DengAl: Predicting Disease Spread

Prepared for

Dr. Paige Rutner

ISQS 6349

Prepared by:

Tahrima Mustafa

Rushan Shakya

Cam Moy

December 11, 2017

**Problem Statement**

Dengue fever or dengue is a virus spread through mosquito bites. This virus present itself in the form of fever, headaches, vomiting, nausea and in some cases it can lead to death. Since this illness is carried by mosquitos it is heavily dependent in tropical and subtropical climates. By forecasting the spread of this disease the health agencies can better organize their preventive measures such as vaccination and provide information to the public about this illness.

Our goal is to predict the total\_cases label for each (city, year, weekofyear) in the test set. There are two cities, San Juan and Iquitos, with test data for each city spanning 5 and 3 years respectively. We will make one submission that contains predictions for both cities. The data for each city have been concatenated along with a city column indicating the source: sj for San Juan and iq for Iquitos. The test set is a pure future hold-out, meaning the test data are sequential and non-overlapping with any of the training data. Throughout, missing values have been filled as NaNs.

**Literature Review**

The first article was, “Disease management with ARIMA model in time series” by Renato Cesar Sato.

This publication presents the reader with the analysis of infectious and noninfectious diseases in a time series through the autoregressive integrated moving average (ARIMA) model. It states that this mode is the preferred method by time series health researches due to its ability to forecast diseases it enables healthcare provider to prepare and manage their resources in the most efficient way.

The ARIMA model is defined as a mean to describe changes in a time series while trying to produce a model which values are as close as possible to the observed values. It identifies ARIMA’s methodology as having three phases, identification, estimation and testing, and application.

The identification phase can be described as the recognition of seasonal patterns, cycles, or any other patterns that are shown throughout the time series. It is suggested to use the autocorrelation function (ACF) to identify any type of pattern within the data set. Followed by the partial ACF (PACF) to measure the associative degree among observations.

The estimation and testing phase heavily depends on the values generated by the ARIMA model. The best model is often identified to be the one with the least sum of squared values and the least autocorrelation. It is also suggested to further evaluate your results with Protmanteau test; if this test returns a positive value the author encourages you to re-evaluate your model for further improvement.

For the last face Renato advices to ensure that your prediction model is not for more than 12 months. He argues that since this specific method is dealing with diseases without the proper historic data that have contributed to the spread of these illnesses a forecast greater than a year will be inaccurate.

He concludes by stating that even though ARIMA models have proved to be efficient tools for forecasting epidemics or the spread of a certain type of virus there is no replacement for the health providers experience.

The second article was, “Modeling and forecasting Australian domestic tourism”. (Athanasopoulos & Hyndman, 2008)

For this article, the authors developed three different statistical models for forecasting Australian domestic tourism. The authors first used a regression model to forecast tourism demand, this model just captured the static properties of the data. The second model the authors used was the Innovations state space model (time series model), which captured the dynamic properties of the data. The third model was the innovations state space model with exogenous variables to include both the static and dynamic properties of the data. The third model was the blend of the regression model and the time series model.

The regression model used 8 predictor variables which included some economic variables, some dummy variables, seasonal dummy variables and a random error term. The visual inspection of the dependent variable suggested them to use the stationarity tests. The authors used the augmented Dickey Fuller test, the KPSS test, and the Modified Phillips Perron test. Before applying the KPSS test, the data was seasonally adjusted using an additive moving average method. Given the quarterly frequency of the data- lag two, three, and four of each regressor were tested. The lags were found insignificant. The authors then removed the variables that were found to be statistically insignificant at 10% level of significance. The R2 values were taken to see the model fit. The authors simplified the models by eliminating insignificant variables one at a time, selecting the coefficient with the highest p-value among all insignificant coefficients at the 5% level of significance. Then, since the authors expected the errors across the equations to be correlated and because all equations include the same regressors, the system was estimated efficiently using the SUR estimation method.

Another model used by the authors was innovation state space model (innovation state space model encapsulates the notion of exponential smoothing in a state space framework). The optimal forecasts from innovations state space models are identical to those obtained using exponential smoothing methods. The models are estimated using the “forecast” R package. Akaike information criterion (AIC) and Bayesian Information Criterion(BIC) were used for the model selection criterion.

The third model used by the authors was the Innovations state space model with exogenous variables. The third model combines the previous two models, giving ETS models with exogenous variables. This model was estimated via a two-step procedure. First, the authors identified the exogenous variables to be included in the model. These were the variables that were found to be statistically significant for each equation in the SUR estimation results. Because the seasonal component in each of the time series models was deterministic, the authors modelled seasonality using seasonal dummies in the set of exogenous variables.

The authors then compared all 3 models. The Innovation state space model with exogenous variables was the best fitted model. The authors evaluated the accuracy measures using the Root mean squared error (RMSE), the mean error(ME), the mean absolute error(MAE) and the mean absolute percentage error (MAPE).

The third article we are going to discuss is “Time series forecasting using a hybrid ARIMA and neural network model by G. Peter Zhang.

Autoregressive integrated moving average (ARIMA) is one of the popular linear models in time series forecasting during the past three decades. Although ARIMA models are quite flexible in that they can represent several different types of time series, their major limitation is the pre-assumed linear form of the model. That is, a linear correlation structure is assumed among the time series values and therefore, no nonlinear patterns can be captured by the ARIMA model.

On the other hand, recent research activities in forecasting with artificial neural networks (ANNs) suggest that ANNs can be a promising alternative to the traditional linear methods. The major advantage of neural networks is their flexible nonlinear modeling capability. With ANNs, there is no need to specify a particular model form. Rather, the model is adaptively formed based on the features presented from the data. This data-driven approach is suitable for many empirical data sets where no theoretical guidance is available to suggest an appropriate data generating process.

ARIMA models and ANNs are often compared with mixed conclusions in terms of the superiority in forecasting performance. In this paper, a hybrid methodology that combines both ARIMA and ANN models is proposed to take advantage of the unique strength of ARIMA and ANN models in linear and nonlinear modeling.

The proposed methodology of the hybrid system consists of two steps. In the first step, an ARIMA model is used to analyze the linear part of the problem. In the second step, a neural network model is developed to model the residuals from the ARIMA model. Since the ARIMA model cannot capture the nonlinear structure of the data, the residuals of linear model will contain information about the nonlinearity. The results from the neural network can be used as predictions of the error terms for the ARIMA model. The hybrid model exploits the unique feature and strength of ARIMA model as well as ANN model in determining different patterns. Thus, the authors claim that it could be advantageous to model linear and nonlinear patterns separately by using different models and then combine the forecasts to improve the overall modeling and forecasting performance.

In this study, all ARIMA modeling is implemented via SAS/ETS system while neural network models are built using the GRG2-based training system mentioned earlier. Only the one-step-ahead forecasting is considered. The mean squared error (MSE) and mean absolute deviation (MAD) are selected to be the forecasting accuracy measures. Results show that for short-term forecasting (1 month), both neural network and hybrid models are much better in accuracy than the ARIMA- random walk model. For longer time horizons, the ANN model gives a comparable performance to the ARIMA model. The hybrid model outperforms both ARIMA and ANN models consistently across three different time horizons and with both error measures although the improvement for longer horizons is not very impressive.

Experimental results with real data sets indicate that the combined model can be an effective way to improve forecasting accuracy achieved by either of the models used separately.

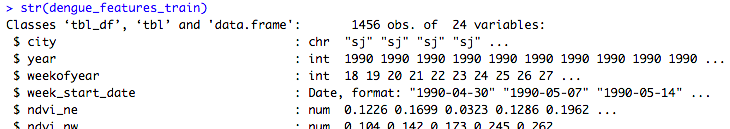
**Data Mining/Cleaning**

The Dengi data was collected from drivendata.org. The data includes environmental data from San Juan, Puerto Rico and Iquitos, Peru. The data was collected from various U.S. Federal Government agencies—from the Centers for Disease Control, Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce.

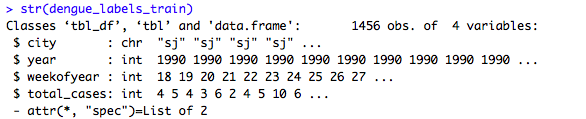
The data is provided in two different datasets the “dengue\_features\_train”, and the “dengue\_labels\_train”.



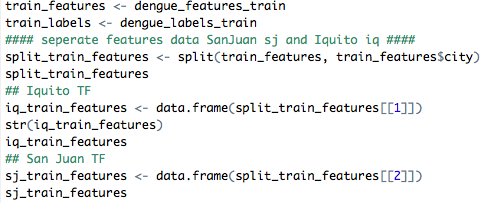
The first dataset contains daily climate information, precipitation, vegetation index and climate forecasts for both, San Juan and Iquito.

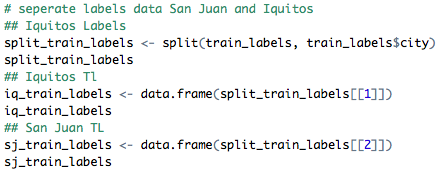


The second dataset contains four variables regarding the time of diagnosed cases of dengue fever for both cities.



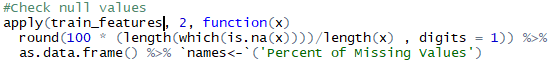
Both datasets contain information about two different cities it is necessary to divide them according to a single city in order to generate a dengue fever forecast per city. In other words it is necessary to divide “training\_features” into “sj\_trian\_features” and “iq\_train\_features” as well as the given data frame “training\_labels” into “sj\_train\_labels” and “iq\_train\_labels” to separate the values of each city; preparing the data for forecasting.

****

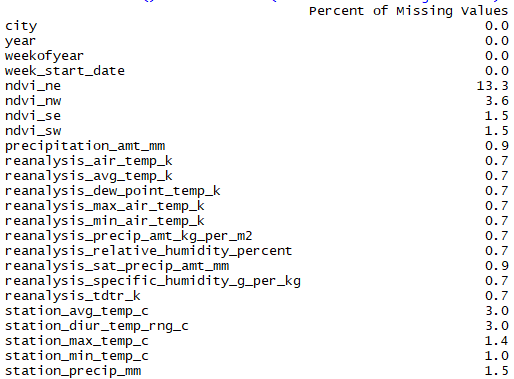
****

The next step is to check if any of the variables contain null values.

R-code:

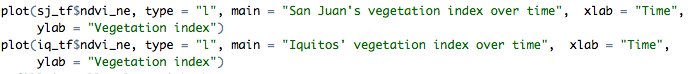


Output:

****

There are several values missing for each variable.

Climate factors directly influence the vegetation index of a location. Looking at the plots for the vegetation index for San Juan and Iquitos we can see that some of the values are missing from these time series.

****

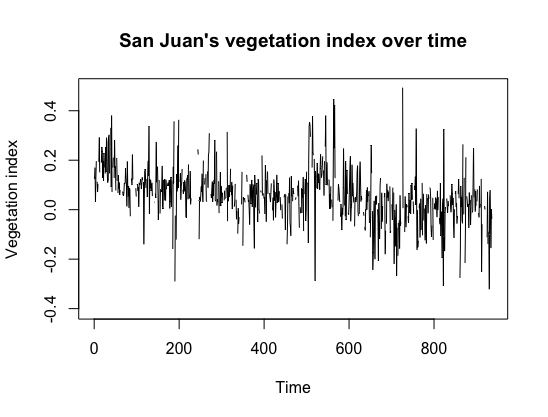
****

Figure 1. San juan’s vegetation index over time

There are missing values throughout the San Juan time series. The most noticeable gap of information is in the beginning of 200 and as the series gets close to the 900 mark.

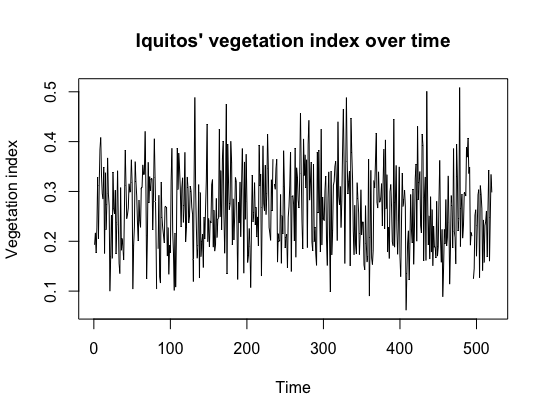
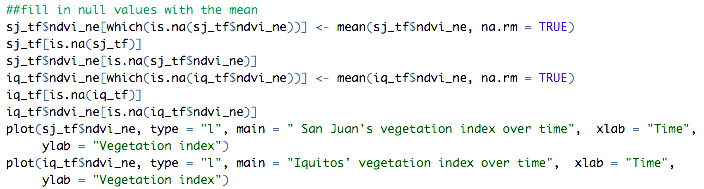


Figure 2. Iquitos’ vegetation index over time

In order to have a more complete dataset we have replaced the null values for the vegetation index (VI) with the average VI throughout the series (without considering null entries).



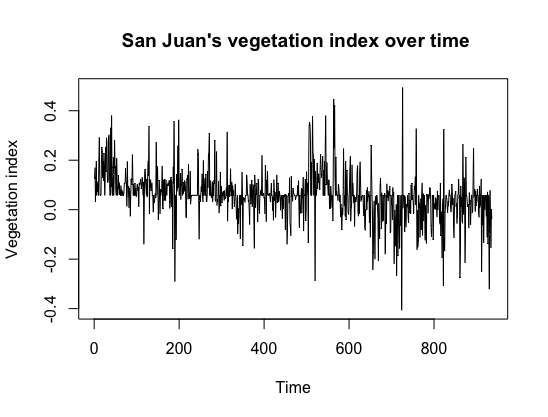


Figure 3. San Juan’s vegetation index over time

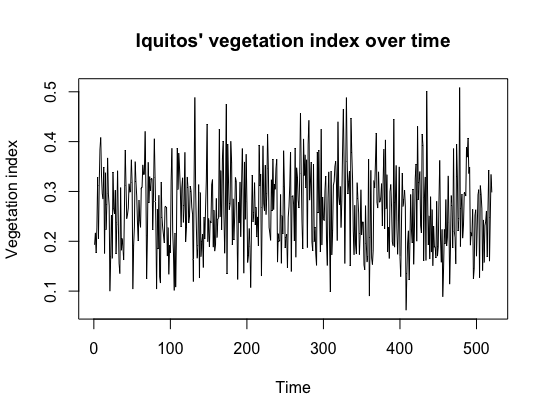
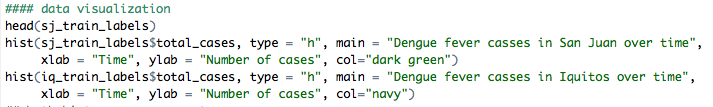


Figure 4. Iquitos’ vegetation index over time

The time series regarding the vegetation index for both cities are now continuous.

**Data Visualization**

Analysing the way the data behaves can give us an insight into what variables are needed for an accurate forecast as well as how what to expect from the forecast model. Visualizing the datasets “sj\_train\_labels” and “iq\_train\_labels” we can see the pattern of dengue fever cases over time.

****

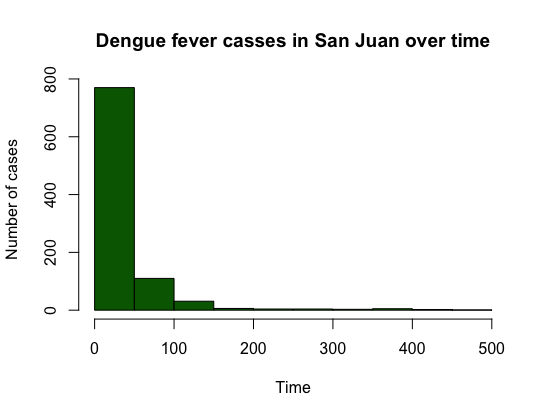


Figure 5. Dengue fever cases in San Juan over time

The data starts on the 1990s and continues on until 2008. Dengue fever cases have steadily declined since the 90s.

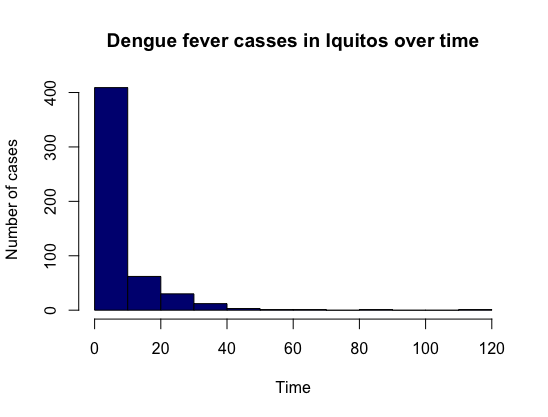
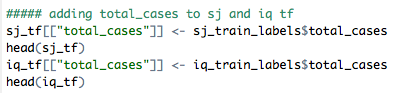


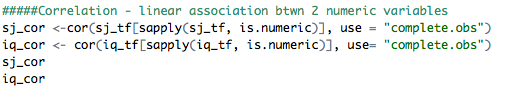
Figure 6. Dengue fever cases in Iquitos over time

For the city of Iquitos the time of the collected data ranges from 2000s to 2010. Iquitos also projects a decreasing pattern over time.

The next visualizations correspond to the datasets containing the features of San Juan and Iquitos with the addition of the column “total\_cases” from the labels datasets.

The addition of this variable will enable us to calculate its correlation with the features of each city. We look into the correlation between the variables because we want to understand the relationship between each of the variables with total\_cases variables.







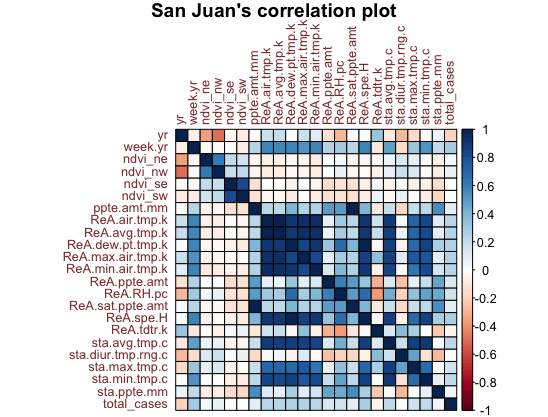
****

Figure 7.a. San Juan’s correlation plot

Looking into the correlation plot, we can see that the total\_cases at San Juan is showing positive correlation with most of our climate variable, mostly the temperature data.

****

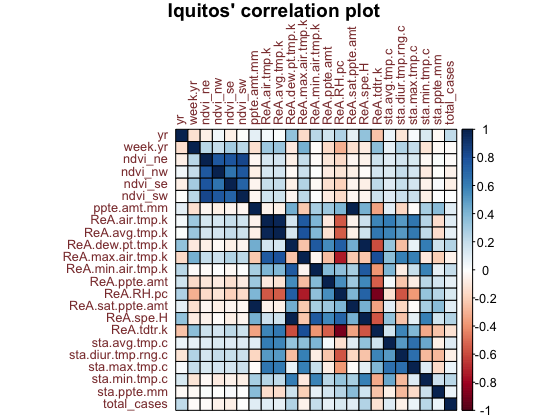
****

Figure 7.b. Iquitos’ correlation plot

Looking into our correlation plot for Iquitos city, we can still find positive correlation between our total\_cases variable and most of our climate variable, with strong correlations with our temperature data. Looking into the correlation between the various variables gives us better understanding of our model.

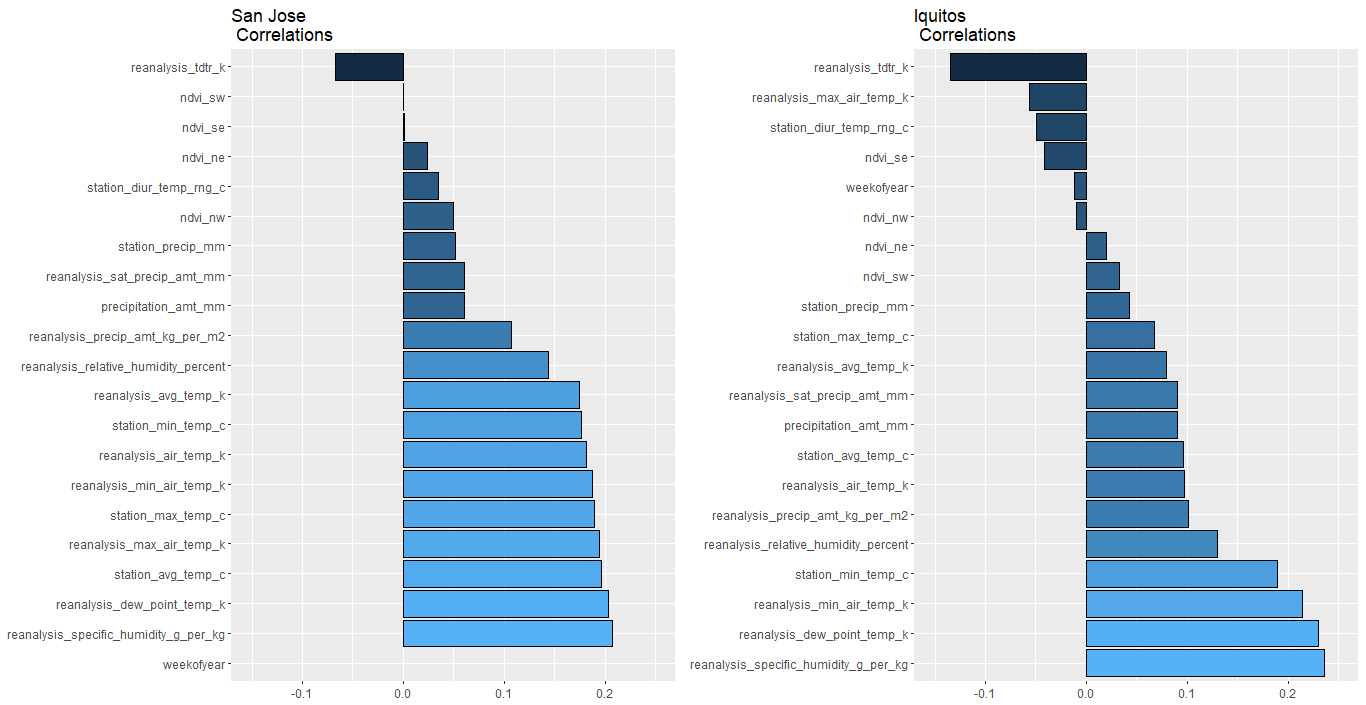


Figure 8. Correlations of various variables with Total\_cases variable in Barplot

From the figure 8, we can observe that reanalysis\_specific\_humidity\_g\_per\_kg and reanalysis\_dew\_point\_temp\_k are most strongly correlated with Total\_cases variable.

Also, relative\_humidity\_percent, station\_min\_temp\_c, reanalysis\_min\_air\_temp\_k, and station\_average\_temp\_c has positive correlation with total\_cases. Also, reanalysis\_tdtrk\_k shows the some negative correlation with total\_cases variable.

Figure 9.a.(below), shows the Boxplot of our dependent and independent variables at Iquitos city. Starting from the left of our figure, we have total\_cases, reanalysis\_specific\_humidity\_g\_per\_kg, station\_max\_temp\_c, reanalysis\_tdtr\_k, reanalysis\_dew\_point\_temp\_k and reanalysis\_max\_air\_temp\_k variables. We can see that total\_cases is right skewed. Reanalysis\_specific\_humidity\_g\_per\_kg looks slightly left skewed. Station\_max\_temp\_c looks like it is normally distributed. Reanalysis\_tdtr\_k looks like it is left skewed. Reanalysis\_dew\_point\_temp\_k looks like it is right skewed and reanalysis\_max\_air\_temp\_k variable looks like it is normally distributed.

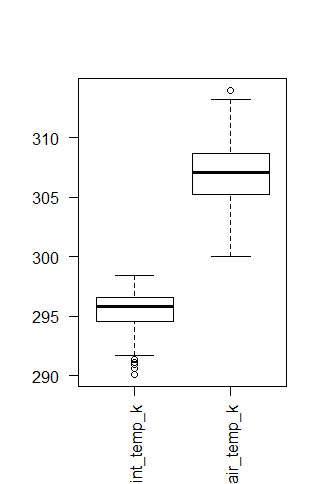
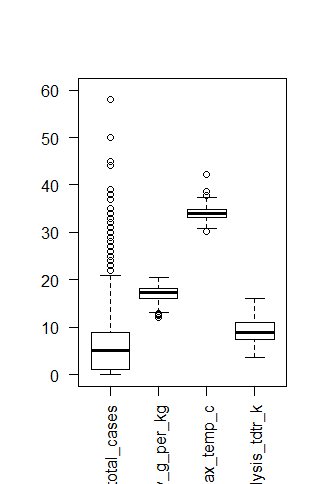


Figure 9.a. Box-plot for Iquitos

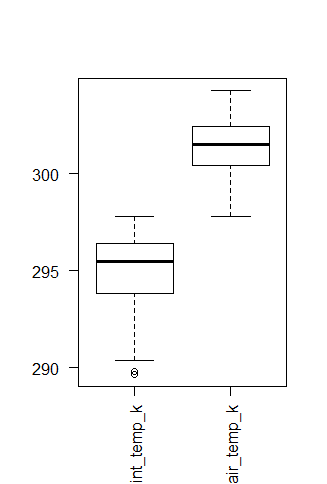
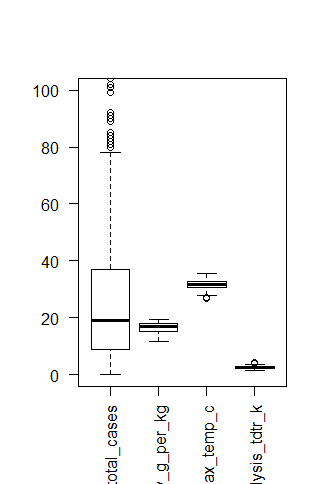


Figure 9.b. Box-plot for San Juan

From figure 9.b., we can see that total\_cases is right skewed. Reanalysis\_specific\_humidity\_g\_per\_kg, Station\_max\_temp\_c, and Reanalysis\_tdtr\_k looks like it is normally distributed. Reanalysis\_dew\_point\_temp\_k looks like it is right skewed and reanalysis\_max\_air\_temp\_k variable looks like it is normally distributed.

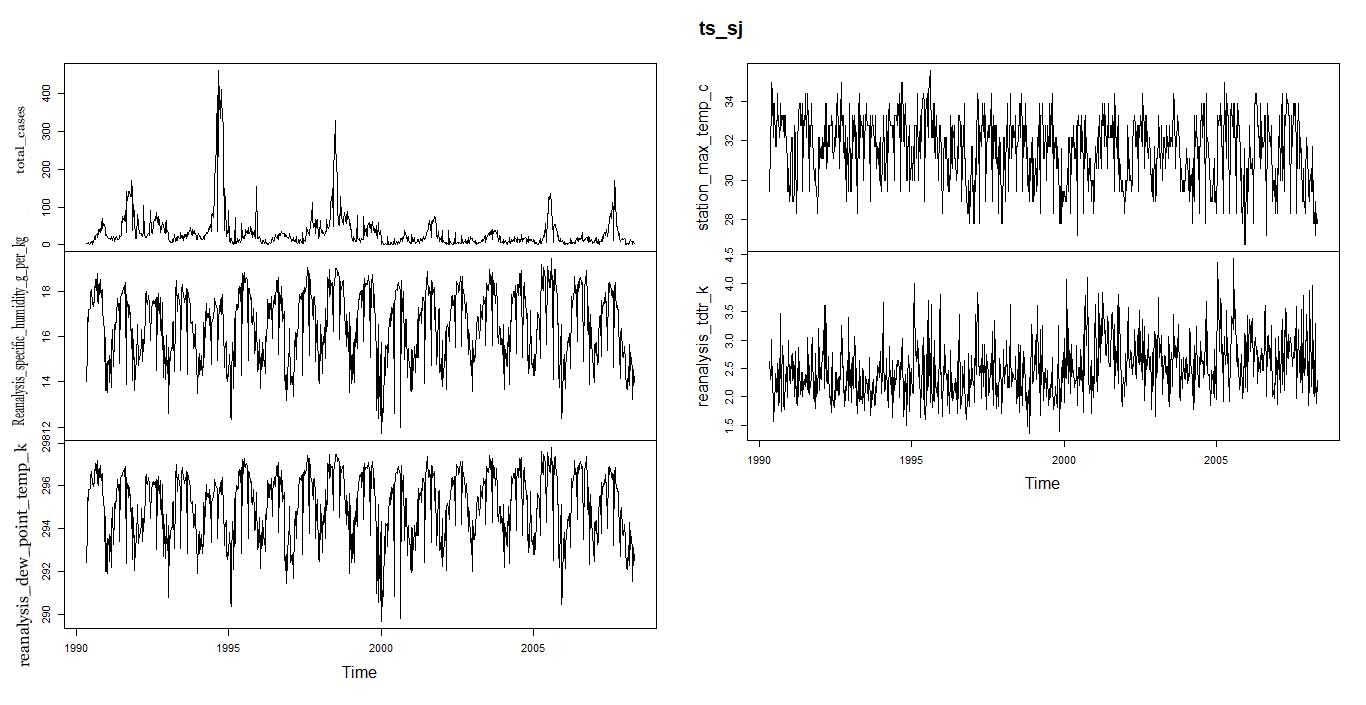


Figure 10.a. Time plot for San Juan variables: total\_cases, reanalysis\_specific\_humidity\_g\_per\_kg, reanalysis\_dew\_point\_temp\_k, station\_max\_temp\_c and reanalysis\_tdtr\_k

From figure 10.a., we can observe that all our variables have strong seasonality. We cannot see any trend in our data. We can also see that all our variables except total\_cases look stationary.

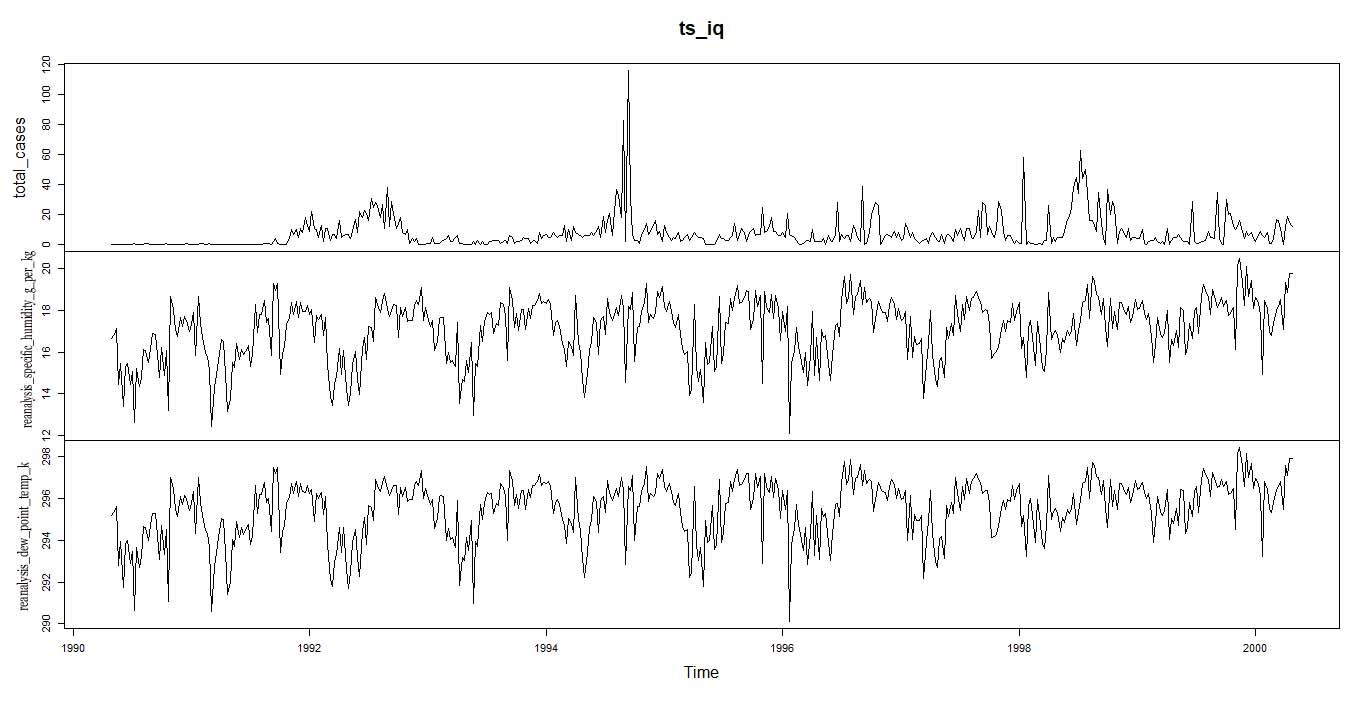


Figure 10.b. Time series for Iquitos variables: total\_cases, reanalysis\_specific\_humidity\_g\_per\_kg, and reanalysis\_dew\_point\_temp\_k

From figure 10.b., we can see some strong seasonality in our variables. We cannot see any trend in our data set. We can also see that all our variables except total\_cases look stationary.

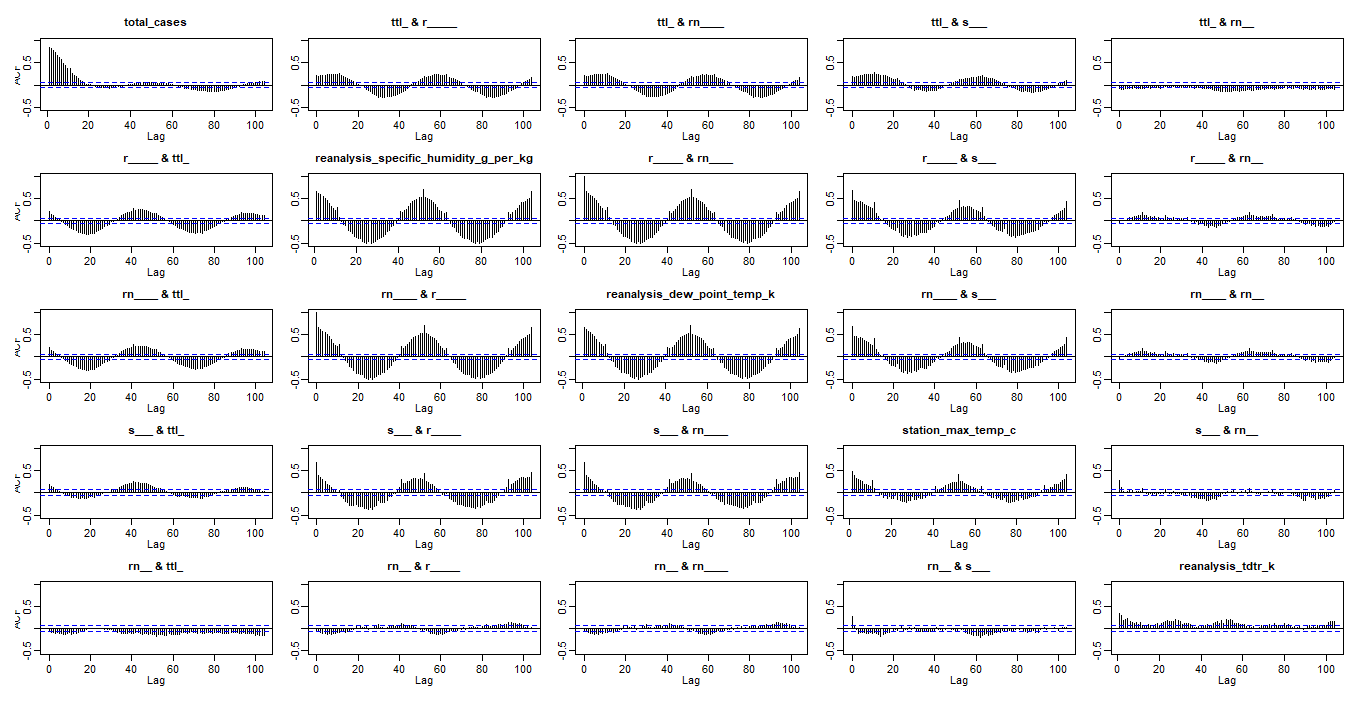


Figure 11.a. ACF plot for San Juan Dengi variables

From the above figure, we can see that there is positive correlation between the lags of each of our predictor variables. We might have to take some transformations with the data when we use these predictors in our time series models. We can also see some seasonality within the variables.

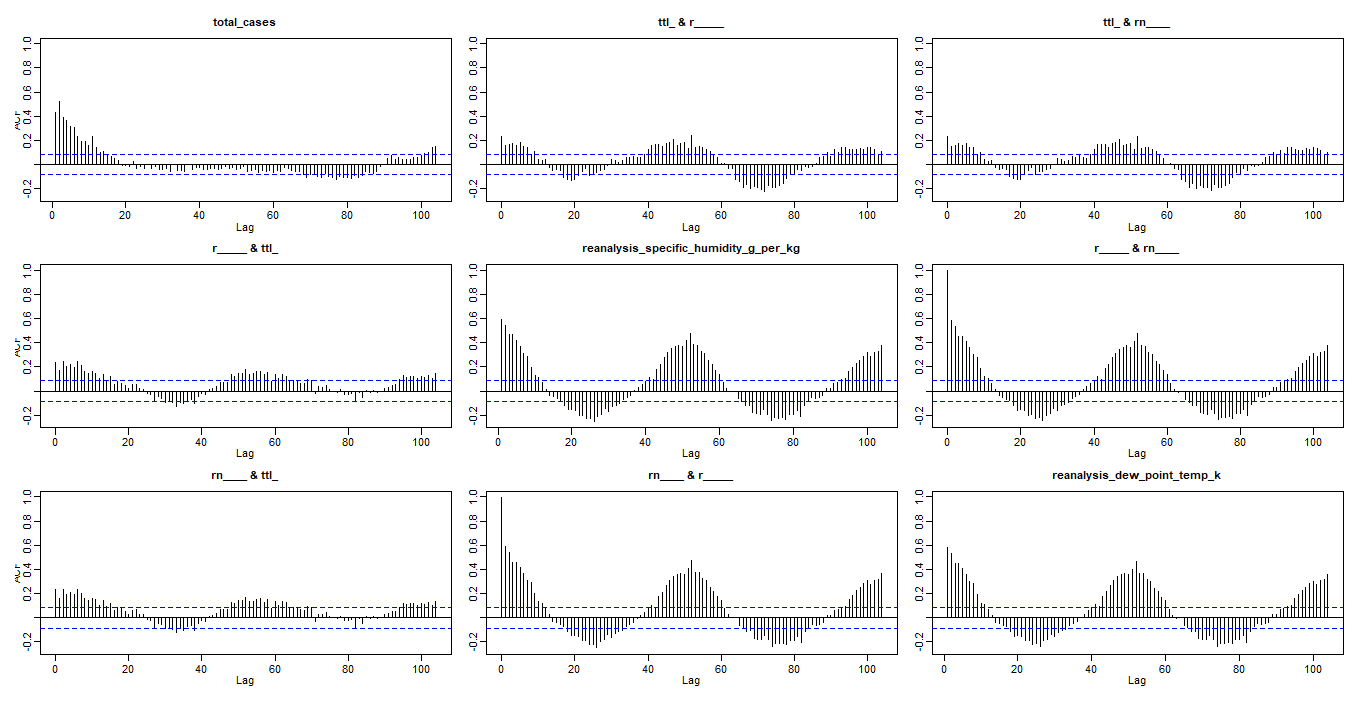


Figure 11.b. ACF plot for Iquitos Dengi variables

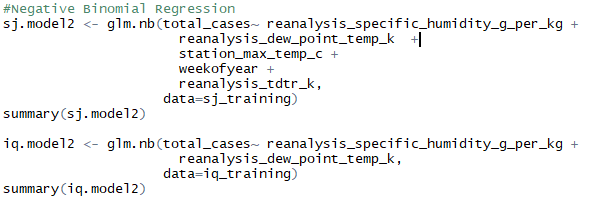
From the above figure, we can see that there is positive autocorrelation within the lags of each of our predictor variables. We can also see some seasonality in the plots for all our predictor variables.

**Predictive Analysis**

1. **Negative Binomial Regression**

We used the Negative Binomial Regression from the MASS package.

R-code:

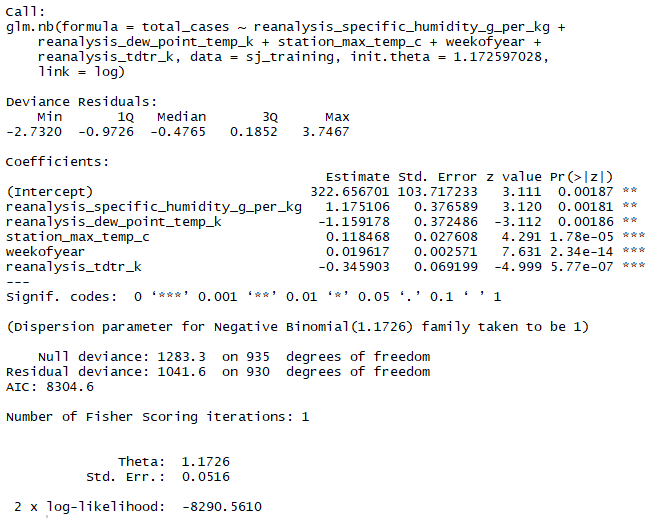


We found that, for the Negative Binomial Regression, only reanalysis\_specific\_humidity\_g\_oer\_kg and reanalysis\_dew\_point\_temp\_k independent variables were significant variables for both our cities. Station\_max\_temp\_c and weekofyear, and reanalysis\_tdtr\_k independent variables were only significant for San Juan.

We started with twelve variables for each of our two cities to fit a negative binomial regression model. But we ended up with just five independent variables for San Juan and 2 independent variable for Iquitos.

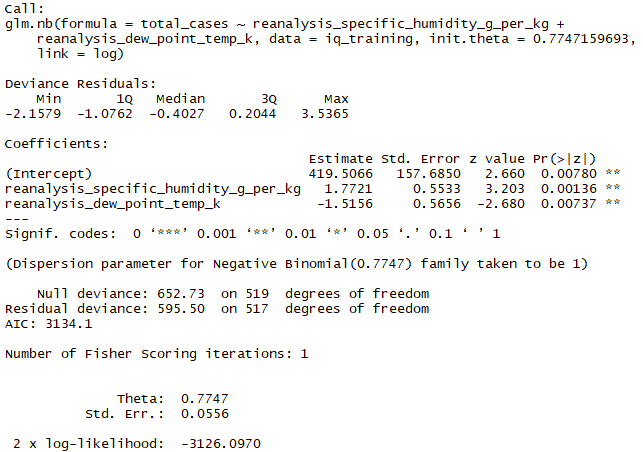
For San Juan, our AIC with the model came up to be 8304.6 and our model had an overall significance level of about 2%.

Output:



Next, we fitted our model for Iquitos. We only had 2 independent variable that were significant to our model. Our model had a AIC score of 3134.1 and our over all significance level for our model was 8%.

Output:



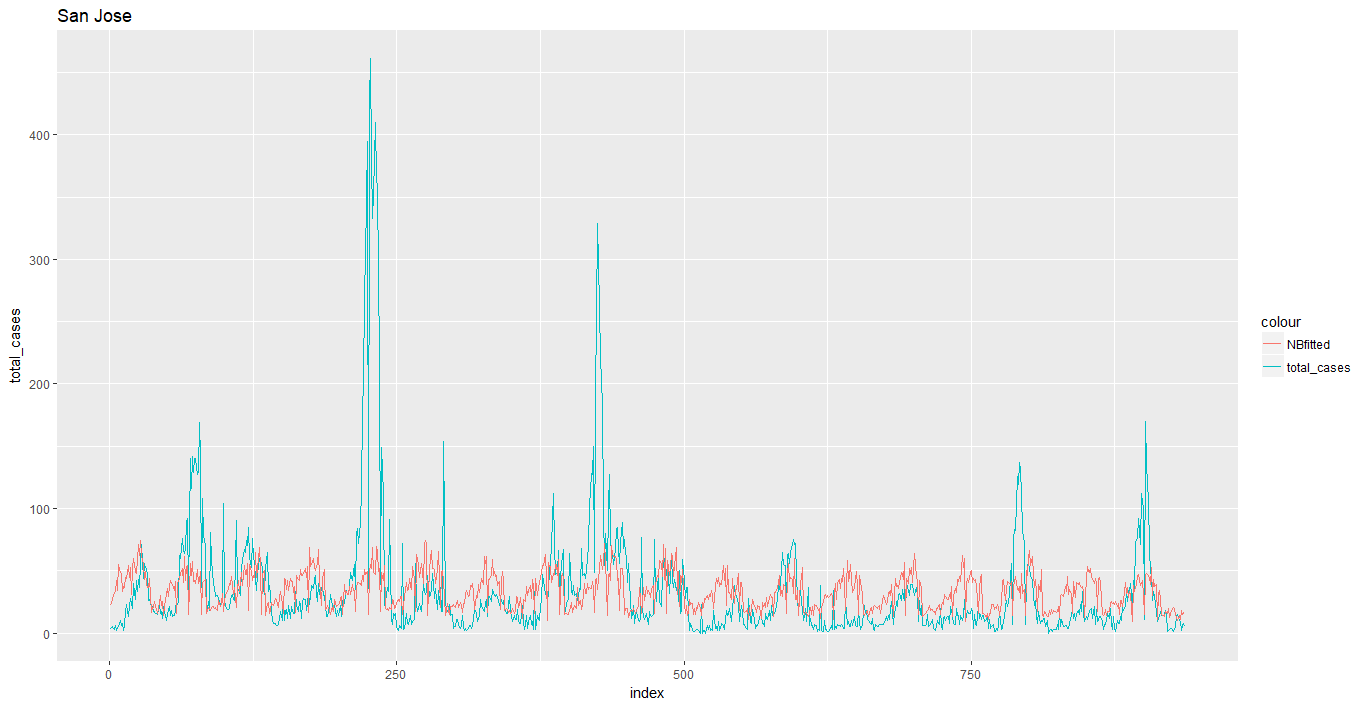


Figure 12.a. Plot of Actual vs Predicted values (Negative Binomial Regression)

From figure 12.a., we can see that our predicted values do not perfectly match our actual values. Even Though, our model does predict the seasonality of data with respect to the actual data, we can see that in some cases the predicted values exceeded the actual values most of the time. There is still room for improvement on our model for San Juan.

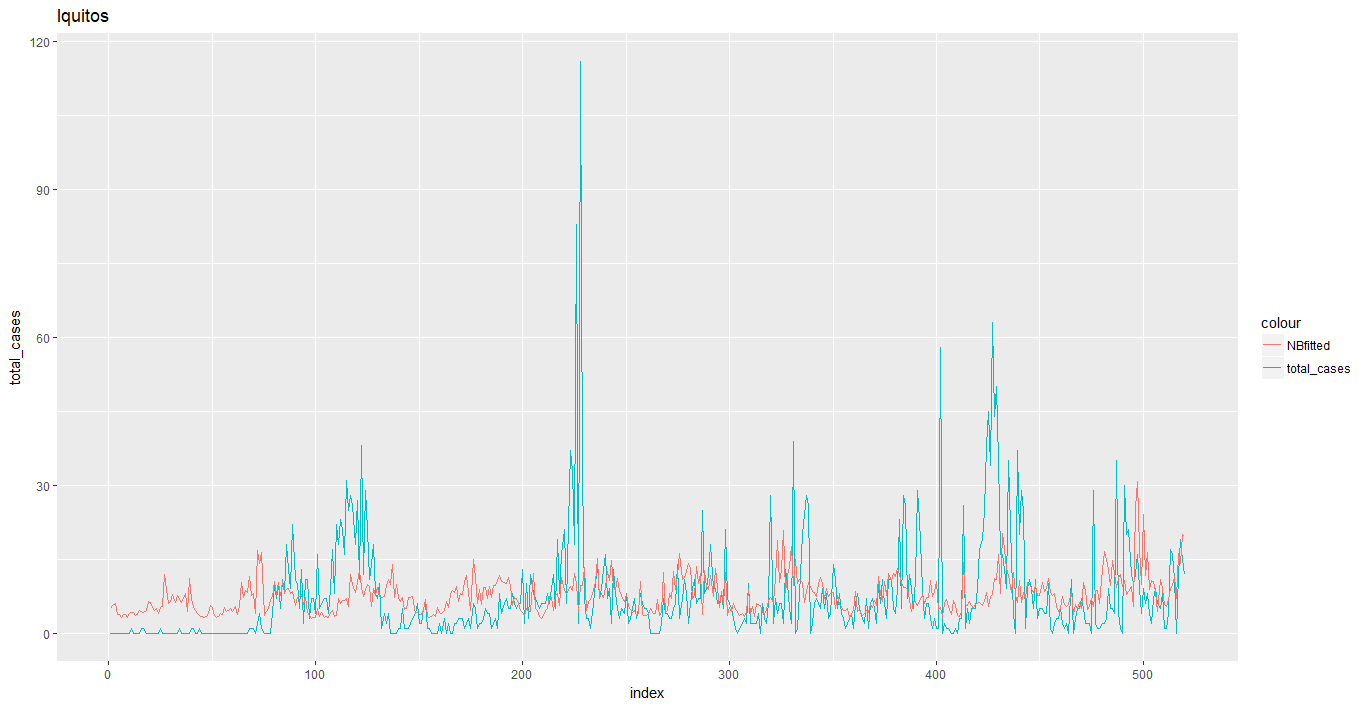


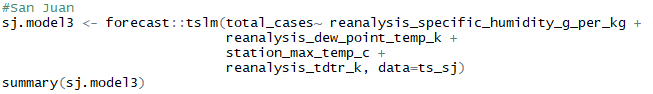
Figure 12.b. Plot of Actual vs Predicted values (Negative Binomial Regression)

From figure 12.b., we can see that our predicted values do not perfectly match our actual values. Even Though, our model does predict the seasonality of data with respect to the actual data, we can see that in some cases the predicted values exceeded the actual values most of the time. There is still room for improvement on our model for Iquitos.

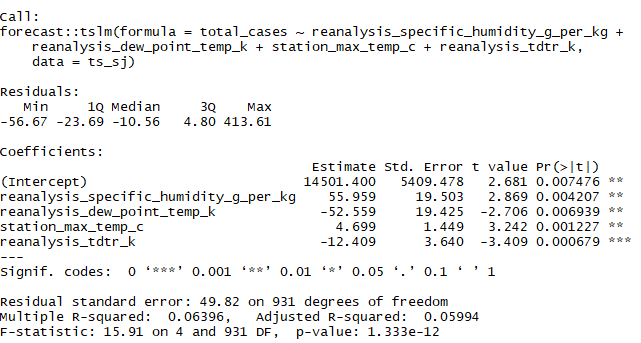
**2. Linear Regression with Time Series Data**

For Linear Regression with Times Series Data for San Juan, we took reanalysis\_specific\_humidity\_g\_per\_kg, reanalysis\_dew\_point\_temp\_k, station\_max\_temp\_c and reanalysis\_tdtr\_k as our independent predictor variables based on the predictor variables for San Juan from our Negative Binomial Regression model.

R-code:



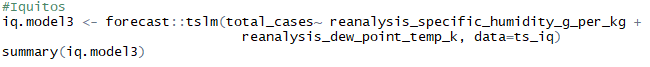
Output:



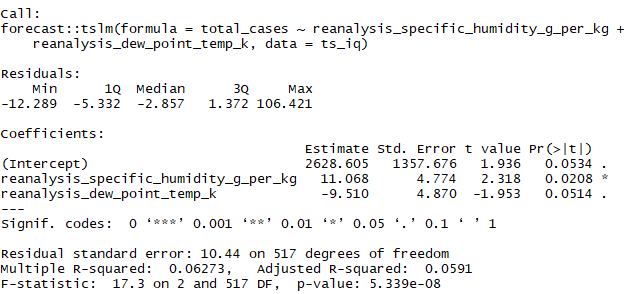
We can see that our Adjusted R^2 for our model is 0.05994. Our F-statistics looks significant.

For Iquitos data, we took reanalysis\_specific\_humidity\_g\_per\_kg and reanalysis\_dew\_point\_temp\_k to predict total\_cases, based on our results from our Negative Binomial Regression model.

R-code:



Output:



We can see from our output, that the R^2 value is 0.0591 and a very significant F-statistics. Since, our R^2 value is only about 6% for both San Juan and Iquitos, our Linear Regression with Time Series Data model is not an appropriate model for making our prediction.

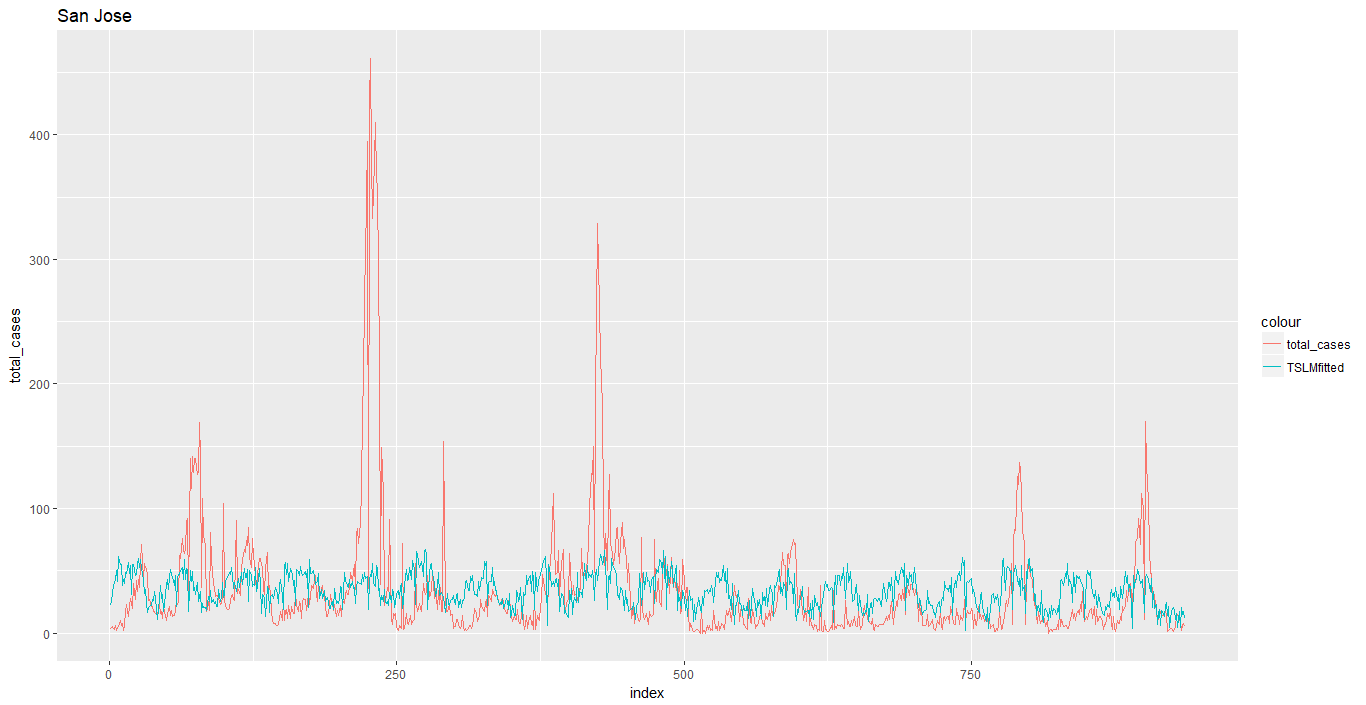


Figure 13.a. Time Series Actual vs Predicted (San Juan)

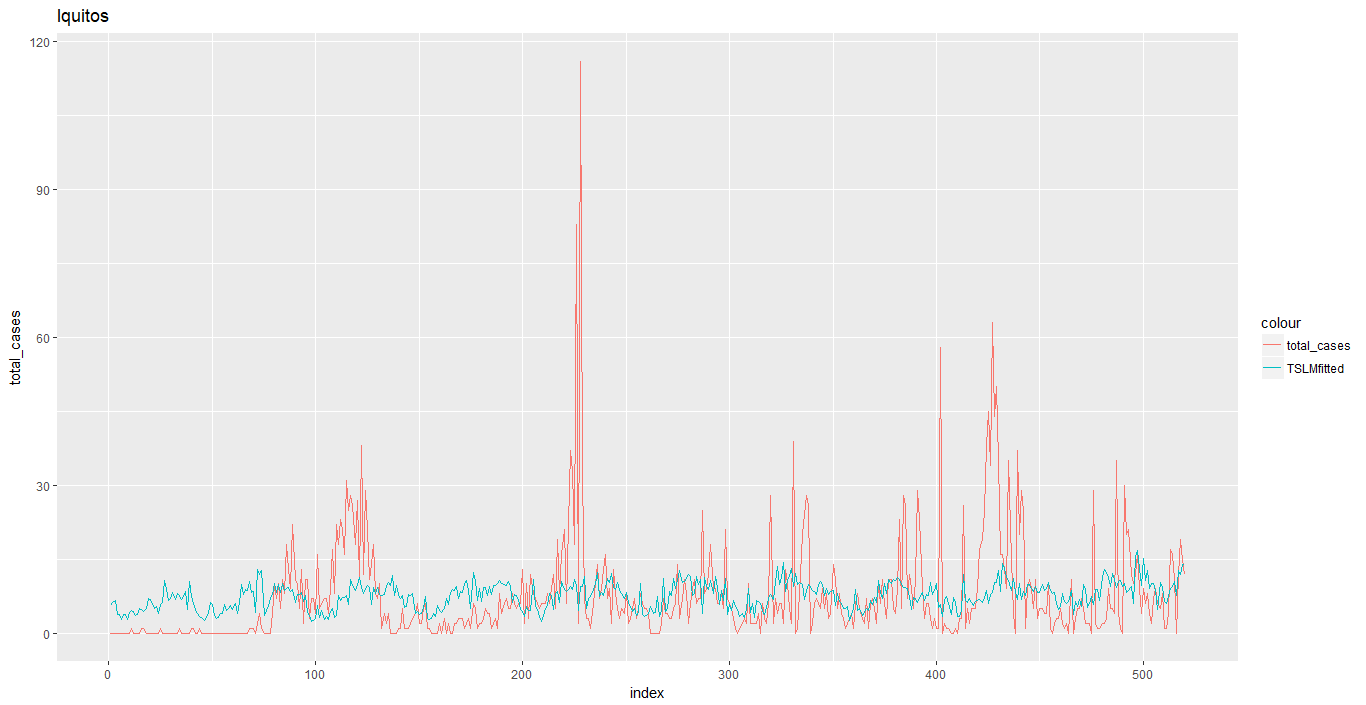


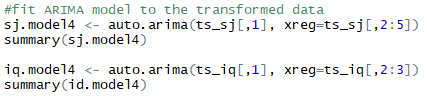
Figure 13.b. Time Series actual vs predicted (Iquitos)

From Figure 13, we can see that our predicted values were not able to truly capture the variation with our actual data movements. Our predicted values did capture the seasonality present in the actual data. There are still some room for improving our models.

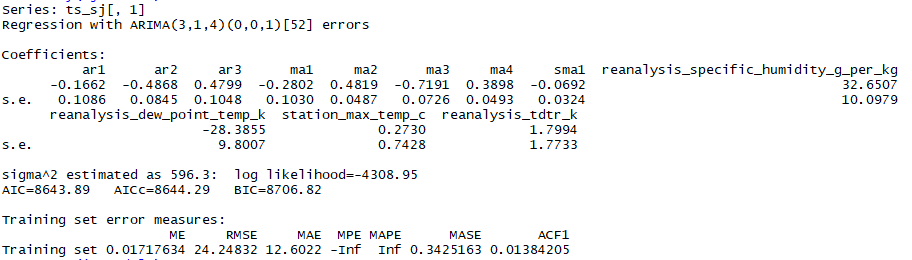
**3. ARIMA**

As discussed in [1], [3] we have used ARIMA with exogenous variables (Dynamic Regression), for San Juan we use reanalysis\_specific\_humidity\_g\_per\_kg, reanalysis\_dew\_point\_temp\_k, station\_max\_temp\_c and reanalysis\_tdtr\_k as our independent predictor variables. For Iquitos, we use reanalysis\_specific\_humidity\_g\_per\_kg and reanalysis\_dew\_point\_temp\_k as our independent predictor variable. We use total\_cases as our dependent variable for both San Juan and Iquitos.

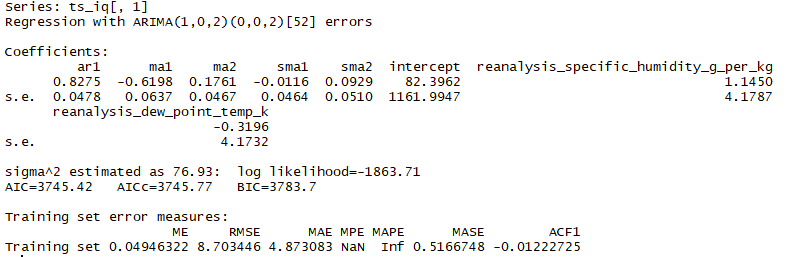
R-code:



Output for San Juan:



From the output above, we can see that auto.arima() applied ARIMA(3,1,4)(0,0,1) to our San Juan data. Our RMSE came to be 24.25, AIC came to be 8643.89 and AICc came to be 8644.29.

Output for Iquitos:  


From our output above, we can see that auto.arima() applied ARIMA(1,0,2)(0,0,2) to our Iquitos data. Our RMSE came to be 8.07, AIC came to be 3745.42 and AICc came to be 3745.77.

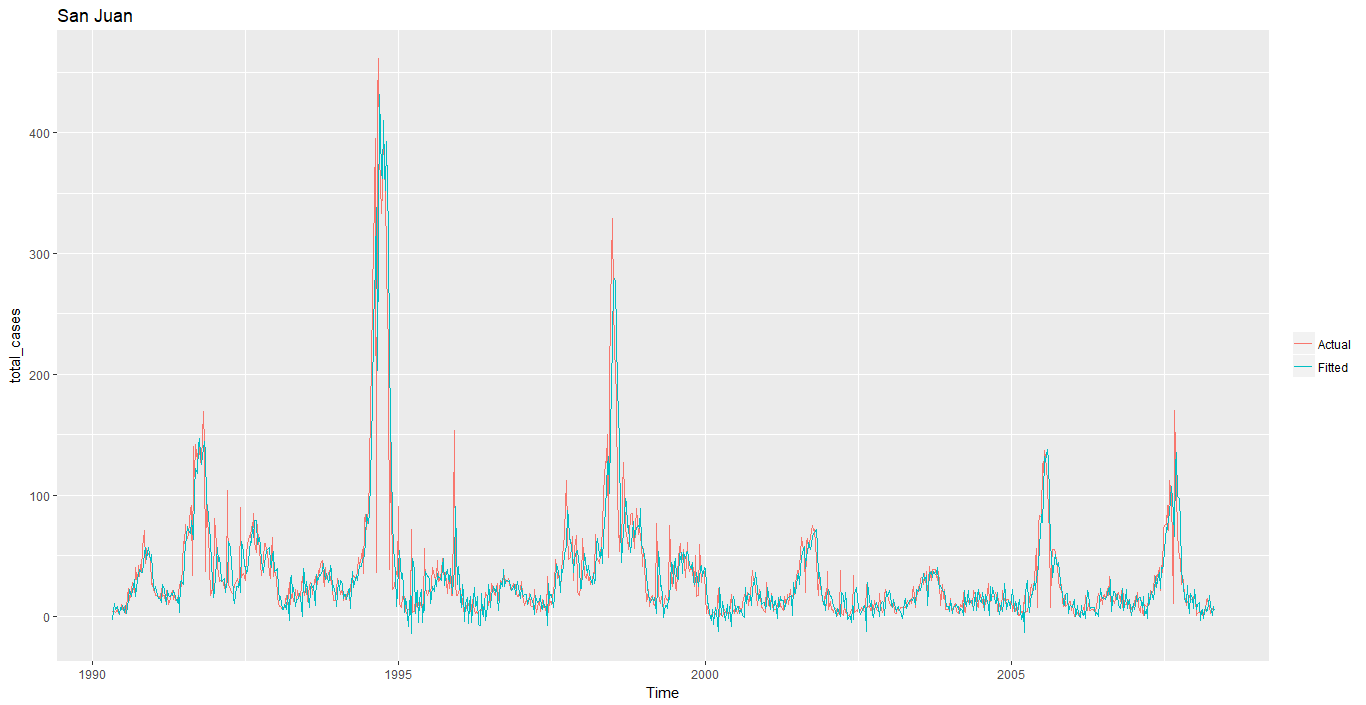


Figure 14.a. Fitted vs actual values for San Juan

From the above figure, we can see that our fitted value almost perfectly fit well with our actual values for San Juan.

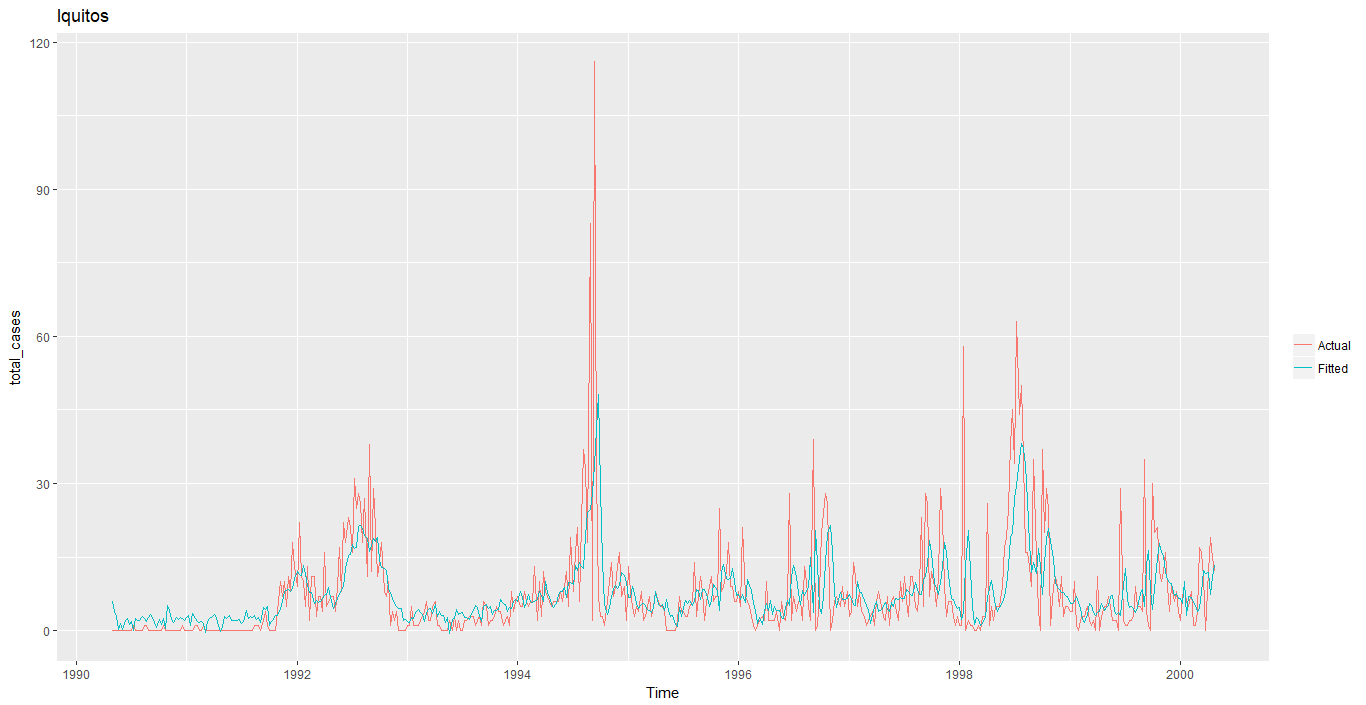


Figure 14.b. Fitted vs actual for Iquitos

From the above figure we can see that our fitted values capture the actual values of our model.

**Discussion and Recommendation**

1. **Residual Diagnostics for Negative Binomial Regression Model**

San Juan

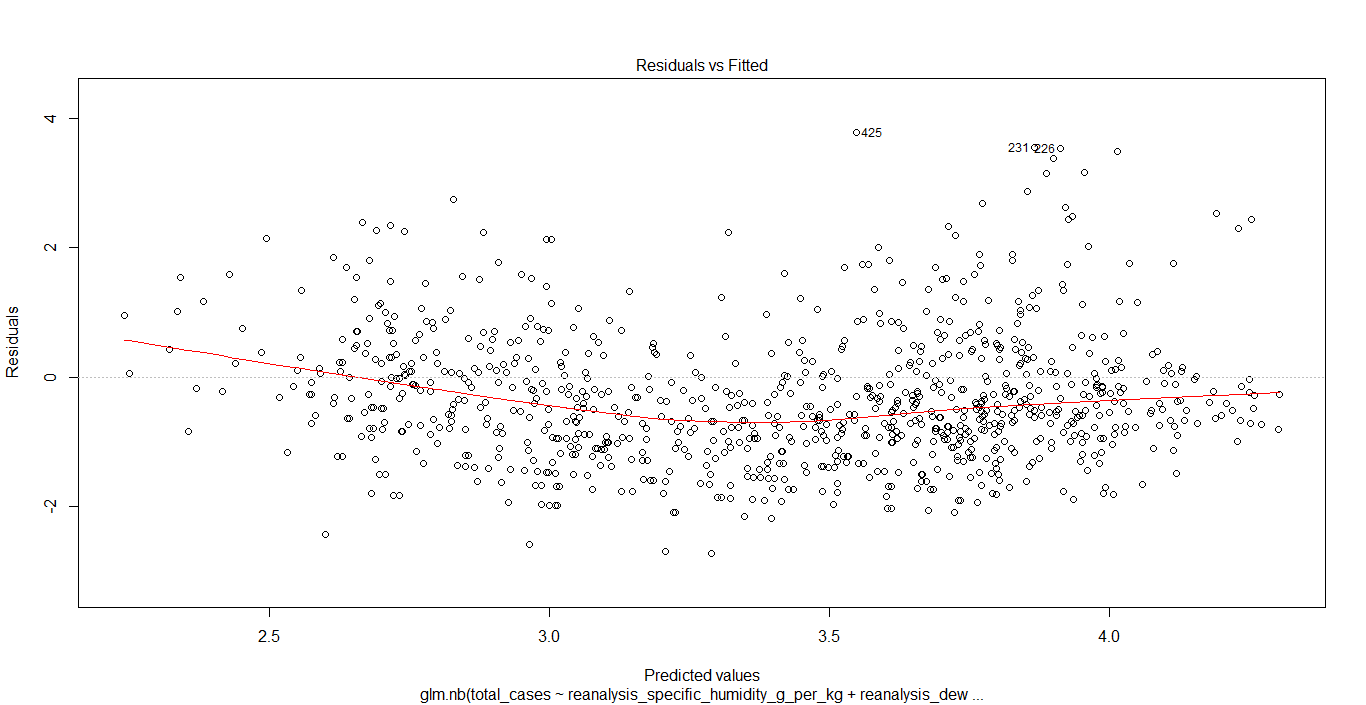


Figure 15.a. Residuals vs Fitted values for San Juan

Residual vs Fitted Values plot show that there is some observable pattern, i.e. we can see some heteroscedasticity in our residuals. The variance of the residuals is not constant.

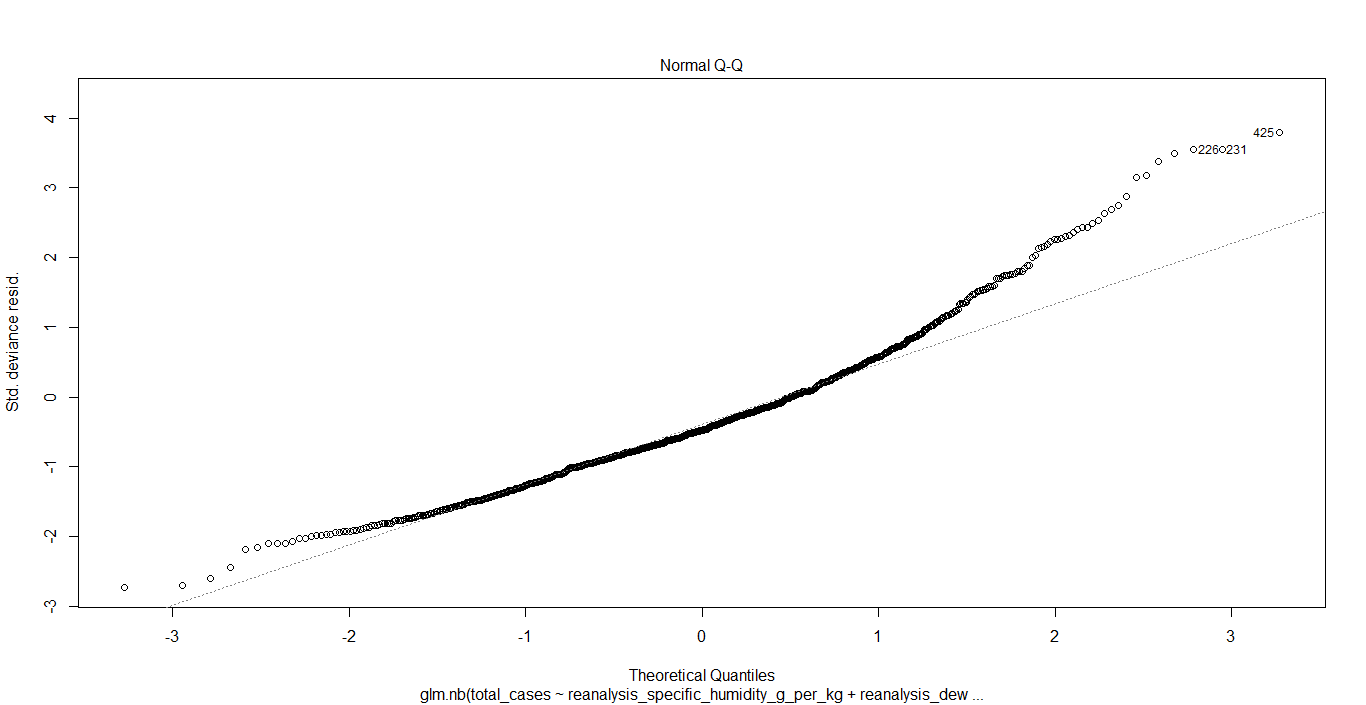


Figure 15.b.. Q-q plot for San Juan

The Q-Q plot shows that we can see that the residuals data are not normally distributed, and we may have an outlier in our data.

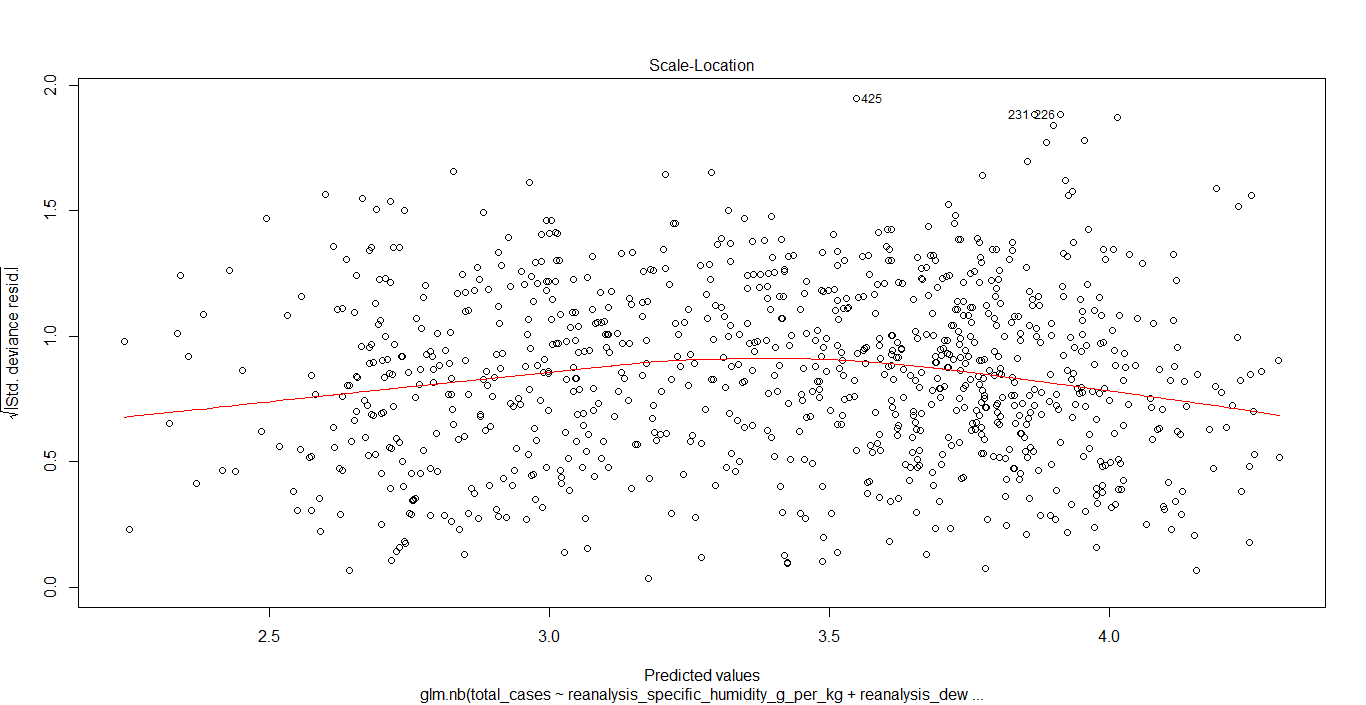


Figure 15.c. Predicted values vs Standard Residuals

Figure 15.c. is quite similar to Figure 15.a. There are some patterns in this graph, suggesting heteroscedasticity in our residuals.

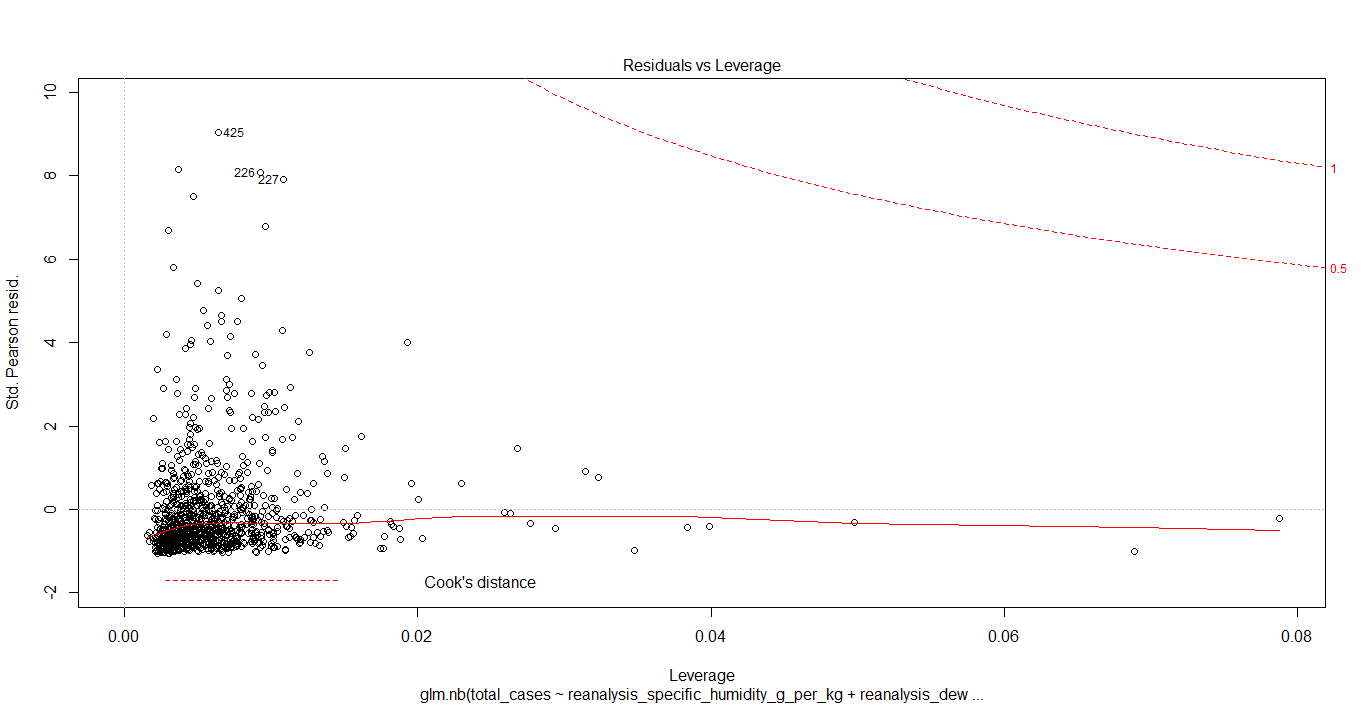


Figure 15.d. Leverage plot for San Juan

We can see that all the residuals are within the Cook’s distance.

Iquitos

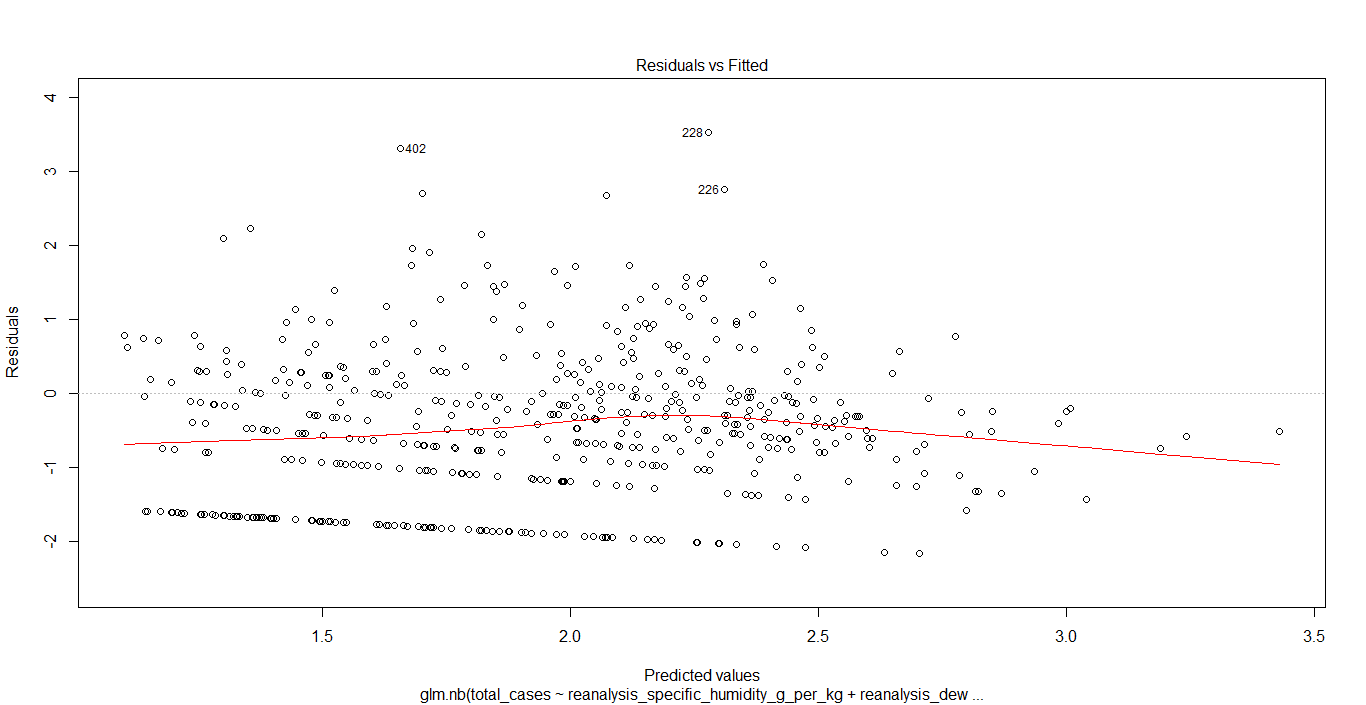


Figure 16.a. Residuals vs fitted values for Iquitos

Residual vs Fitted Values plot show that there is some observable pattern, i.e. we can see some heteroscedasticity in our residuals. The variance of the residuals is not constant.

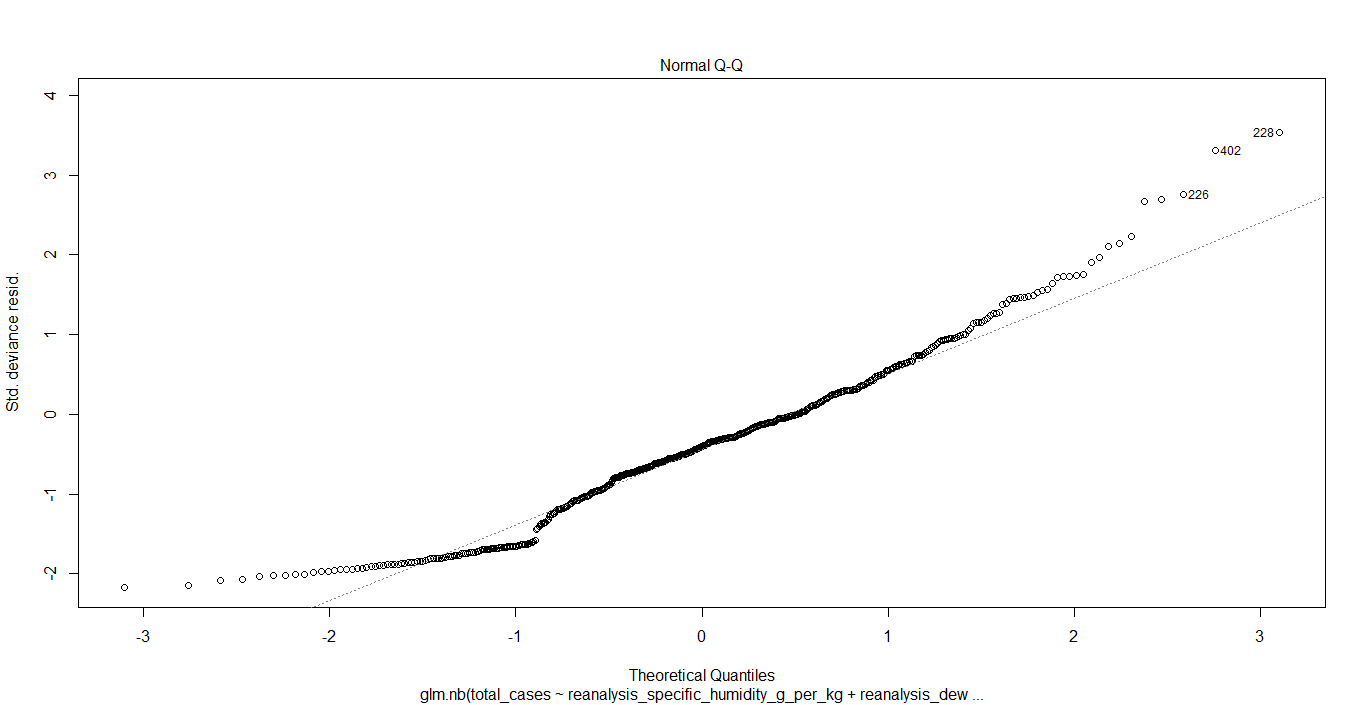


Figure 16.b. Q-q plot for Iquitos

From the q-q plot we can see that our residual data is not normally distributed.

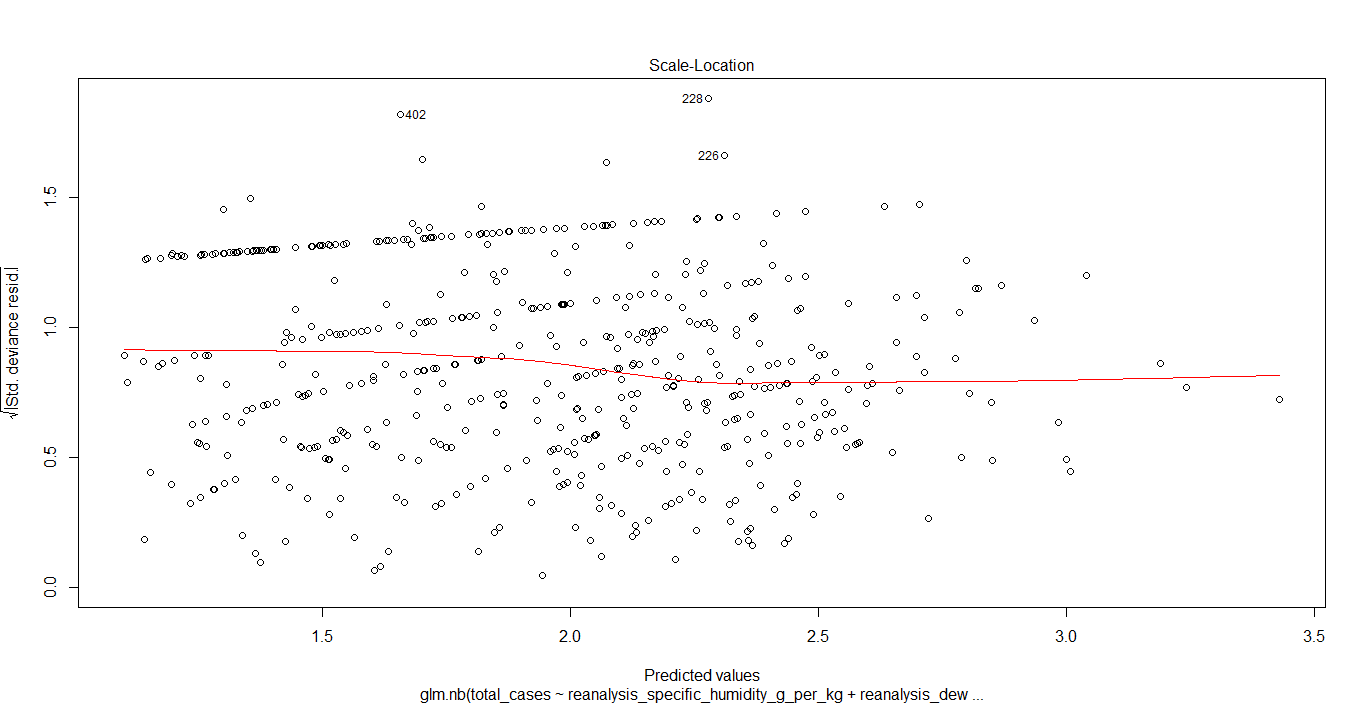


Figure 16.c. Predicted values vs Standard Residuals for Iquitos

This plot is similar to Figure 16.b. It shows that there are some patterns in this graph, suggesting heteroscedasticity in our residuals.

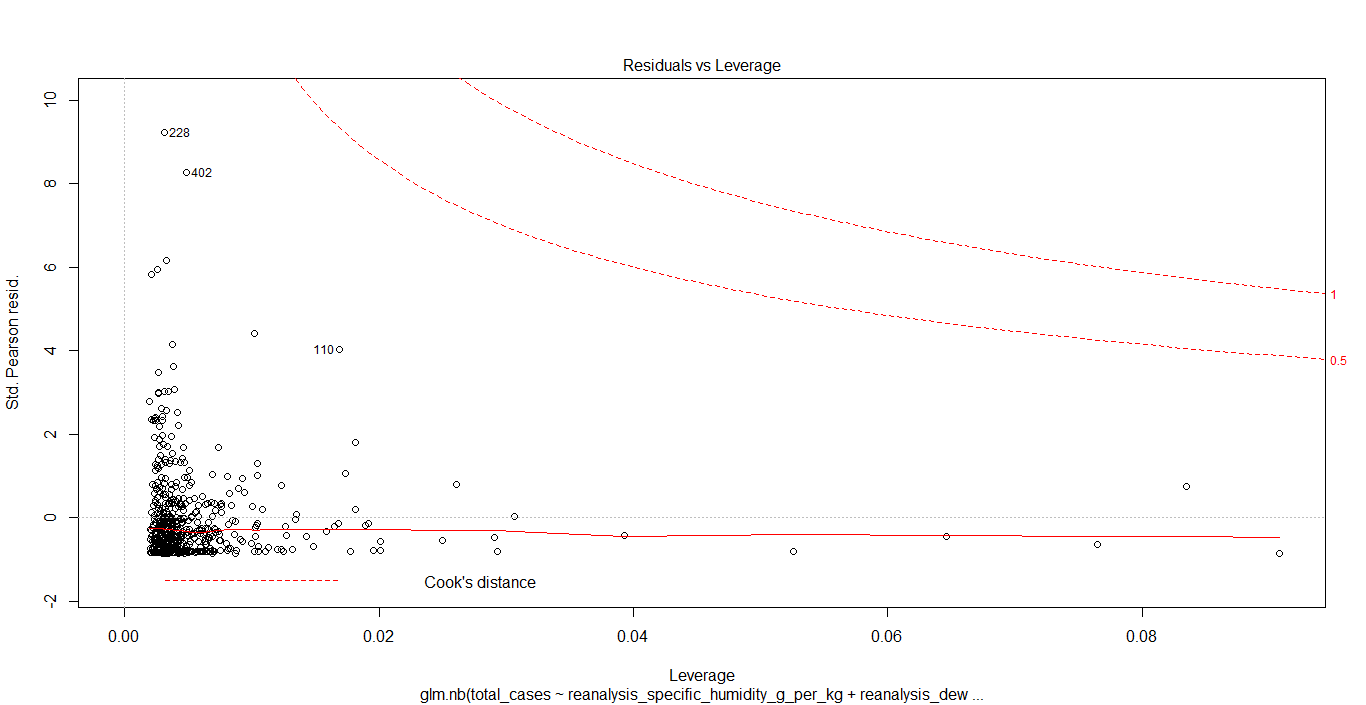


Figure 16.d. Leverage plot for Iquitos

For this graph, we can see that all our residuals are within the Cook’s distance.

**2. Residual diagnostics for Linear Regression with Time Series Data**

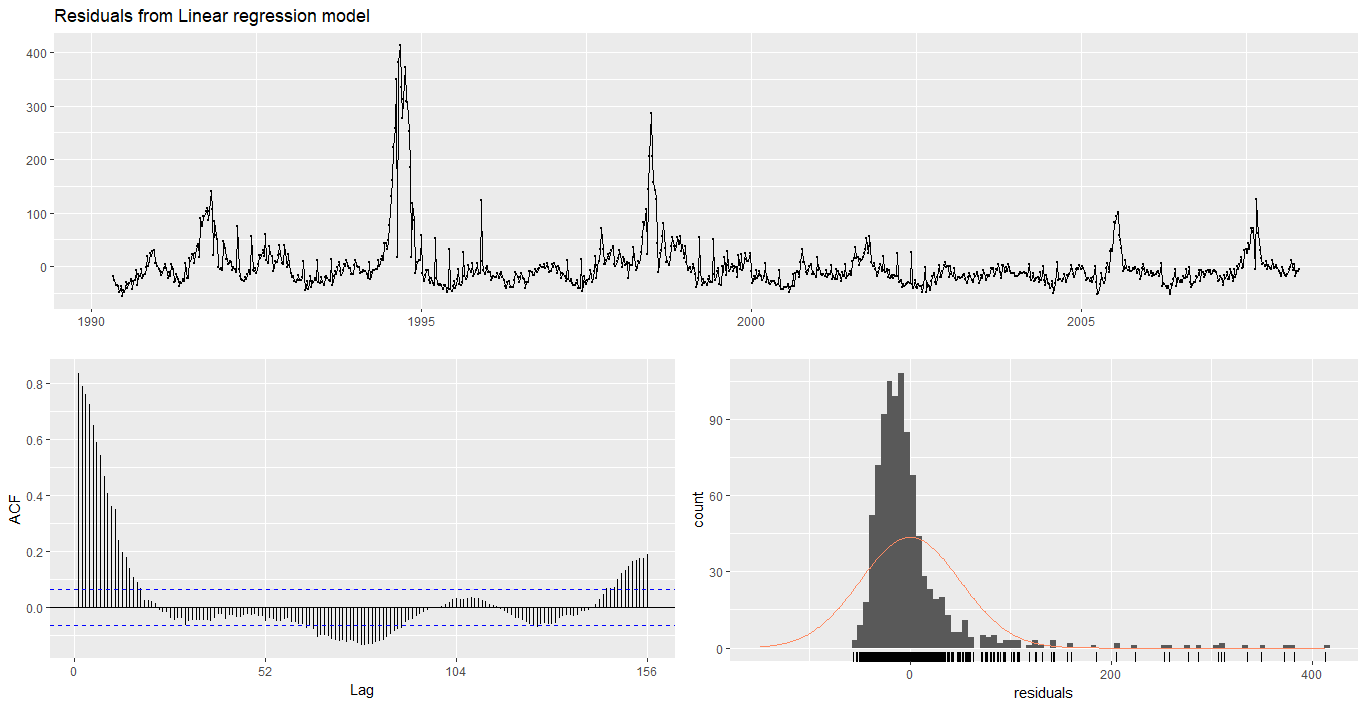
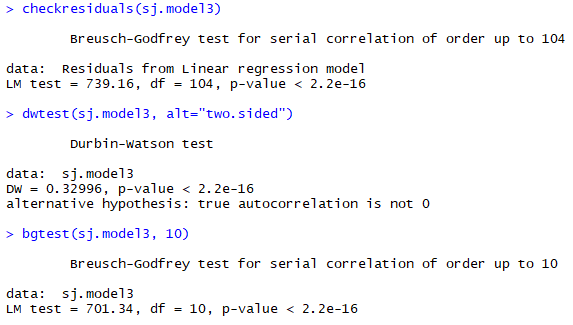
****

Figure 17.a. Residual Diagnostics for San Juan

From the above figure we can see that our residuals don’t have constant variance. We can also see from the Acf plot that our residuals are autocorrelated, meaning that there is information left in the residuals which should be used in computing forecasts. We can also see that our residuals are not normally distributed.

Output:



The Breusch-Godfrey test has a p-value of 2.2e-16, which is highly significant. Thus, we can conclude that our residuals are serially correlated.

The Durbin-Watson test has a DW value of 0.033 i.e. <2. Thus, we can conclude that there is positive autocorrelation in our residuals.

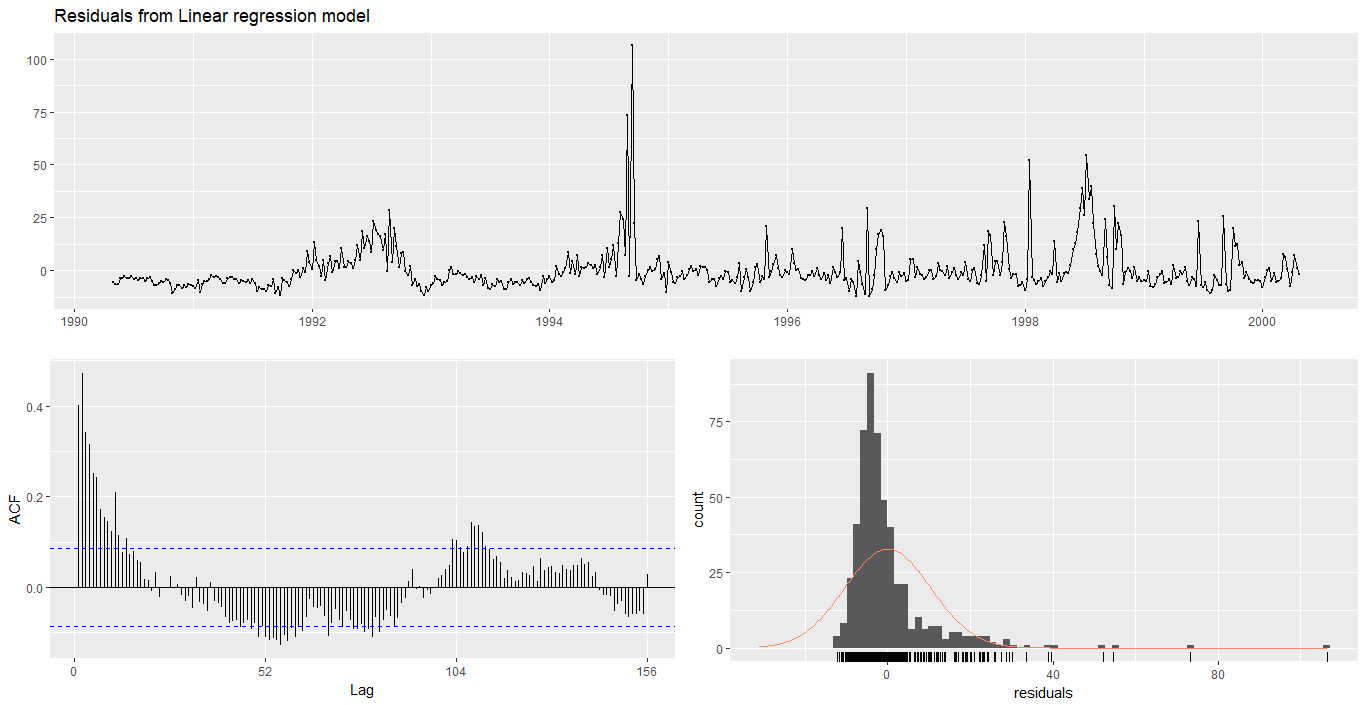
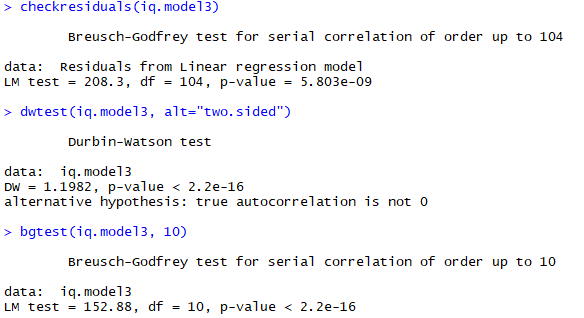


Figure 17.b. Residual Diagnostics for Iquitos

From figure 17.b., we can see that our residuals do have have constant variance. We can also see that the residuals are highly autocorrelated. This means that there is information left in the residuals which should be used in computing forecasts. Also, we can see that the residuals are not normally distributed.

Output:



The Breusch-Godfrey test has a p-value of 5.0803e-09, which is highly significant. Thus, we can conclude that our residuals are serially correlated.

The Durbin-Watson test has a DW value of 1.1982 i.e. <2. Thus, we can conclude that there is positive autocorrelation in our residuals.

**3. Residual diagnostics for ARIMA model with exogenous variables**

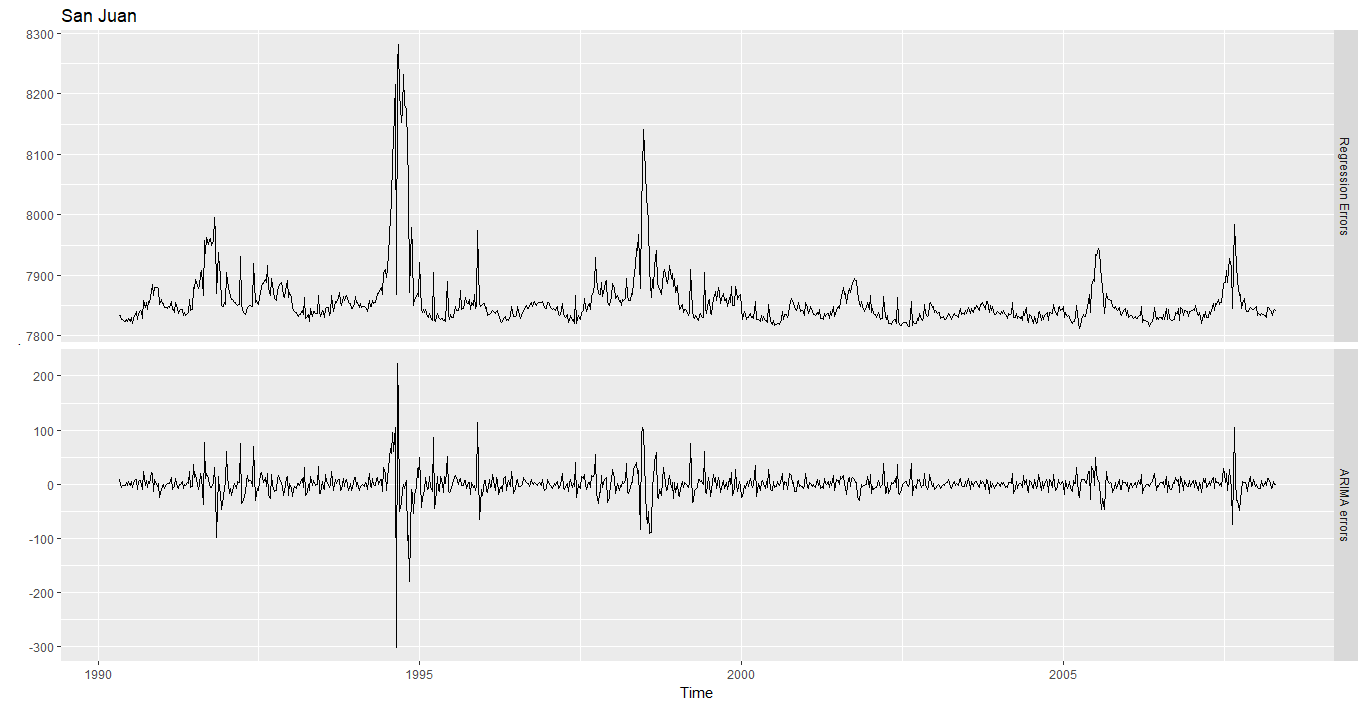
****

Figure 18.a. Regression errors and ARIMA errors from the fitted model.

The residuals of the ARIMA model should look like white noise. Our residuals do not perfectly look like white noise. Our model had a regression component and an ARIMA component.

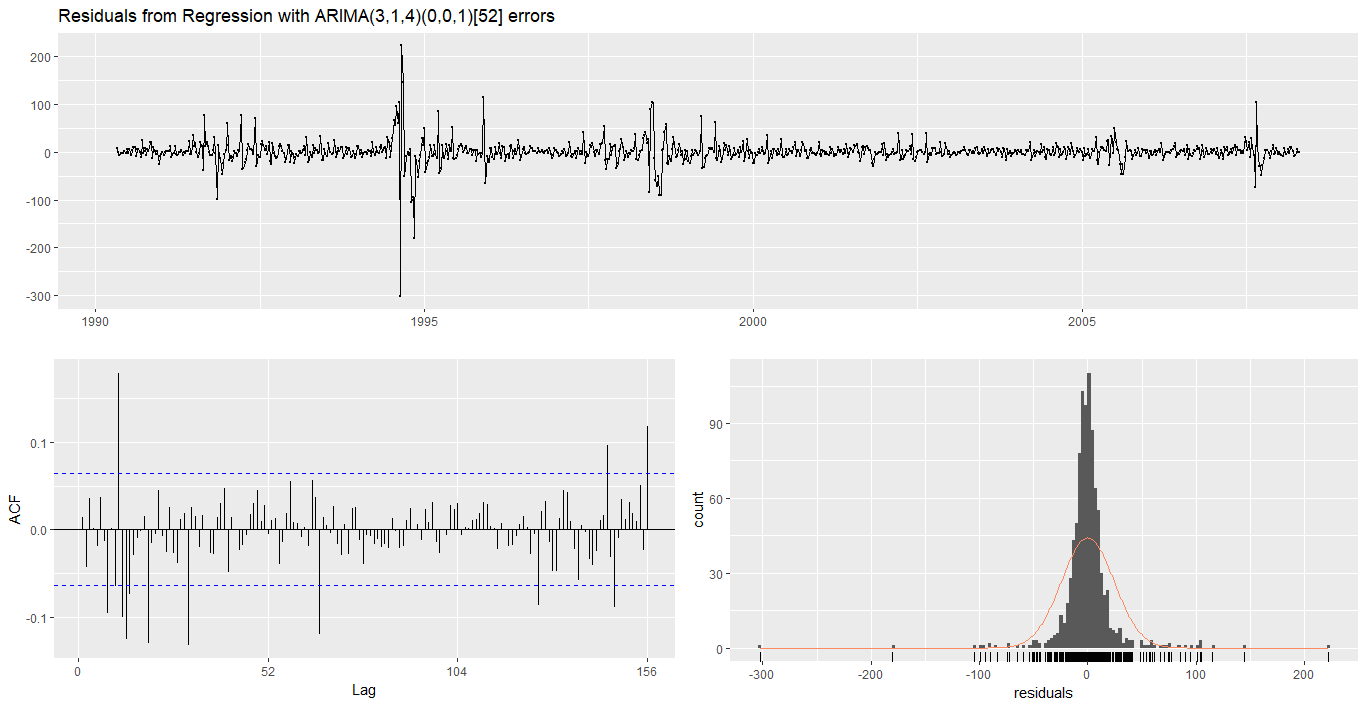
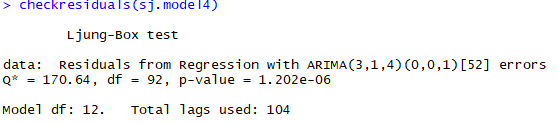


Figure 18.b. Residual diagnostics for ARIMA model for San Juan

When we look at our residuals, we can see that the residuals are not white noise. Ther are some lags in our ACF plot are well outside our critical values. There was a lag within the first few lags that went outside our critical point. Our residual almost normally distributed. The AIC score for our ARIMA model is more than that for Negative Binomial model.

Output:



Our Ljung-Box test had a p-value of 1.202e-06 i.e. <0.05. Thus, we reject our null hypothesis that the data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process). Thus, the data are not independently distributed; they exhibit serial correlation.

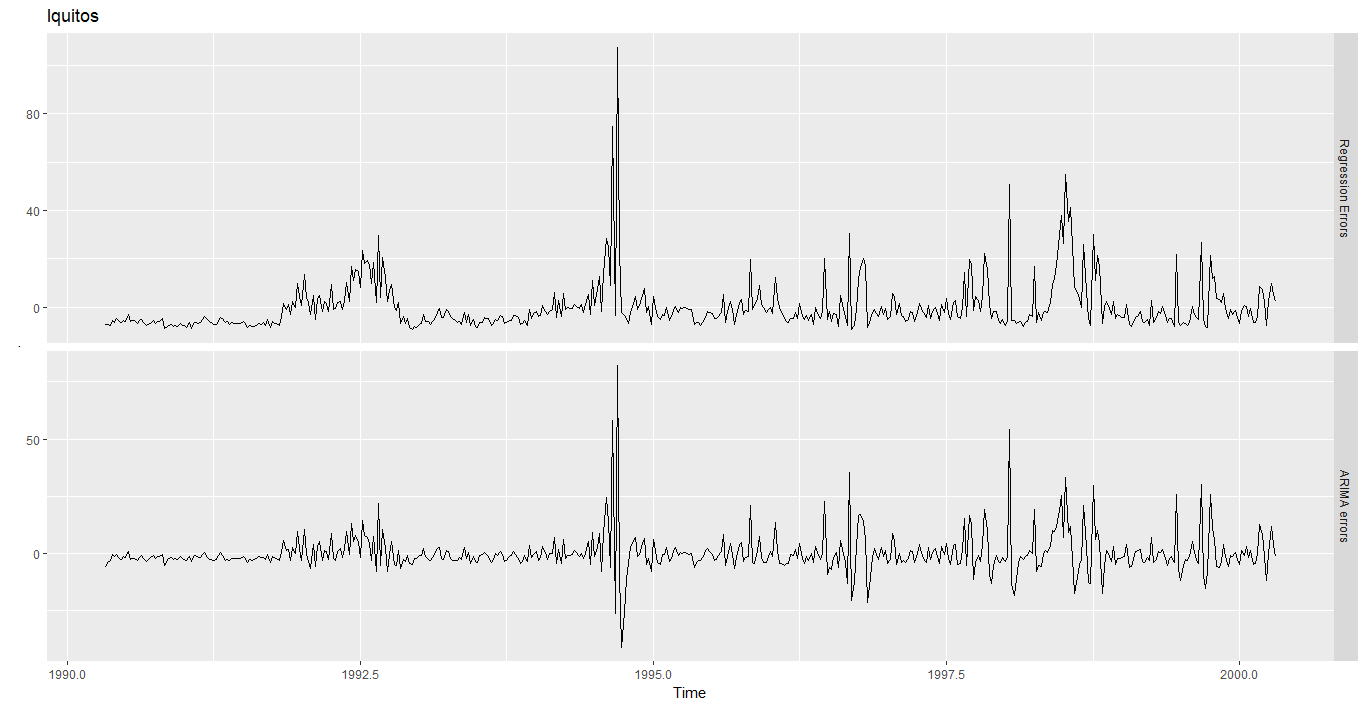
****

Figure 19.b. Regression errors and ARIMA errors from the fitted model.

We can see that our residuals for our ARIMA model is not a perfect white noise. Our ARIMA had a regression component and a ARIMA component.

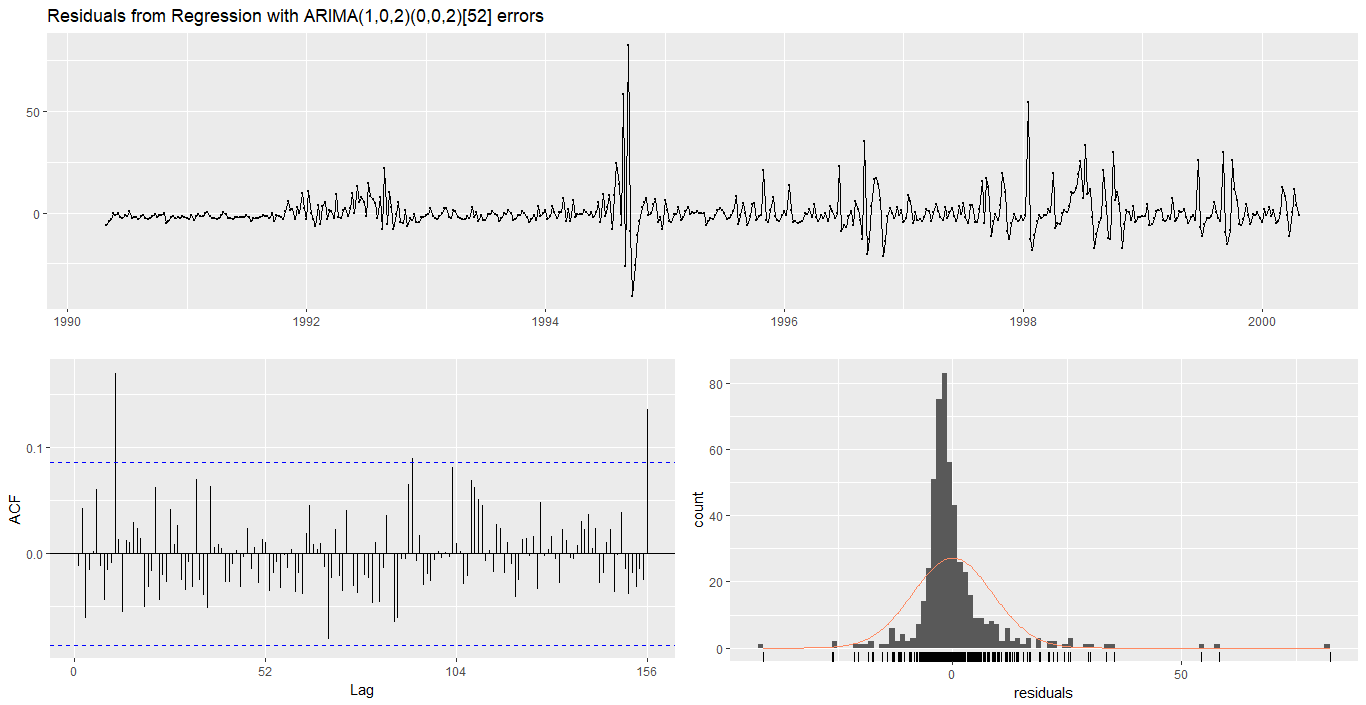
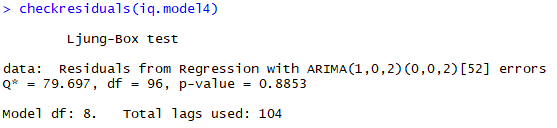


Figure 19.b. Residual diagnostics for ARIMA model for Iquitos

From the above figure, we can see that our residual are not white noise. The leading lag for our model was well outside our critical value. Our residuals were somewhat normally distributed with some values reaching far at the corners.

Output:



Our Ljung-Box test had a p-value of 0.8853 i.e. >0.05. Thus, we accept our null hypothesis that the data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

So far among the three models we have tried, the negative binomial regression performed best with lowest AIC score. To improve the model even more, we can check for autocollinearity in the data variables and residuals for future improvement.

**Driven Data Submission**

****

Figure 20.a. DriveData submission for Negative Binomial Regression

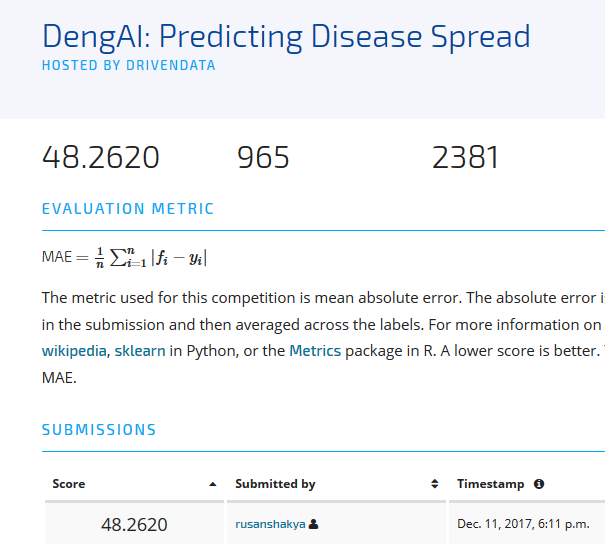


Figure 20.b. DrivenData submission for ARIMA model

**References**

[1] Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management*, 19–31.

[2] Disease management with ARIMA model in time series” by Renato Cesar Sato.

[3] Time series forecasting using a hybrid ARIMA and neural network model

G. Peter Zhang, Department of Management, J. Mack Robinson College of Business, Georgia State University, University Plaza, Atlanta, GA 30303, USA. Received 16 July 1999; accepted 23 November 2001