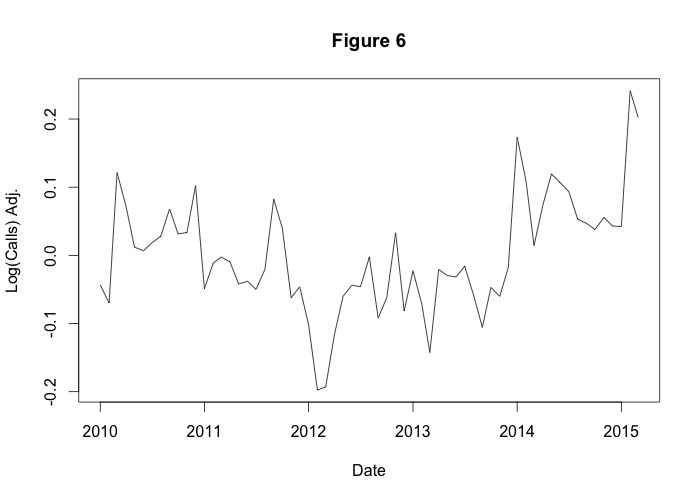
|  |  |
| --- | --- |
| Ljung-Box Tests | p-values |
| Box.test(resid, lag=12, type = "Ljung-Box", fitdf=4) | .84 |
| Box.test(resid, lag=24, type = "Ljung-Box", fitdf=4) | .47 |
| Box.test(resid, lag=36, type = "Ljung-Box", fitdf=4) | .15 |
| Box.test(resid, lag=48, type = "Ljung-Box", fitdf=4) | .36 |

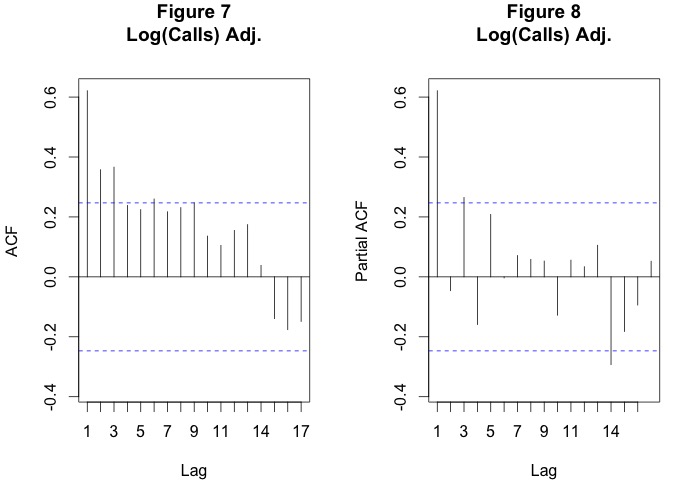


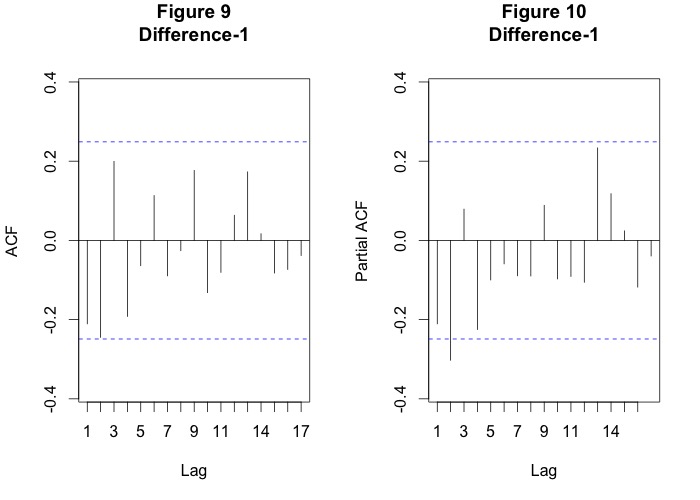


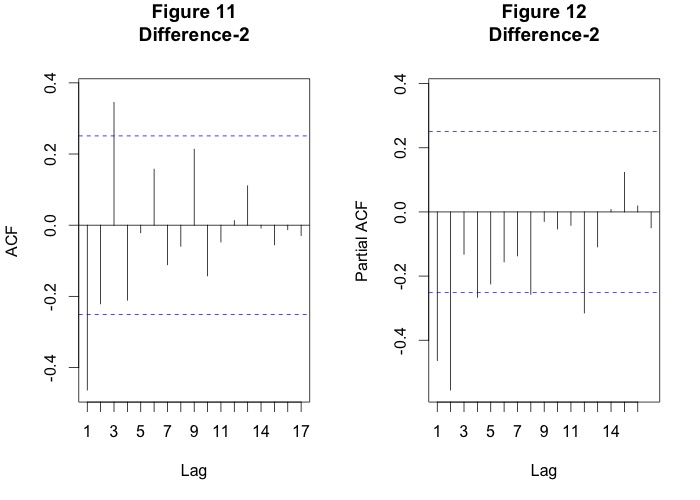


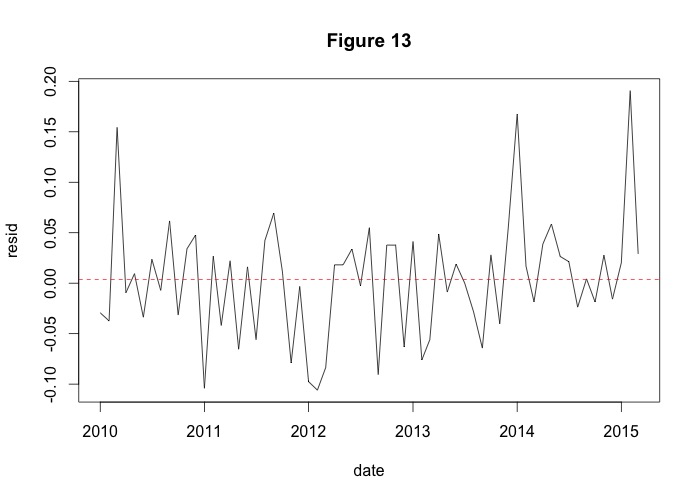


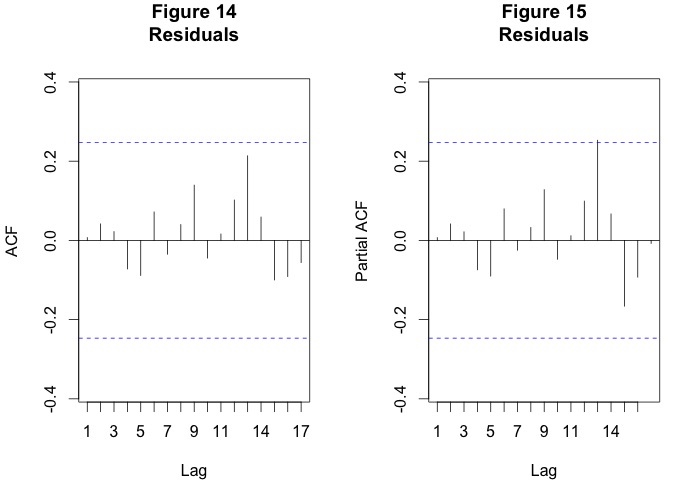


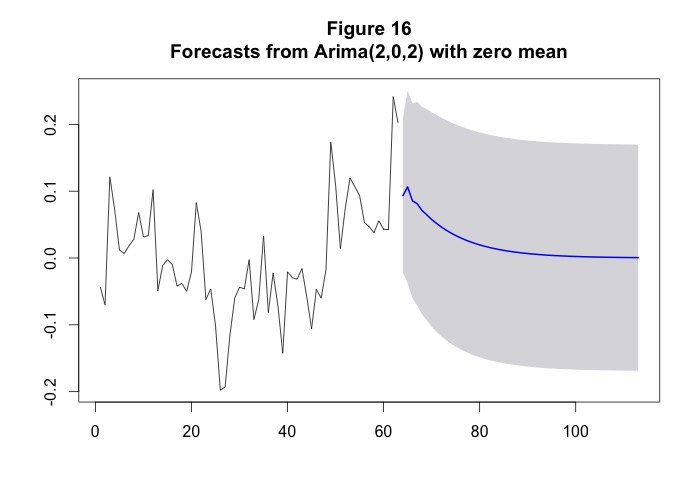












1. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9 [↑](#footnote-ref-1)