DengAI Project

By Mimi Trinh

INTRODUCTION

This report includes the purpose, methodology, results, and recommendations for DengAI, a competition on drivendata.org. Predictive models are evaluated by submitting them to drivendata.org, which will provide a public score for each model based on a portion of the validation dataset. At the conclusion of the competition, the score will be updated based on the holdout dataset.

REPORT

Section 1: Problem Definition and Significance

The name of the competition DengAI comes from the dengue fever, a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world where mosquito population is bigger than other places around the globe. In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death. Historically, the disease has been most prevalent in Southeast Asia and the Pacific Islands. However, in recently years, dengue fever has been spreading. These days many of the nearly half billion cases per year are occurring in South America. As a result, this is a very significant global issue. Being able to predict the cases for a region can help the government and medical staff of that country better prepare and prevent the disease.

This disease also has a personal significant impact on me since I grew up in Vietnam where dengue fever is a real threat and fear of many people, including my family. Every night we covered up our skin and burned incense sticks to keep the mosquitos away to prevent dengue fever.

Because dengue fever is carried by mosquitoes, the transmission dynamics of the disease are related to climate variables. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide. Using environmental data describing changes in temperature, precipitation, vegetation, and more, which is collected by various US Federal Government agencies (Centers for Disease Control and Prevention and the National Oceanic and Atmospheric Administration), we will predict the number of dengue fever cases reported each week in Iquitos, Peru and San Juan, Puerto Rico.

Section 2: Descriptive Statistics, Data Visualization, and Processing

The raw data comes in four different files 1) response variable (total cases), city, year, week of the year 2) predictor variables (environmental factors) 3) forecast predictor variables 4) submission template. The first two files are used as train dataset to build the model. The third file is used as the test dataset to generate the forecasts. The fourth file is used to submit the results to drivendata.org. The goal is to predict the total cases label for each (city, year, week of year) in the test set ("iq" is Iquitos and "sj" is San Juan). Iquitos and San Juan have test data spanning three and five years respectively. The test set is a pure future holdout set, which means that test data is sequential and non-overlapping with any of the train data. Missing values have been filled as "NaN." By combining the first two files into one data frame in R, we generate the list of variables in exhibit 1.

City and date indicators

- city – City abbreviations: sj for San Juan and iq for Iquitos
- week_start_date – Date given in yyyy-mm-dd format

NOAA's GHCN daily climate data weather station measurements

- station_max_temp_c – Maximum temperature
- station_min_temp_c – Minimum temperature
- station_avg_temp_c – Average temperature
- station_precip_mm – Total precipitation
- station_diur_temp_rng_c – Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- precipitation_amt_mm – Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- reanalysis_sat_precip_amt_mm – Total precipitation
- reanalysis_dew_point_temp_k – Mean dew point temperature
- reanalysis_air_temp_k – Mean air temperature
- reanalysis_relative_humidity_percent – Mean relative humidity
- reanalysis_specific_humidity_g_per_kg – Mean specific humidity
- reanalysis_precip_amt_kg_per_m2 – Total precipitation
- reanalysis_max_air_temp_k – Maximum air temperature
- reanalysis_min_air_temp_k – Minimum air temperature
- reanalysis_avg_temp_k – Average air temperature
- reanalysis_tdtr_k – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index(0.5x0.5 degree scale) measurements

- ndvi_se – Pixel southeast of city centroid
- ndvi_sw – Pixel southwest of city centroid
- ndvi_ne – Pixel northeast of city centroid
- ndvi_nw – Pixel northwest of city centroid

Then we separate the train data into two data frames: one for Iquitos and one for San Juan. Each data frame has 25 variables, as outlined above and in exhibit 1. There are 520 observations in Iquitos and 936 observations in San Juan. Exhibit 2 shows the descriptive statistics for these two data frames. If the mean and median are close to each other, that means the variable has normally distributed data. This is not the case with some of the variables using the descriptive statistics in exhibit 2, so we can conclude that there's skewness in some of the variables. In addition, all of the predictors have missing values in both data frames, so we will need to address this issue to clean the data prior to building predictive models. There's also skewness toward the right from observing exhibit 3 since the response variable has many 0's in both cities (96 in Iquitos and 184 in San Juan). In Iquitos, the mean is 7.565, and the median is 5 whereas in San Juan, the mean is 34.18, and the median is 19. Since the mean is higher than the median, we can confirm there's positive skewness in the data.

From the correlation plots in exhibit 4, there's strong multicollinearity among the predictors in both data frames with dark blue indicates a strong positive relationship, and dark red indicates a strong negative relationship. Moreover, many variables capture the same information since they represent the same environmental factors but collected by different US agencies. Therefore, during the model development process, feature engineering is utilized since we don't need all predictors in the models. The biggest concern during the data processing and cleaning procedure is the large number of missing values in the dataset. Specifically, there's no predictor with a large portion of missing values, but each predictor has some missing values. Thus, we can't completely drop a variable but will need to address the issue. During the exploratory data analysis (EDA) process, we recognize that some predictors have skewness, so instead of the mean, we replace missing values with the median of that variable.

The final part of the data cleaning and processing aspect is to convert the two data frames into two time series with frequency of 52 to represent 52 weeks in a year: Iquitos from 7/1/2000 to 6/25/2010 and San Juan from 4/30/1990 to 4/22/2008. From the auto plots in exhibit 5, both time series have cyclical behavior but no strong seasonality. Also, there's no trend in either series. Interestingly, there's a pike in the middle of the year 2004 in Iquitos as a significant outlier with almost 120 cases whereas in San Juan, there's a pike in the last quarter of the year 1995 with a significant outlier with more than 400 cases.

Section 3: Literature of Peer Reviewed Journals

This model utilizes multiple techniques in build models: naïve, seasonal naïve, drift, ARIMA, and neural networks. The following five peer reviewed journals show how these types of models are used in other situations to forecast time series data.

- Article 1: use naïve method for "an informed trader in after-hours trading to earn abnormal returns from identifying and investing in stocks with a positive earnings surprise" (Billingsley and Resnick, 2014). Naïve method in this example is not the same as naïve method in our model. Specifically, naïve method in the article utilizes moving average of the time series data on earnings, with the condition of analyst forecast. Meanwhile, naïve method in this project is simply the same as the last observation. Thus, the method in the article is more similar to the seasonal naïve method in this project in which the forecasts are generated based on the last observations of the same seasons.
- Article 2: use neural networks to forecast time series data of the S&P 500 index, exchange rate, and interest rate while utilizing Akaike's information criterion (AIC) and Bayesian information criterion (BIC) as performance measurement metrics to determine the best model on both train set and test set (Qi and Zhang, 2001).

- Article 3: use neural networks to forecast four time series 1) monthly critical radio frequency in Washington DC 2) monthly number of pigs slaughtered in Victoria 3) Mackey-Glass chaotic series 4) daily number of human faces that entered a sports store (Pereira, Cortez, and Mendes, 2017).
- Article 4: use ARIMA model to evaluate the effectiveness of intervention programs. Specifically, the research question asks if the district-wide drug abuse prevention program interrupts the series of drug-abuse behaviors observed in middle school children. If the trend or seasonality is broken, it supports the positive impact of the program (Braden and Gonzalez, 1990).
- Article 5: use ARIMA model to measure changes in behavior occurring during nursing period. "A goal of clinical research is to identify, describe, explain, and predict the effects of processes that bring about therapeutic change over an entire course of treatment... Research questions are often: Can patterns of change be reliably identified? Are these patterns of change related to the outcome?" (Jensen, 1990).

Section 4: Programming Formulation in R and Model Development

Typically, Python is better than R to generate neural networks because with Python, we can specify the number of hidden layers whereas R doesn't allow this. However, majority of machine learning problems can be solved with only one hidden layer in neural networks, which R supports. We don't handle complex data and issues such as image processing and natural language processing, so there's no need to utilize more than one hidden layer. Therefore, R is the chosen program for this project. Six sets of predictive models are developed to forecast total cases, one set for Iquitos and one set for San Juan.

- Model #1 naïve method: forecasts are generated based on the last observation
- Model #2 seasonal naïve method: forecasts are generated based on the last observations of each season
- Model #3 drift method: forecasts are generated by drawing a line between the first and last observations and extrapolating it into the future
- Model #4 ARIMA method: use auto.arima() to let R determine the best ARIMA model. The chosen models are ARIMA(1,0,4) for Iquitos and ARIMA(1,1,1) for San Juan
- Model #5 neural network method with no Box Cox transformation: use nnetar() and lambda=NULL to let R determine the best neural network model. The chosen models are NNAR(5,1,6)[52] for Iquitos and NNAR(14,1,10)[52] for San Juan
- Model #6 neural network method with Box Cox transformation: use nnetar() and lambda=auto to let R determine the best neural network model. The chosen models are NNAR(6,1,6)[52] for Iquitos and NNAR(16,1,12)[52] for San Juan.

For model #6, because some of the Box Cox transformation, such as log transformation, can't handle values of 0's, so we add 1 to each predictor and response variable. When forecasts are generated on the test set, we subtract 1 from each predicted value.

Exhibit 6 shows the auto plots of the first three models. Naïve and drift methods only have a straight line in forecast, which means they're useless for this project. Seasonal naïve has some variation in forecasts, but it only mirrors the previous season. Therefore, though we already expect the first three models to perform poorly, it's always good to try. The first three models use only the response variable whereas the last three models use the predictors to forecast the response variable. However, as mentioned earlier, we won't use all predictors to develop models because there's multicollinearity as shown in the correlation plots in the exhibit and also because R takes a long time to process all predictors. As part of feature engineering, by removing duplicate predictors that measure the same thing but are collected by different US agencies and those that are highly correlated with others, we narrow down to the following five predictors: ndvi_ne, precipitation_amt_mm, reanalysis_avg_temp_k, reanalysis_specific_humidity_g_per_kg, reanalysis_tdtr_k.

Section 5: Model Performance and Key Learnings

Using checkresiduals() function in R, we generate the residual plots in exhibit 7 and p-values of the Ljung-Box test. Small p-values (less than 0.05 alpha) indicate that residuals don't resemble white noise. The first three models don't have white noise residuals. Model #4 has white noise residuals. Model #5 doesn't generate p-value, but from the ACF plots, though there are some pikes out of bounds, residuals still mostly follow white noise. Model #6 doesn't work on checkresiduals() function because ggtsdisplay for histogram is only for univariate time series. These are the first key learnings.

Using accuracy() function, we generate the output in exhibit 8, which includes a list of error terms serving as metrics to measure model performance. The lower the error term, the more accurate the model. Models # 4, 5, 6 perform much better than models # 1, 2, 3 as expected, which means that the chosen predictors are significant. Among the last three models, model #5 performs the best in both time series train set based on majority of the error terms. These are the second key learnings.

Applying the predicted explanatory variables, we generate forecasts for the test set. The auto plots of the forecast, as shown in exhibit 9, show that models using predictors are much better than those without. After submitting the predicted values of each model to drivendata.org, we get the following MAE for each model applied on the test set: 33.1875 in model #4, 32.0865 in model #5, and 31.5625 in model #6. Therefore, the best model with the lowest error term on the test set is model #6 using neural networks and Box Cox transformation (Iquitos is NNAR(6,1,6)[52] and San Juan is NNAR(16,1,12)[52]). Model #5 performs best on train set but not test set, which is probably due to overfitting, a common issue in neural networks. These are the third key learnings.

Section. 6: Model Limitations and Future Recommendations

Below are the limitations and future recommendations for each limitation of this project.

The best model #6 ranks us at 1743 out of 5192 competitors, which is the top 33.6% or one-third of the competition. The #1 ranked model on drivendata.org posted on the leaderboard of this competition has MAE of 13.0144 with 22 entries. One limitation of this project is the high MAE of the test set with only four entries. Thus, it's recommended to continue the project with more entries to increase the MAE.

Another limitation of this project is the fact we drop many variables from the model development process due to the time constraint in R using a personal computer. Therefore, it’s recommended to use more predictors in the model with a machine with more processing power.

Values of the predictors used in the test set in this case are already given, which limits the data scientists from understand how these are generated. Future researchers should develop additional models to forecast the chosen predictors in the models.

Finally, the model separates Iquitos and San Juan from each other since the beginning to conduct the analysis, which may lose insight into any correlation between the two cities. Future researchers are recommended to analyze any relationship of both response and predictor variables between the two cities.

CONCLUSION

In conclusion, this final project walks through the scope, methodology, results, key learnings, and recommendations for predictive models to forecast dengue fever in Iquitos, Peru and San Juan, Puerto Rico. Six models are developed in the analysis. Neural networks model with no Box Cox transformation performs the best on the train set whereas neural networks model with Box Cox transformation performs the best on the test set. It’s evidenced in the results that the predictors are significant to generate prediction for the total cases of dengue fever in the two cities. However, the models developed in this project still have many limitations, which can be addressed based on the recommendations outlined above for future researchers.

References

Billingsley, R. S., & Resnick, B. G. (2014). A Trading Strategy to Profit from Overly Aggressive Downward Earnings Guidance. *Journal of Portfolio Management*, *40*(2). Retrieved from https://search-proquest-com.turing.library.northwestern.edu/docview/1496996171/abstract/72AC313DF17C41D6PQ/1?accountid=12861

Braden, Jeffrey P., and Gerardo M. Gonzalez. "Use of time-series, ARIMA designs to assess program efficacy." *School Psychology Review*, vol. 19, no. 2, 1990, pp. 224+, web.a.ebscohost.com.turing.library.northwestern.edu/ehost/detail/detail?vid=0&sid=5c3ebd0d-7d9a-4b23-a3e6-04575ccf1151%40sessionmgr4009&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#AN=9607015442&db=tfh

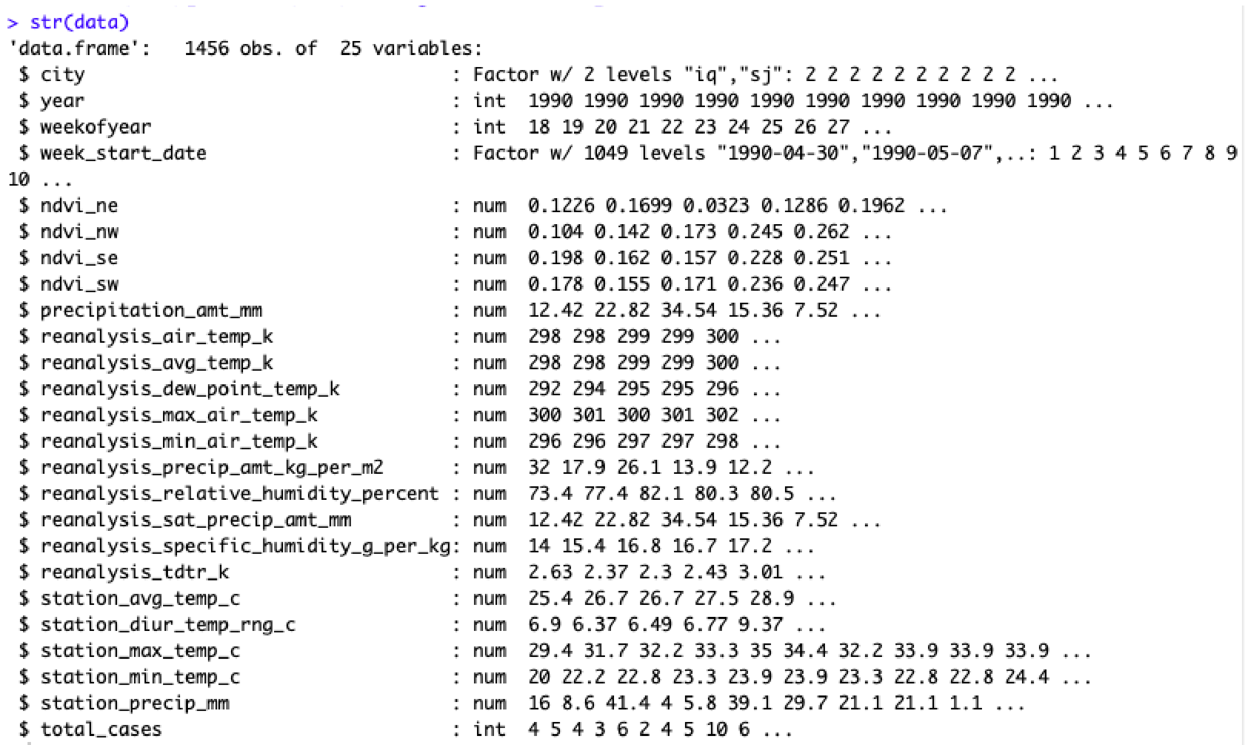
Jense, Louise. "Guidelines for the application of arima models in time series." *Research in Nursing & Health*, vol. 13, no. 6, Dec. 1990, pp. 429-35, search.library.northwestern.edu/primo-explore/fulldisplay?docid=TN_wj10.1002%2Fnur.4770130611&context=PC&vid=NULVNEW&lang=en_US&search_scope=NWU&adaptor=primo_central_multiple_fe&tab=default_t

Pereira P.J., Cortez P., Mendes R. (2017) Multi-objective Learning of Neural Network Time Series Prediction Intervals. In: Oliveira E., Gama J., Vale Z., Lopes Cardoso H. (eds) Progress in Artificial Intelligence. EPIA 2017. Lecture Notes in Computer Science, vol 10423. Springer, Cham

Qi, M., & Zhang, G. P. (2001, August 1). An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, *132*(3), 666-680. Retrieved from https://www-sciencedirect-com.turing.library.northwestern.edu/science/article/pii/S0377221700001715

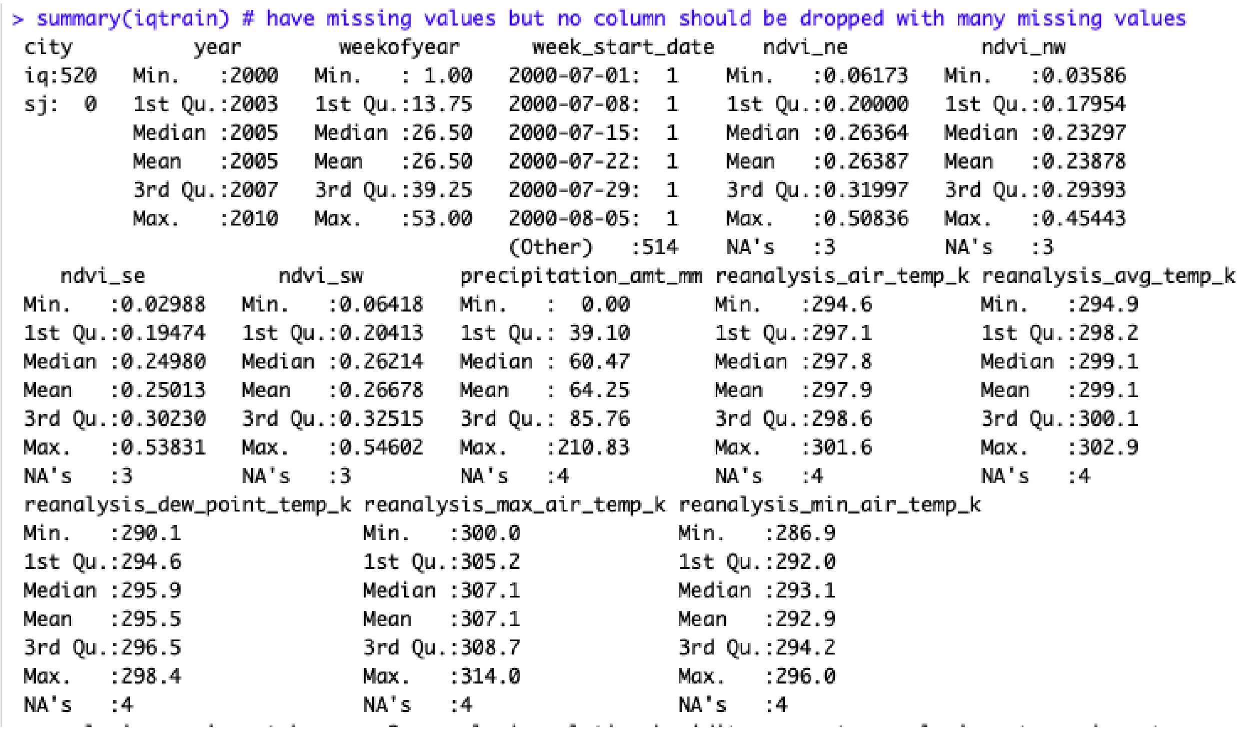
Exhibit

Exhibit 1

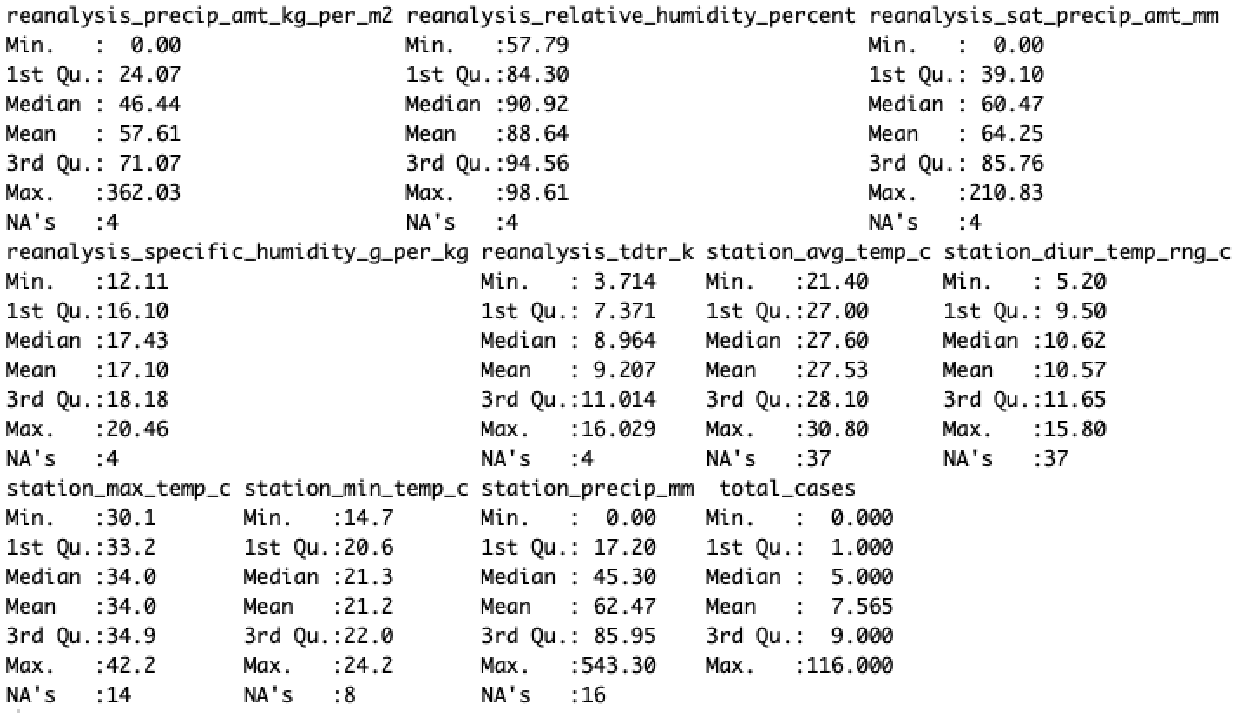


```
> str(data)
'data.frame':	1456 obs. of  25 variables:
 $ city                                 : Factor w/ 2 levels "iq","sj": 2 2 2 2 2 2 2 2 2 2 ...
 $ year                                 : int  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
 $ weekofyear                           : int  18 19 20 21 22 23 24 25 26 27 ...
 $ week_start_date                      : Factor w/ 1049 levels "1990-04-30","1990-05-07",..: 1 2 3 4 5 6 7 8 9
10 ...
 $ ndvi_ne                              : num  0.1226 0.1699 0.0323 0.1286 0.1962 ...
 $ ndvi_nw                              : num  0.104 0.142 0.173 0.245 0.262 ...
 $ ndvi_se                              : num  0.198 0.162 0.157 0.228 0.251 ...
 $ ndvi_sw                              : num  0.178 0.155 0.171 0.236 0.247 ...
 $ precipitation_amt_mm                 : num  12.42 22.82 34.54 15.36 7.52 ...
 $ reanalysis_air_temp_k                : num  298 298 299 299 300 ...
 $ reanalysis_avg_temp_k                : num  298 298 299 299 300 ...
 $ reanalysis_dew_point_temp_k          : num  292 294 295 295 296 ...
 $ reanalysis_max_air_temp_k            : num  300 301 300 301 302 ...
 $ reanalysis_min_air_temp_k            : num  296 296 297 297 298 ...
 $ reanalysis_precip_amt_kg_per_m2      : num  32 17.9 26.1 13.9 12.2 ...
 $ reanalysis_relative_humidity_percent : num  73.4 77.4 82.1 80.3 80.5 ...
 $ reanalysis_sat_precip_amt_mm         : num  12.42 22.82 34.54 15.36 7.52 ...
 $ reanalysis_specific_humidity_g_per_kg: num  14 15.4 16.8 16.7 17.2 ...
 $ reanalysis_tdtr_k                    : num  2.63 2.37 2.3 2.43 3.01 ...
 $ station_avg_temp_c                   : num  25.4 26.7 26.7 27.5 28.9 ...
 $ station_diur_temp_rng_c              : num  6.9 6.37 6.49 6.77 9.37 ...
 $ station_max_temp_c                   : num  29.4 31.7 32.2 33.3 35 34.4 32.2 33.9 33.9 33.9 ...
 $ station_min_temp_c                   : num  20 22.2 22.8 23.3 23.9 23.9 23.3 22.8 22.8 24.4 ...
 $ station_precip_mm                    : num  16 8.6 41.4 4 5.8 39.1 29.7 21.1 21.1 1.1 ...
 $ total_cases                          : int  4 5 4 3 6 2 4 5 10 6 ...
```

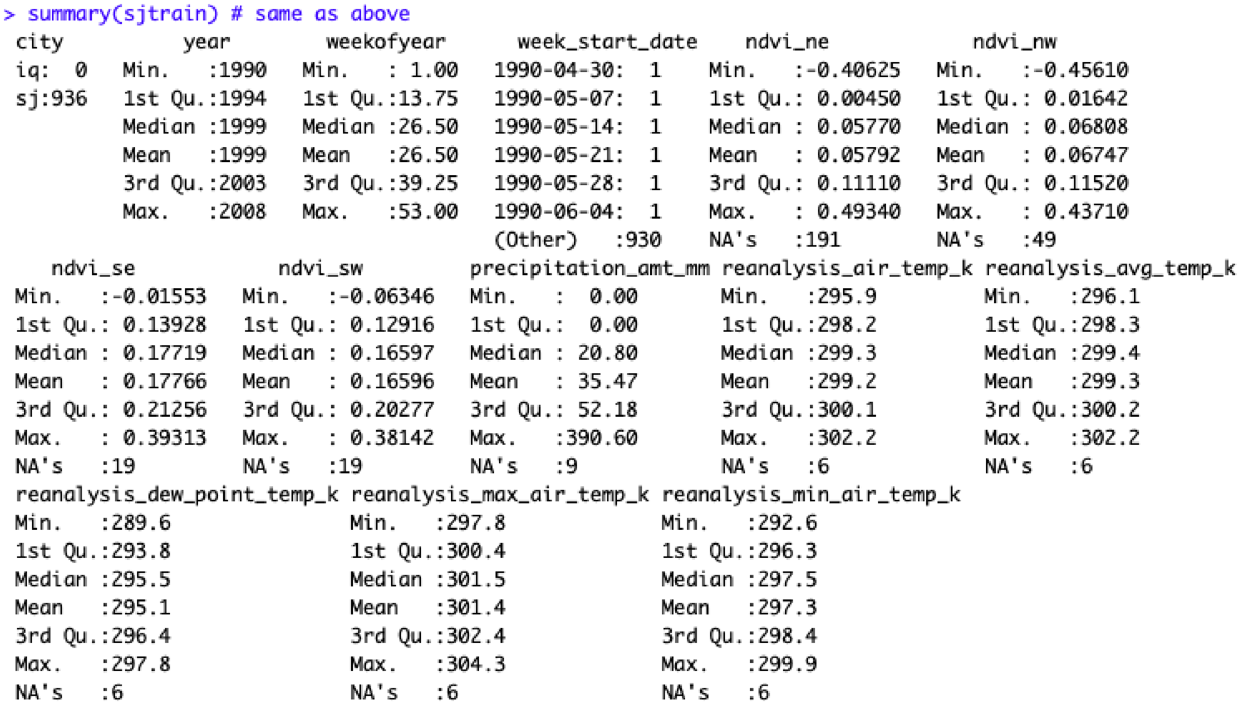
Exhibit 2



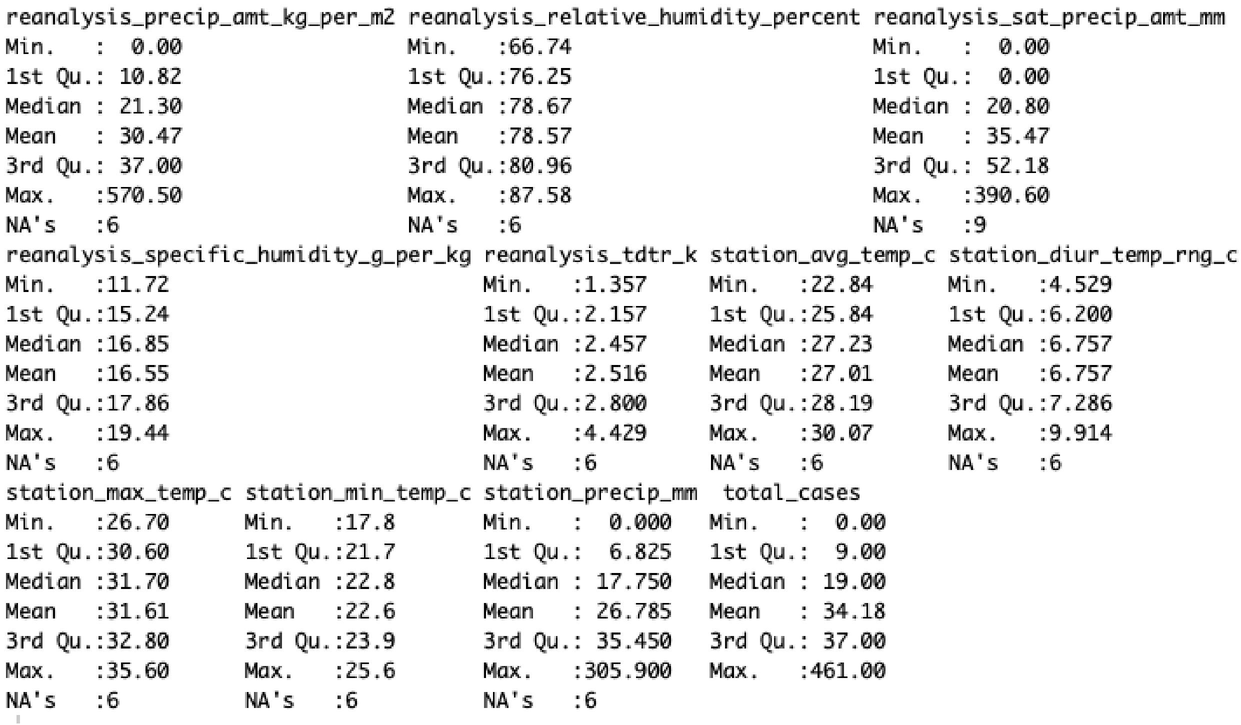
```
> summary(iqtrain) # have missing values but no column should be dropped with many missing values
 city          year        weekofyear      week_start_date     ndvi_ne            ndvi_nw       
 iq:520   Min.   :2000   Min.   : 1.00   2000-07-01:  1     Min.   :0.06173   Min.   :0.03586  
 sj:  0   1st Qu.:2003   1st Qu.:13.75   2000-07-08:  1     1st Qu.:0.20000   1st Qu.:0.17954  
          Median :2005   Median :26.50   2000-07-15:  1     Median :0.26364   Median :0.23297  
          Mean   :2005   Mean   :26.50   2000-07-22:  1     Mean   :0.26387   Mean   :0.23878  
          3rd Qu.:2007   3rd Qu.:39.25   2000-07-29:  1     3rd Qu.:0.31997   3rd Qu.:0.29393  
          Max.   :2010   Max.   :53.00   2000-08-05:  1     Max.   :0.50836   Max.   :0.45443  
                                         (Other)    :514    NA's   :3         NA's   :3        
    ndvi_se            ndvi_sw          precipitation_amt_mm reanalysis_air_temp_k reanalysis_avg_temp_k
 Min.   :0.02988   Min.   :0.06418   Min.   :  0.00       Min.   :294.6         Min.   :294.9        
 1st Qu.:0.19474   1st Qu.:0.20413   1st Qu.: 39.10       1st Qu.:297.1         1st Qu.:298.2        
 Median :0.24980   Median :0.26214   Median : 60.47       Median :297.8         Median :299.1        
 Mean   :0.25013   Mean   :0.26678   Mean   : 64.25       Mean   :297.9         Mean   :299.1        
 3rd Qu.:0.30230   3rd Qu.:0.32515   3rd Qu.: 85.76       3rd Qu.:298.6         3rd Qu.:300.1        
 Max.   :0.53831   Max.   :0.54602   Max.   :210.83       Max.   :301.6         Max.   :302.9        
 NA's   :3         NA's   :3         NA's   :4            NA's   :4             NA's   :4            
 reanalysis_dew_point_temp_k reanalysis_max_air_temp_k reanalysis_min_air_temp_k
 Min.   :290.1               Min.   :300.0             Min.   :286.9            
 1st Qu.:294.6               1st Qu.:305.2             1st Qu.:292.0            
 Median :295.9               Median :307.1             Median :293.1            
 Mean   :295.5               Mean   :307.1             Mean   :292.9            
 3rd Qu.:296.5               3rd Qu.:308.7             3rd Qu.:294.2            
 Max.   :298.4               Max.   :314.0             Max.   :296.0            
 NA's   :4                   NA's   :4                 NA's   :4                
```



```
reanalysis_precip_amt_kg_per_m2 reanalysis_relative_humidity_percent reanalysis_sat_precip_amt_mm
Min.   :  0.00                  Min.   :57.79                        Min.   :  0.00
1st Qu.: 24.07                  1st Qu.:84.30                        1st Qu.: 39.10
Median : 46.44                  Median :90.92                        Median : 60.47
Mean   : 57.61                  Mean   :88.64                        Mean   : 64.25
3rd Qu.: 71.07                  3rd Qu.:94.56                        3rd Qu.: 85.76
Max.   :362.03                  Max.   :98.61                        Max.   :210.83
NA's   :4                       NA's   :4                            NA's   :4
reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k station_avg_temp_c station_diur_temp_rng_c
Min.   :12.11                         Min.   : 3.714    Min.   :21.40      Min.   : 5.20
1st Qu.:16.10                         1st Qu.: 7.371    1st Qu.:27.00      1st Qu.: 9.50
Median :17.43                         Median : 8.964    Median :27.60      Median :10.62
Mean   :17.10                         Mean   : 9.207    Mean   :27.53      Mean   :10.57
3rd Qu.:18.18                         3rd Qu.:11.014    3rd Qu.:28.10      3rd Qu.:11.65
Max.   :20.46                         Max.   :16.029    Max.   :30.80      Max.   :15.80
NA's   :4                             NA's   :4         NA's   :37         NA's   :37
station_max_temp_c station_min_temp_c station_precip_mm  total_cases
Min.   :30.1       Min.   :14.7       Min.   :  0.00    Min.   :  0.000
1st Qu.:33.2       1st Qu.:20.6       1st Qu.: 17.20    1st Qu.:  1.000
Median :34.0       Median :21.3       Median : 45.30    Median :  5.000
Mean   :34.0       Mean   :21.2       Mean   : 62.47    Mean   :  7.565
3rd Qu.:34.9       3rd Qu.:22.0       3rd Qu.: 85.95    3rd Qu.:  9.000
Max.   :42.2       Max.   :24.2       Max.   :543.30    Max.   :116.000
NA's   :14         NA's   :8          NA's   :16
```



```
> summary(sjtrain) # same as above
 city         year        weekofyear      week_start_date     ndvi_ne             ndvi_nw
 iq:  0   Min.   :1990   Min.   : 1.00   1990-04-30:  1    Min.   :-0.40625   Min.   :-0.45610
 sj:936   1st Qu.:1994   1st Qu.:13.75   1990-05-07:  1    1st Qu.: 0.00450   1st Qu.: 0.01642
          Median :1999   Median :26.50   1990-05-14:  1    Median : 0.05770   Median : 0.06808
          Mean   :1999   Mean   :26.50   1990-05-21:  1    Mean   : 0.05792   Mean   : 0.06747
          3rd Qu.:2003   3rd Qu.:39.25   1990-05-28:  1    3rd Qu.: 0.11110   3rd Qu.: 0.11520
          Max.   :2008   Max.   :53.00   1990-06-04:  1    Max.   : 0.49340   Max.   : 0.43710
                                         (Other)   :930    NA's   :191        NA's   :49
    ndvi_se             ndvi_sw          precipitation_amt_mm reanalysis_air_temp_k reanalysis_avg_temp_k
 Min.   :-0.01553   Min.   :-0.06346   Min.   :  0.00       Min.   :295.9         Min.   :296.1
 1st Qu.: 0.13928   1st Qu.: 0.12916   1st Qu.:  0.00       1st Qu.:298.2         1st Qu.:298.3
 Median : 0.17719   Median : 0.16597   Median : 20.80       Median :299.3         Median :299.4
 Mean   : 0.17766   Mean   : 0.16596   Mean   : 35.47       Mean   :299.2         Mean   :299.3
 3rd Qu.: 0.21256   3rd Qu.: 0.20277   3rd Qu.: 52.18       3rd Qu.:300.1         3rd Qu.:300.2
 Max.   : 0.39313   Max.   : 0.38142   Max.   :390.60       Max.   :302.2         Max.   :302.2
 NA's   :19         NA's   :19         NA's   :9            NA's   :6             NA's   :6
 reanalysis_dew_point_temp_k reanalysis_max_air_temp_k reanalysis_min_air_temp_k
 Min.   :289.6               Min.   :297.8             Min.   :292.6
 1st Qu.:293.8               1st Qu.:300.4             1st Qu.:296.3
 Median :295.5               Median :301.5             Median :297.5
 Mean   :295.1               Mean   :301.4             Mean   :297.3
 3rd Qu.:296.4               3rd Qu.:302.4             3rd Qu.:298.4
 Max.   :297.8               Max.   :304.3             Max.   :299.9
 NA's   :6                   NA's   :6                 NA's   :6
```



```
reanalysis_precip_amt_kg_per_m2 reanalysis_relative_humidity_percent reanalysis_sat_precip_amt_mm
Min.   :  0.00                  Min.   :66.74                        Min.   :  0.00
1st Qu.: 10.82                  1st Qu.:76.25                        1st Qu.:  0.00
Median : 21.30                  Median :78.67                        Median : 20.80
Mean   : 30.47                  Mean   :78.57                        Mean   : 35.47
3rd Qu.: 37.00                  3rd Qu.:80.96                        3rd Qu.: 52.18
Max.   :570.50                  Max.   :87.58                        Max.   :390.60
NA's   :6                       NA's   :6                            NA's   :9
reanalysis_specific_humidity_g_per_kg reanalysis_tdtr_k station_avg_temp_c station_diur_temp_rng_c
Min.   :11.72                         Min.   :1.357     Min.   :22.84      Min.   :4.529
1st Qu.:15.24                         1st Qu.:2.157     1st Qu.:25.84      1st Qu.:6.200
Median :16.85                         Median :2.457     Median :27.23      Median :6.757
Mean   :16.55                         Mean   :2.516     Mean   :27.01      Mean   :6.757
3rd Qu.:17.86                         3rd Qu.:2.800     3rd Qu.:28.19      3rd Qu.:7.286
Max.   :19.44                         Max.   :4.429     Max.   :30.07      Max.   :9.914
NA's   :6                             NA's   :6         NA's   :6          NA's   :6
station_max_temp_c station_min_temp_c station_precip_mm  total_cases
Min.   :26.70      Min.   :17.8       Min.   :  0.000   Min.   :  0.00
1st Qu.:30.60      1st Qu.:21.7       1st Qu.:  6.825   1st Qu.:  9.00
Median :31.70      Median :22.8       Median : 17.750   Median : 19.00
Mean   :31.61      Mean   :22.6       Mean   : 26.785   Mean   : 34.18
3rd Qu.:32.80      3rd Qu.:23.9       3rd Qu.: 35.450   3rd Qu.: 37.00
Max.   :35.60      Max.   :25.6       Max.   :305.900   Max.   :461.00
NA's   :6          NA's   :6          NA's   :6
```

Exhibit 3

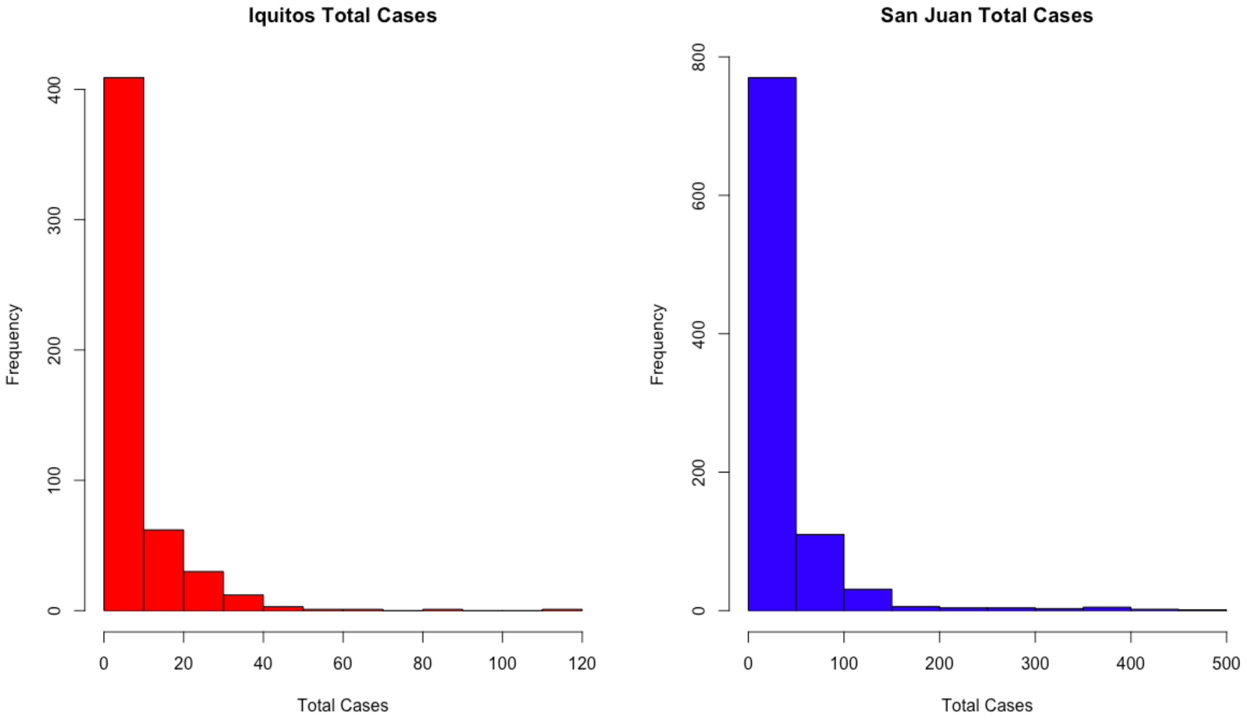
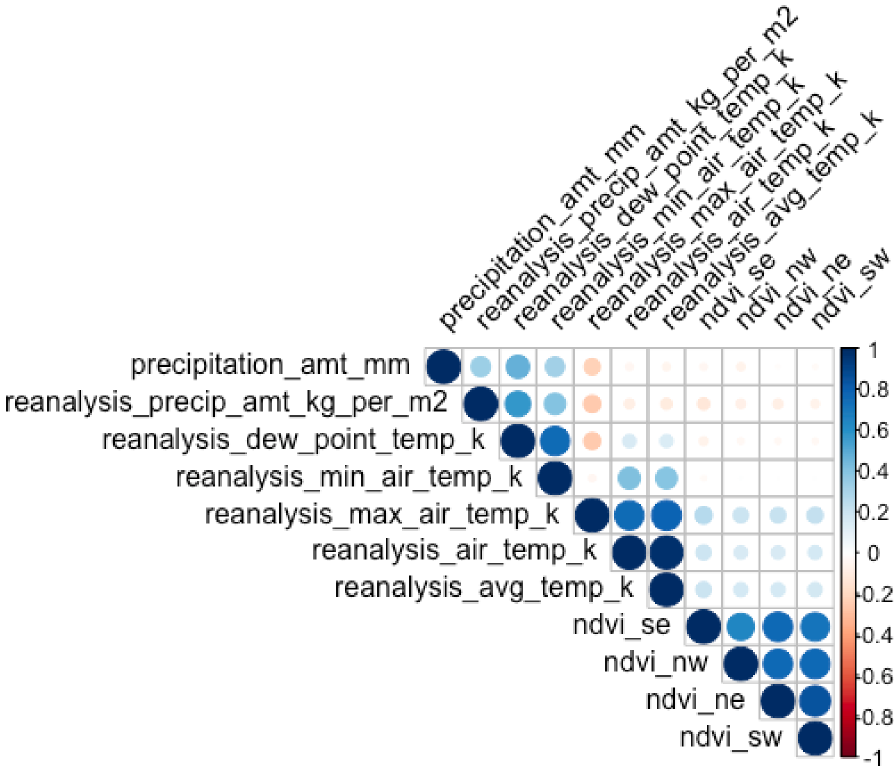
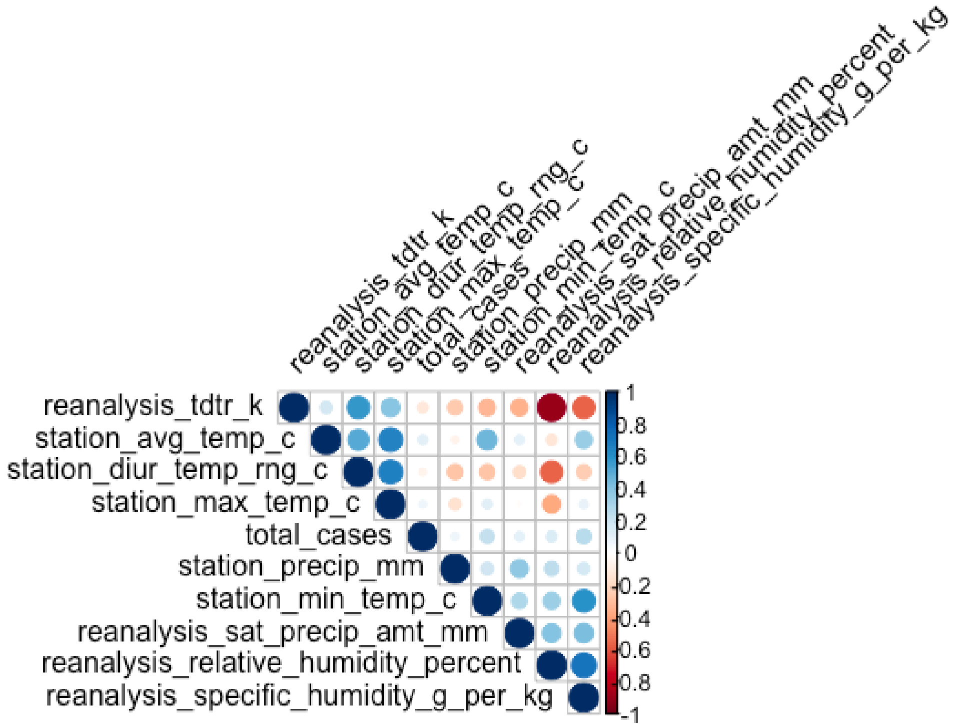
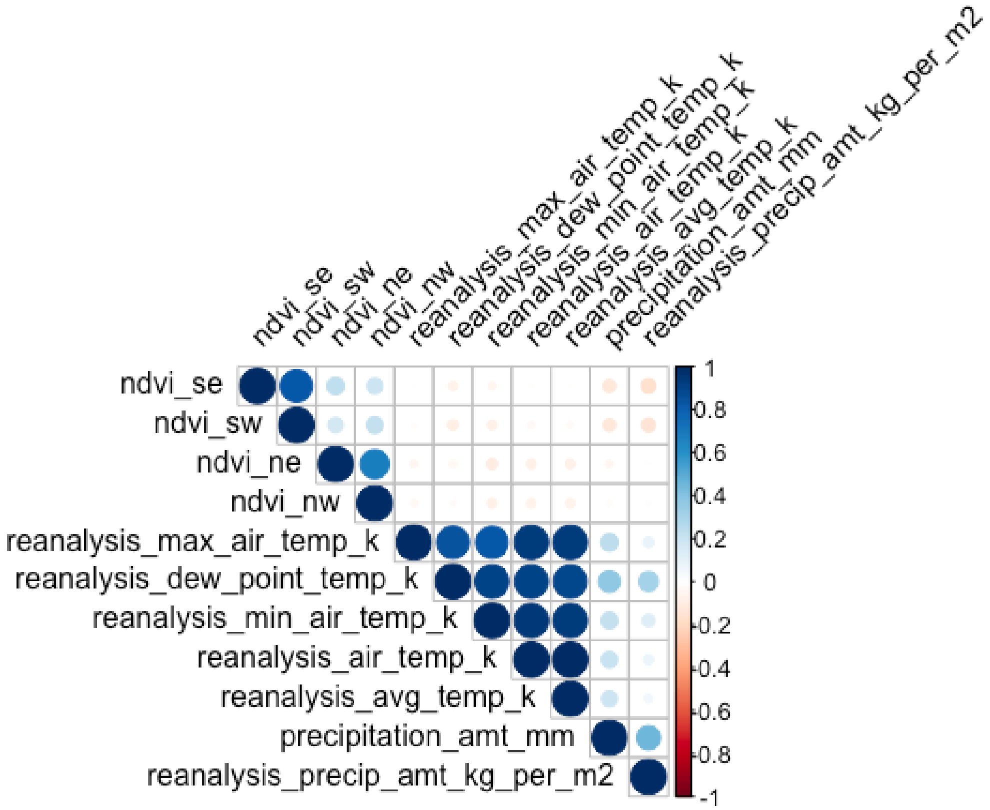


Exhibit 4


precipitation_amt_mm
reanalysis_precip_amt_kg_per_m2
reanalysis_dew_point_temp_k
reanalysis_min_air_temp_k
reanalysis_max_air_temp_k
reanalysis_air_temp_k
reanalysis_avg_temp_k
ndvi_se
ndvi_nw
ndvi_ne
ndvi_sw
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


reanalysis_tdtr_k
station_avg_temp_c
station_diur_temp_rng_c
station_max_temp_c
total_cases
station_precip_mm
station_min_temp_c
reanalysis_sat_precip_amt_mm
reanalysis_relative_humidity_percent
reanalysis_specific_humidity_g_per_kg
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


ndvi_se
ndvi_sw
ndvi_ne
ndvi_nw
reanalysis_max_air_temp_k
reanalysis_dew_point_temp_k
reanalysis_min_air_temp_k
reanalysis_air_temp_k
reanalysis_avg_temp_k
precipitation_amt_mm
reanalysis_precip_amt_kg_per_m2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

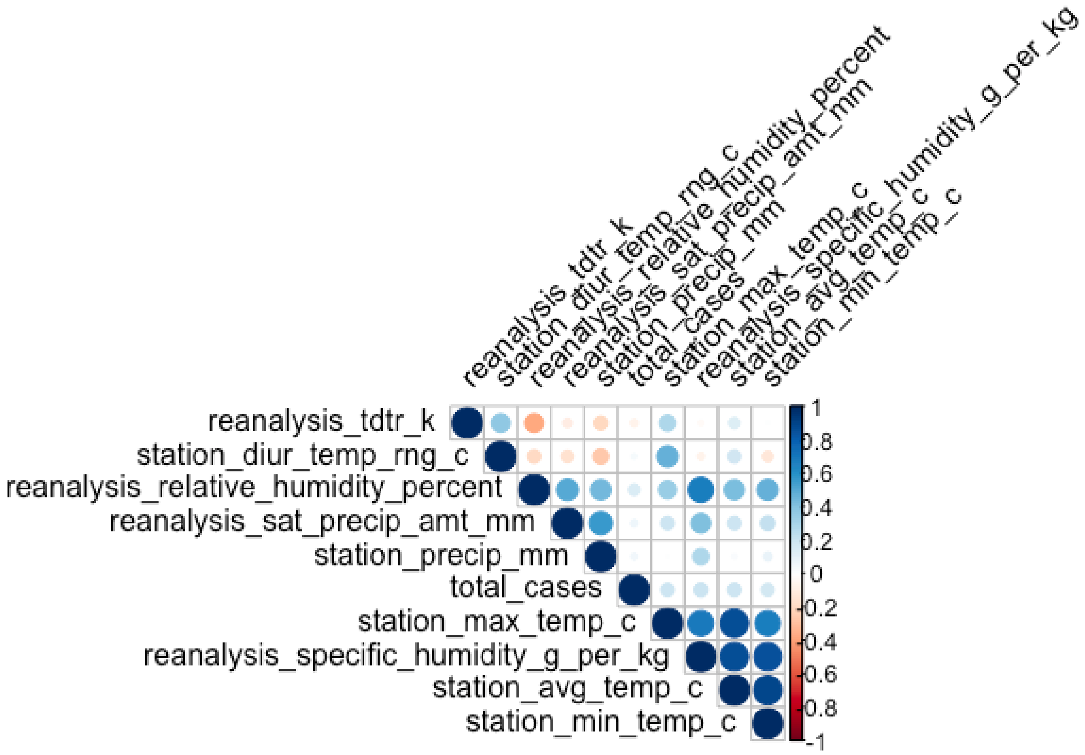
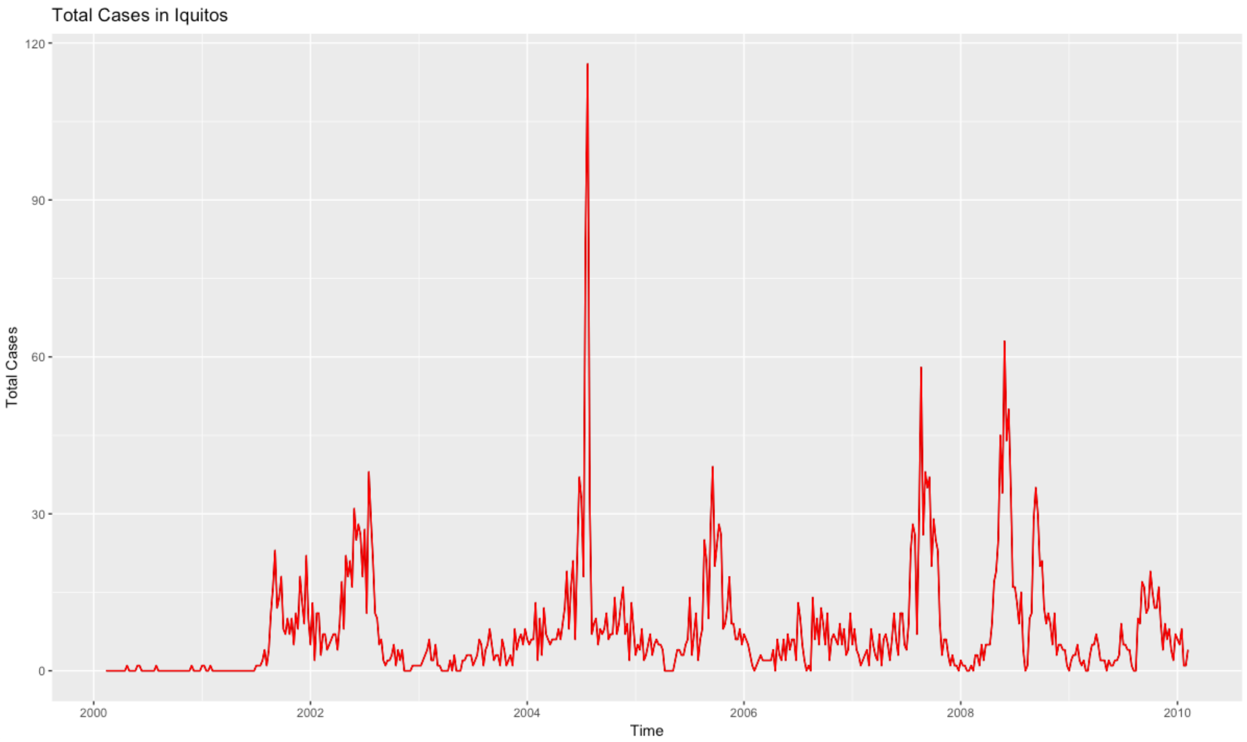

reanalysis_tdtr_k
station_diur_temp_rng_c
reanalysis_relative_humidity_percent
reanalysis_sat_precip_amt_mm
station_precip_mm
total_cases
station_max_temp_c
reanalysis_specific_humidity_g_per_kg
station_avg_temp_c
station_min_temp_c
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Exhibit 5


Total Cases in Iquitos

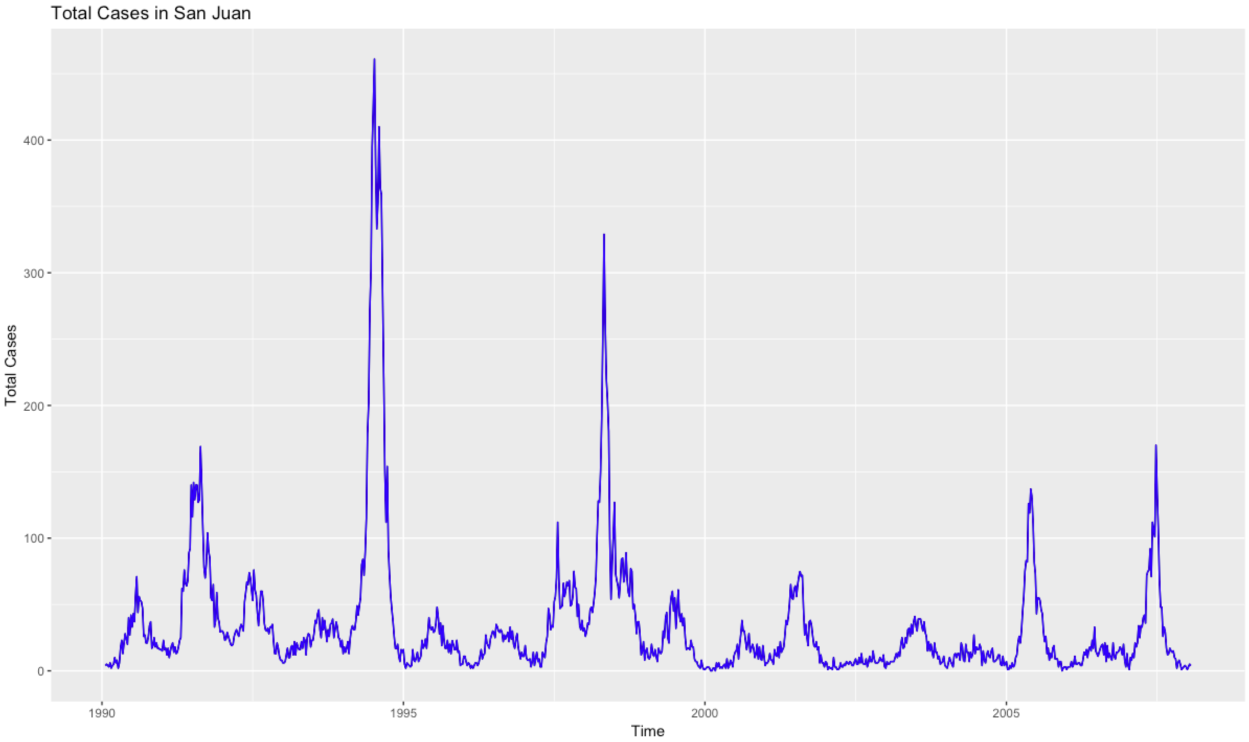
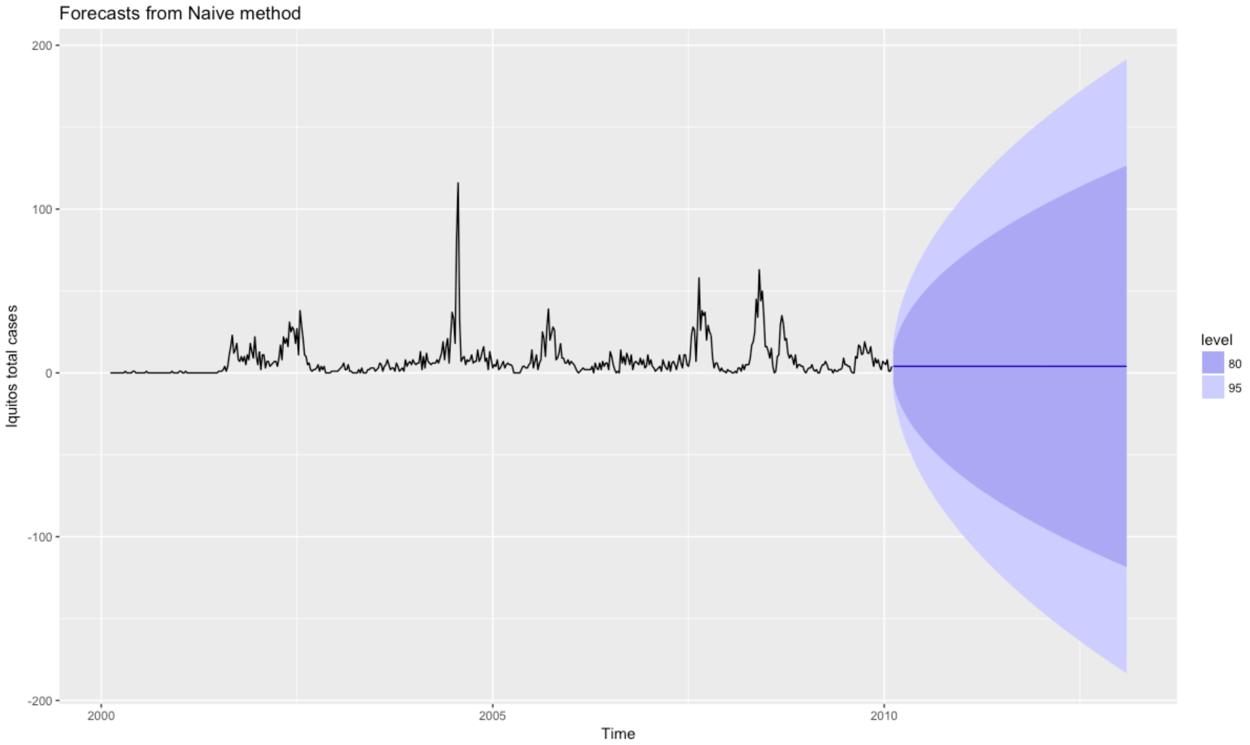
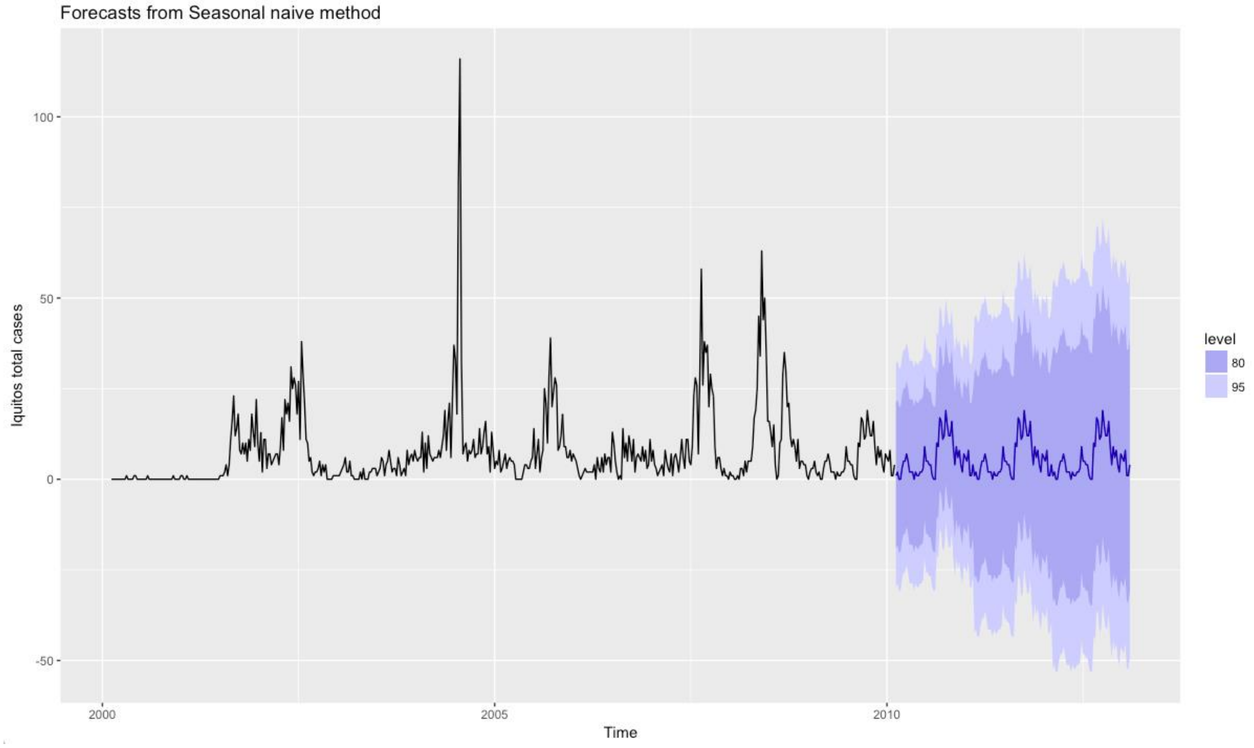
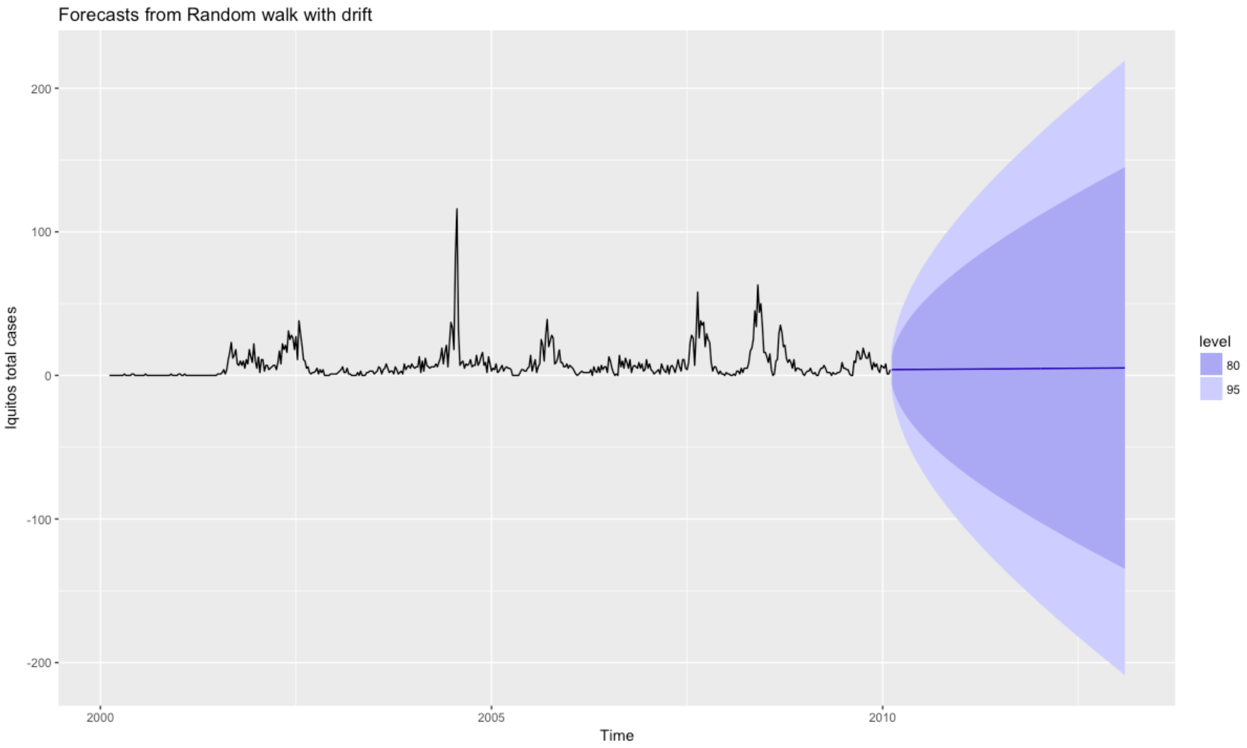
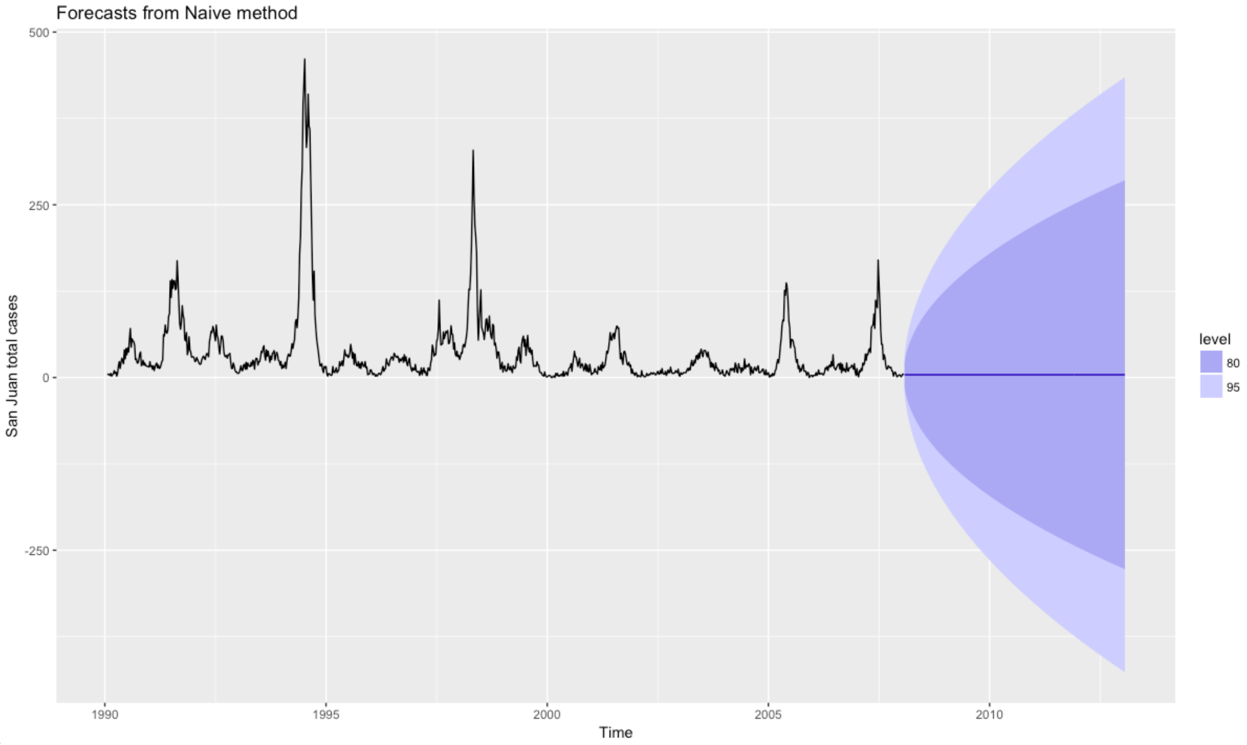

Total Cases in San Juan

Exhibit 6


Forecasts from Naive method
200
100
0
-100
-200
Iquitos total cases
2000
2005
2010
Time
level
80
95


Forecasts from Seasonal naive method
100
50
0
-50
Iquitos total cases
2000
2005
2010
Time
level
80
95


Forecasts from Random walk with drift
200
100
0
-100
-200
Iquitos total cases
2000
2005
2010
Time
level
80
95


Forecasts from Naive method
500
250
0
-250
San Juan total cases
1990
1995
2000
2005
2010
Time
level
80
95

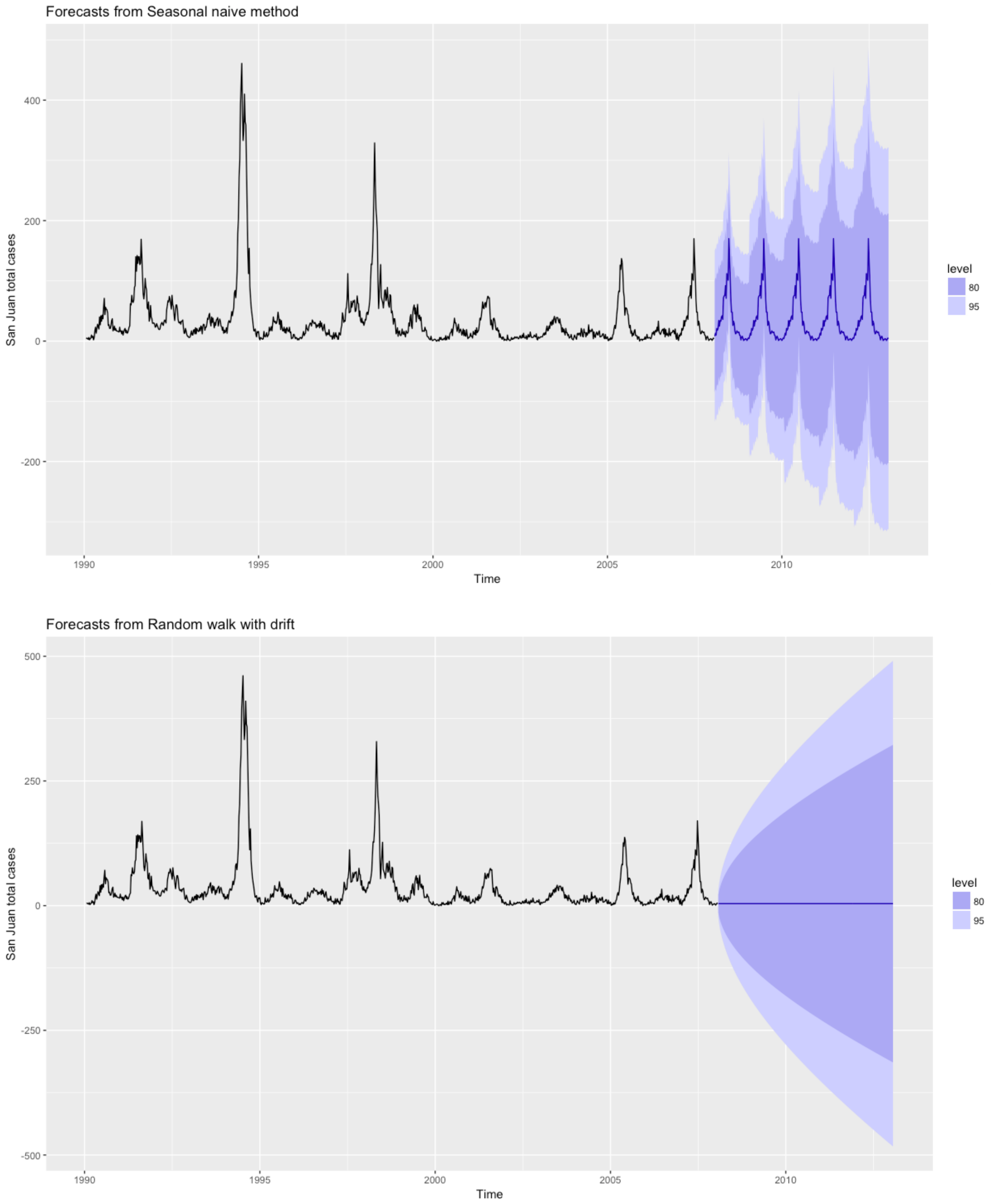
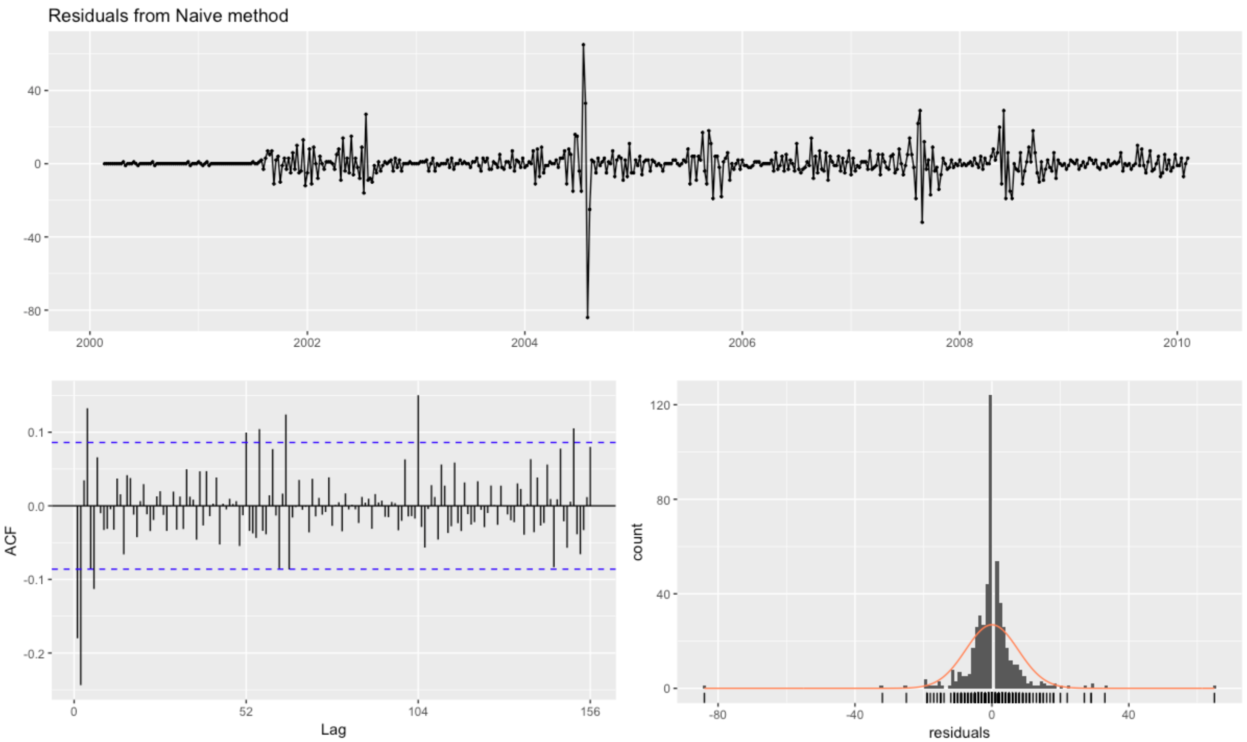
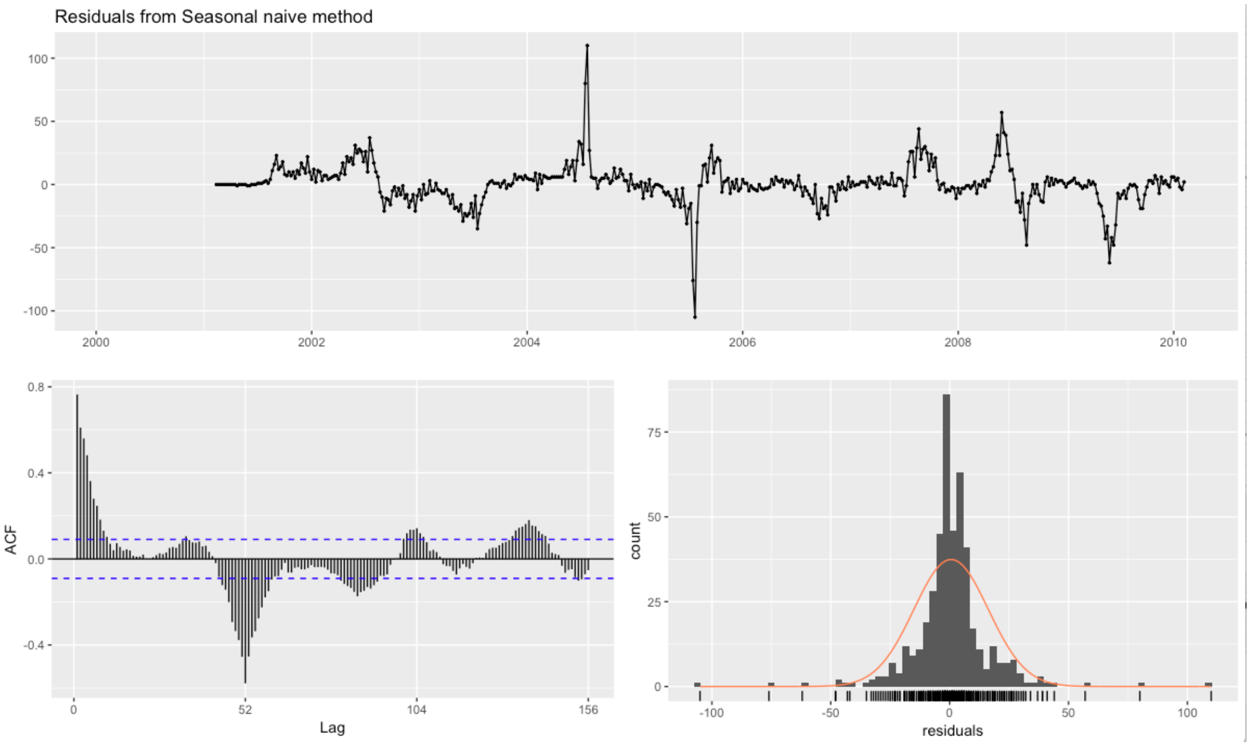
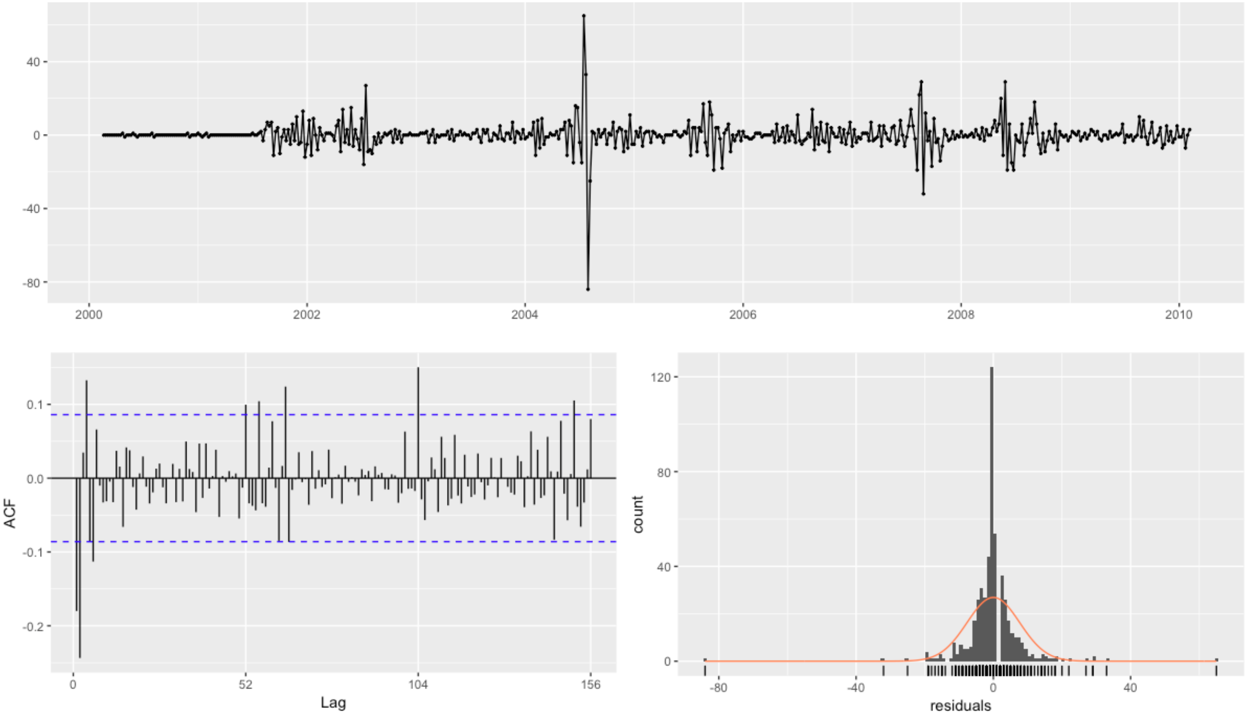

Forecasts from Seasonal naive method
San Juan total cases
400
200
0
-200
1990
1995
2000
2005
2010
Time
level
80
95
Forecasts from Random walk with drift
San Juan total cases
500
250
0
-250
-500
1990
1995
2000
2005
2010
Time
level
80
95

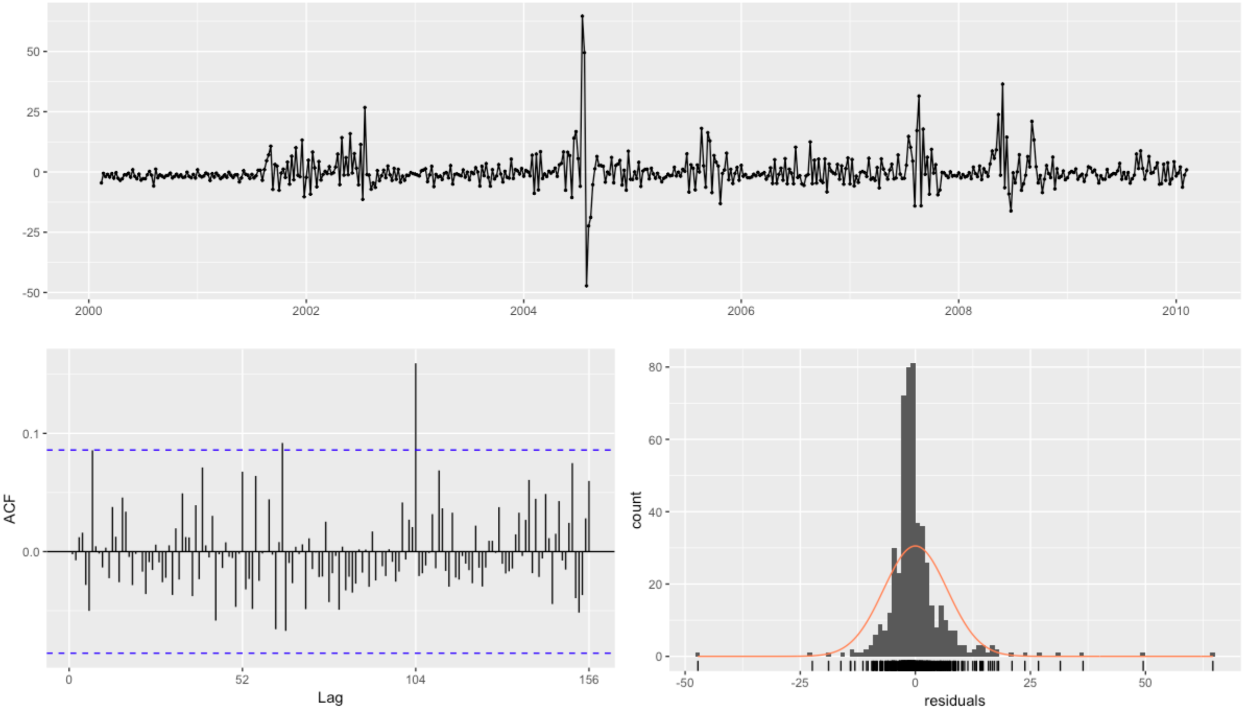
Exhibit 7


Residuals from Naive method


Residuals from Seasonal naive method

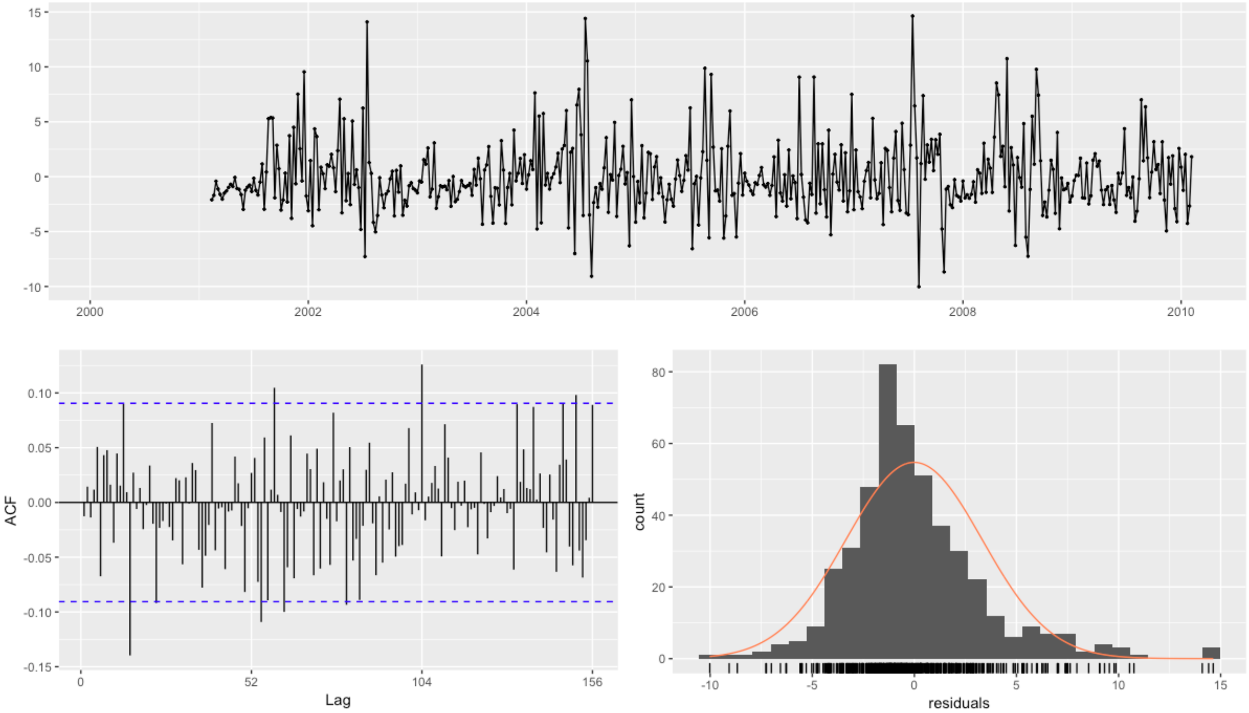


Residuals from Random walk with drift

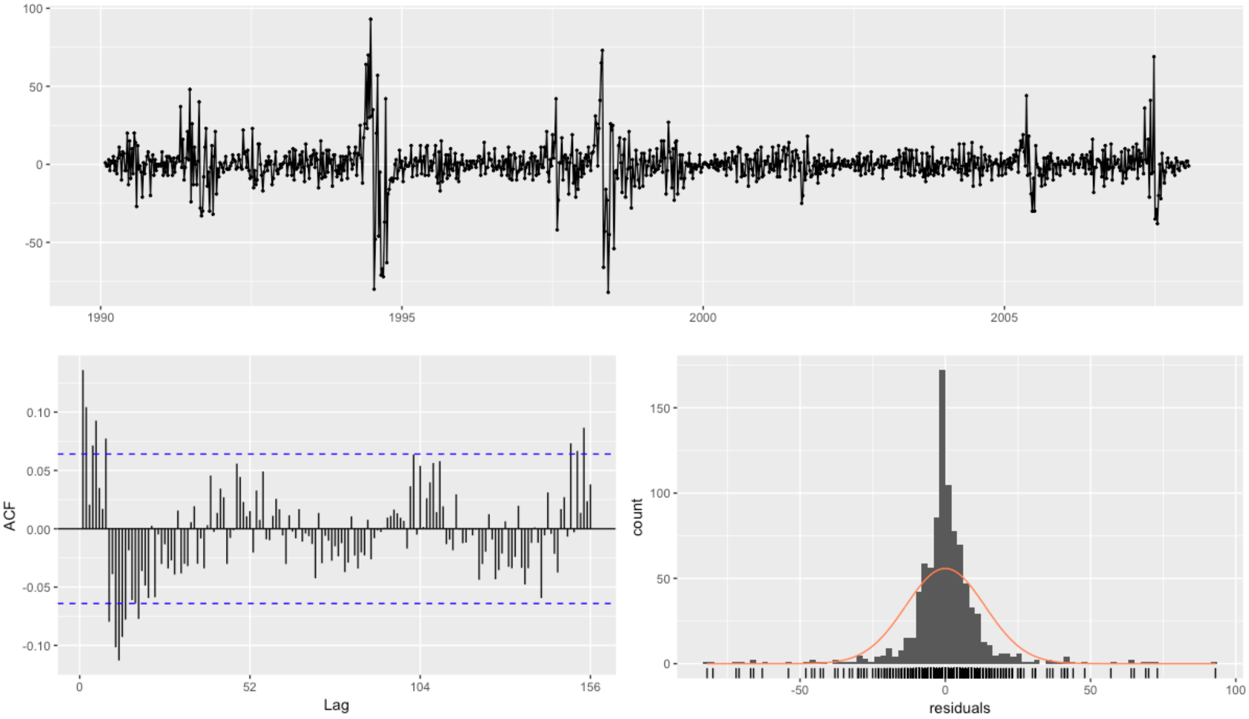


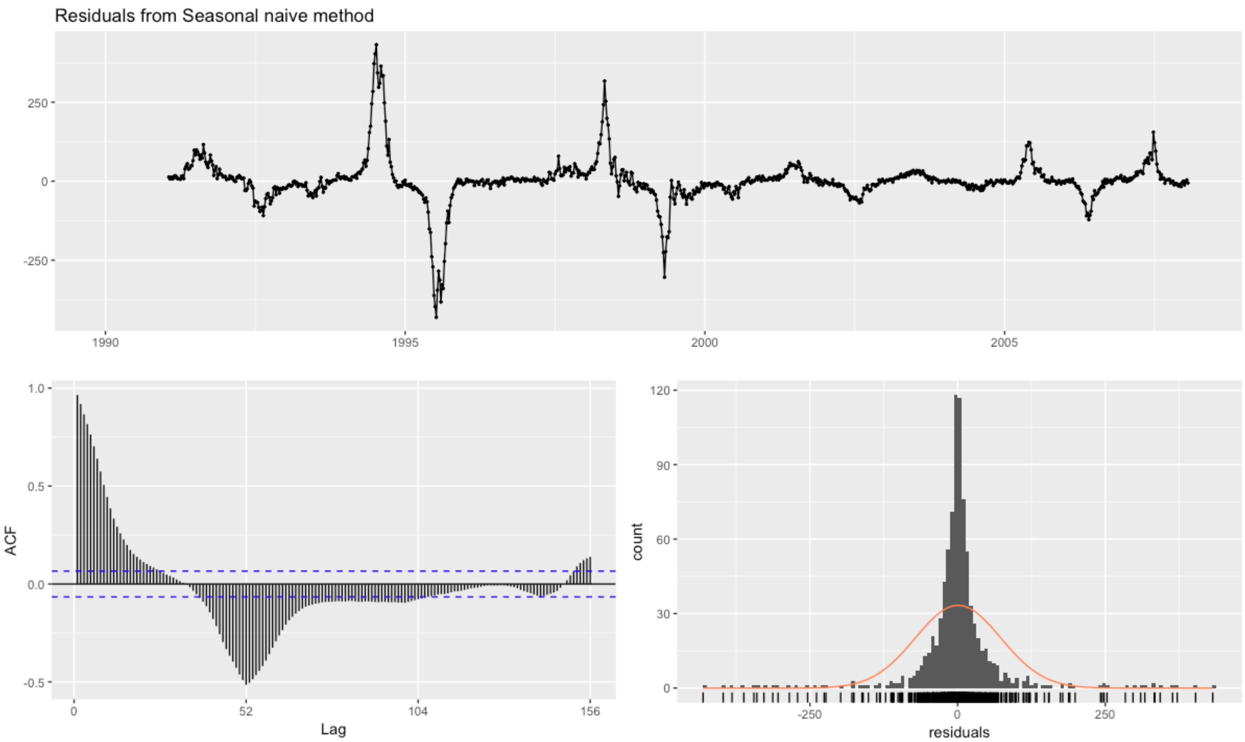
Residuals from Regression with ARIMA(1,0,4) errors

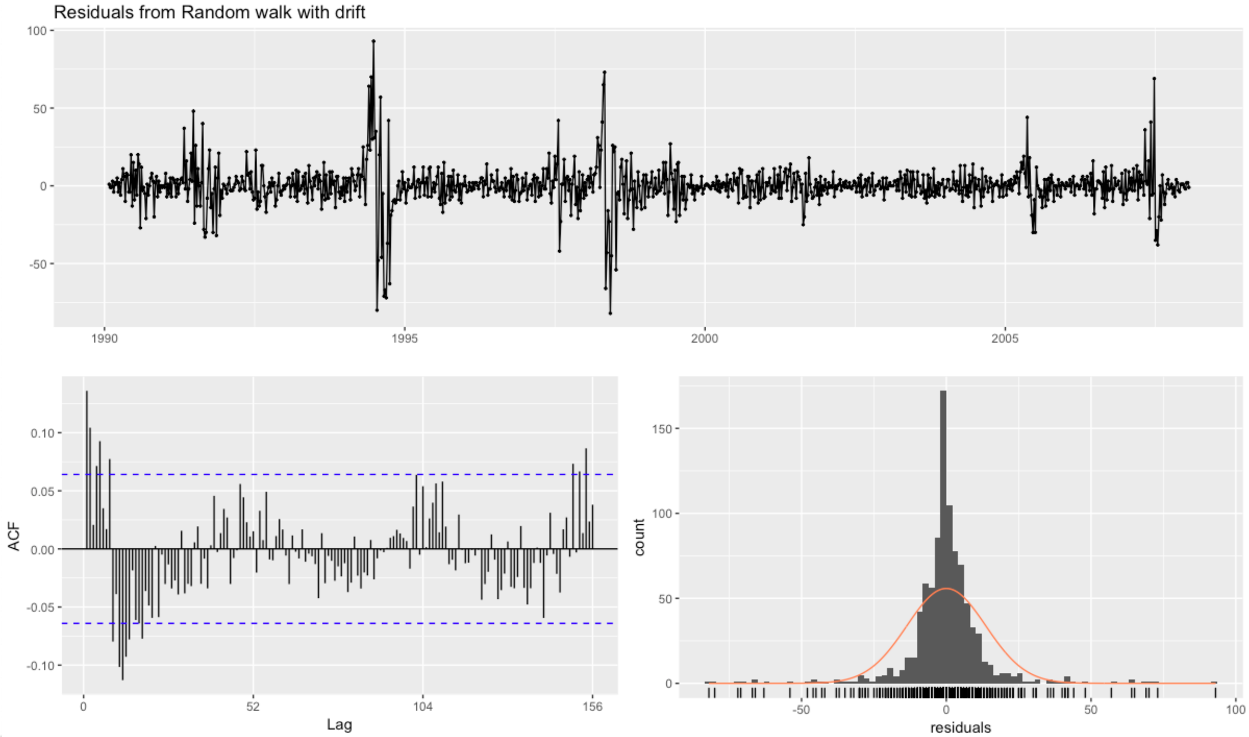
Residuals from NNAR(5,1,6)[52]



Residuals from Naive method




Residuals from Seasonal naive method
250
0
-250
1990
1995
2000
2005
1.0
0.5
0.0
-0.5
ACF
0
52
104
156
Lag
120
90
60
30
0
count
-250
0
250
residuals


Residuals from Random walk with drift
100
50
0
-50
1990
1995
2000
2005
0.10
0.05
0.00
-0.05
-0.10
ACF
0
52
104
156
Lag
150
100
50
0
count
-50
0
50
100
residuals

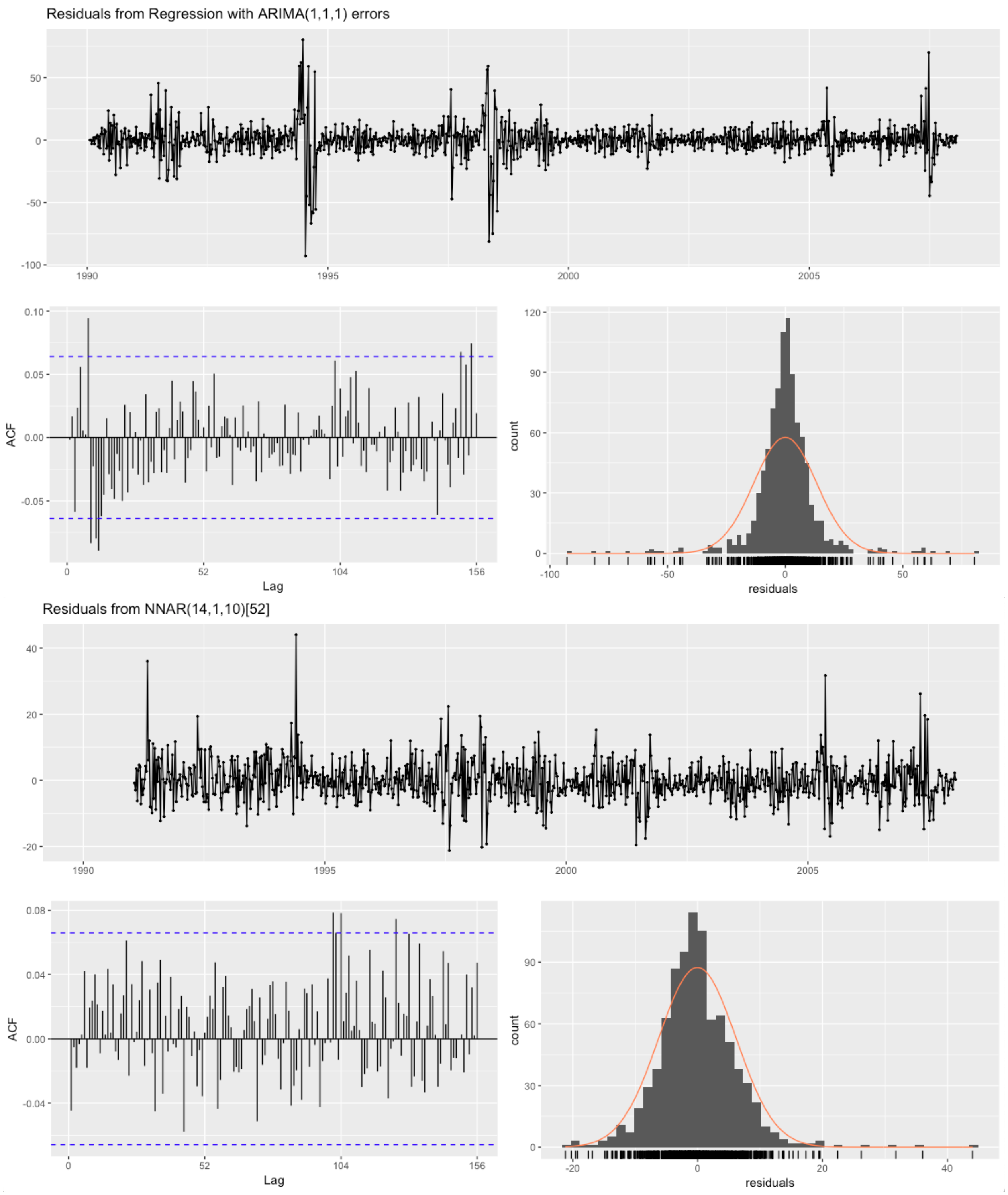
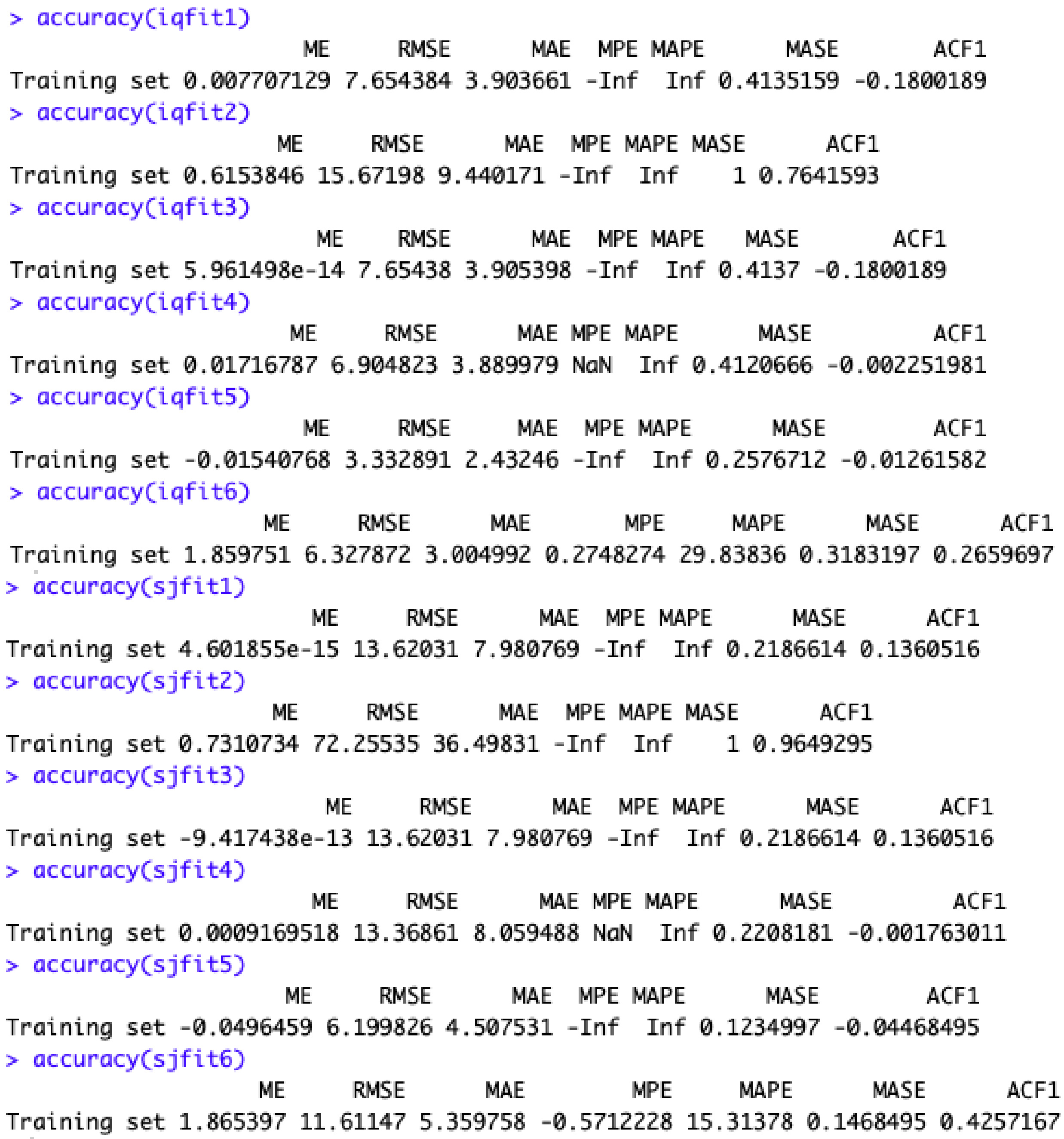
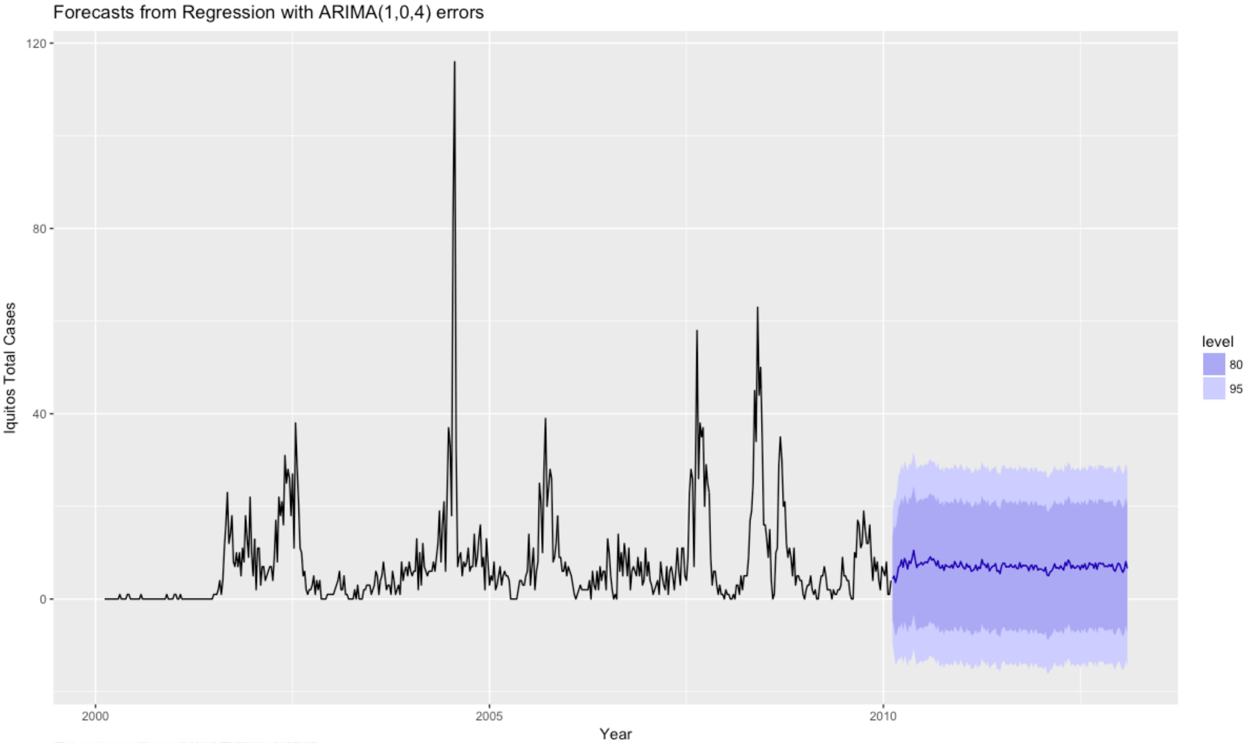

Residuals from Regression with ARIMA(1,1,1) errors
50
0
-50
-100
1990
1995
2000
2005
ACF
0.10
0.05
0.00
-0.05
0
52
104
156
Lag
count
120
90
60
30
0
-100
-50
0
50
residuals
Residuals from NNAR(14,1,10)[52]
40
20
0
-20
1990
1995
2000
2005
ACF
0.08
0.04
0.00
-0.04
0
52
104
156
Lag
count
90
60
30
0
-20
0
20
40
residuals

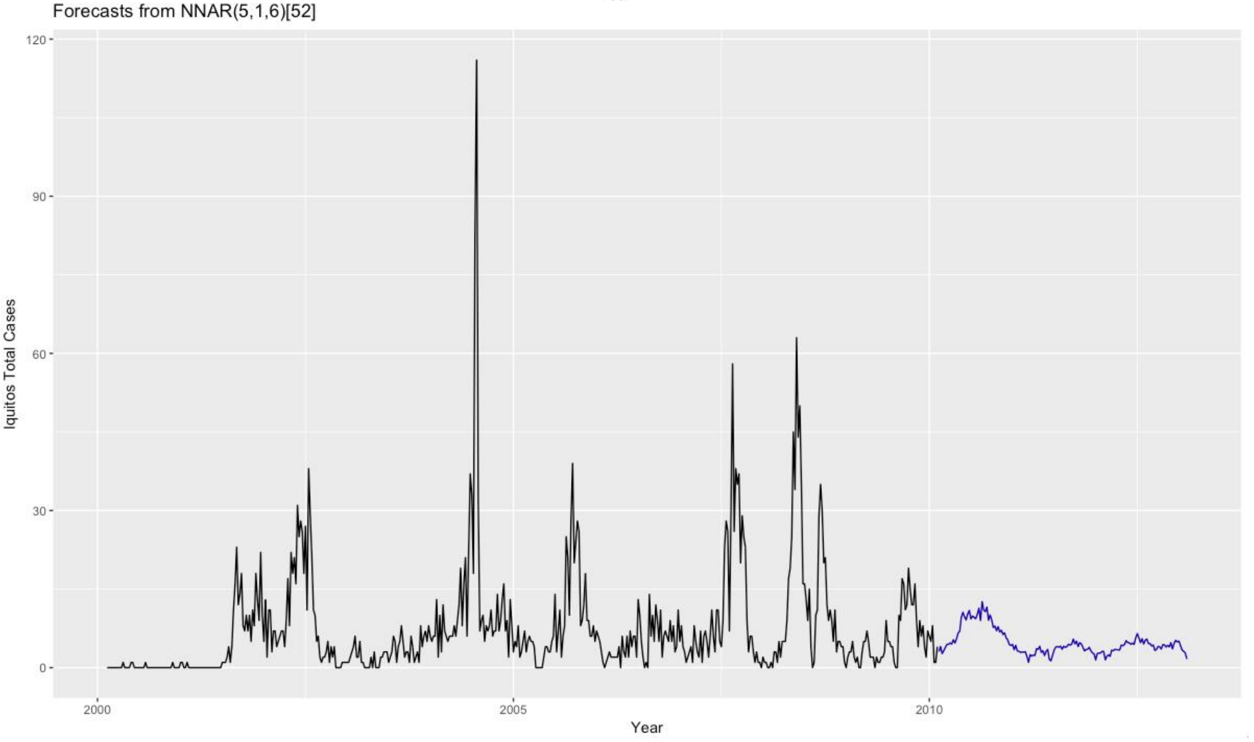
Exhibit 8

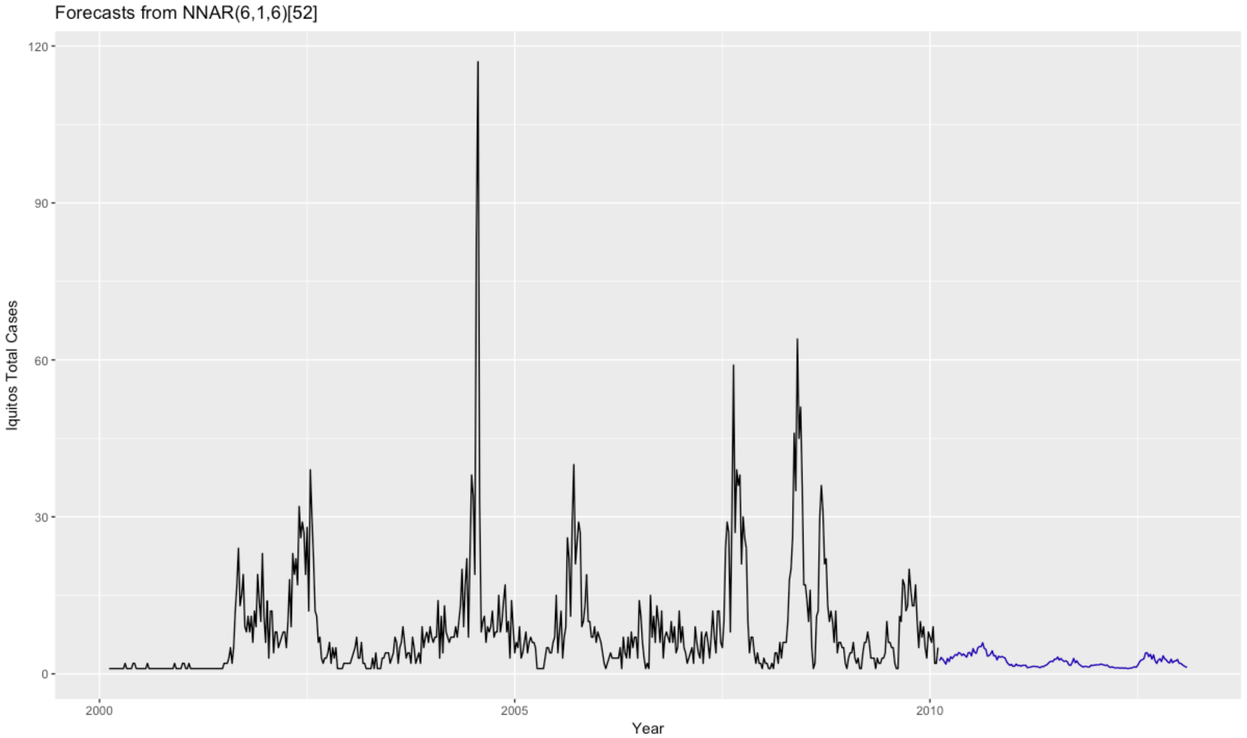


```
> accuracy(iqfit1)
                       ME     RMSE      MAE  MPE MAPE      MASE       ACF1
Training set 0.007707129 7.654384 3.903661 -Inf  Inf 0.4135159 -0.1800189
> accuracy(iqfit2)
                    ME     RMSE      MAE  MPE MAPE MASE      ACF1
Training set 0.6153846 15.67198 9.440171 -Inf  Inf    1 0.7641593
> accuracy(iqfit3)
                       ME    RMSE      MAE  MPE MAPE   MASE       ACF1
Training set 5.961498e-14 7.65438 3.905398 -Inf  Inf 0.4137 -0.1800189
> accuracy(iqfit4)
                     ME     RMSE      MAE MPE MAPE      MASE         ACF1
Training set 0.01716787 6.904823 3.889979 NaN  Inf 0.4120666 -0.002251981
> accuracy(iqfit5)
                      ME     RMSE     MAE  MPE MAPE      MASE        ACF1
Training set -0.01540768 3.332891 2.43246 -Inf  Inf 0.2576712 -0.01261582
> accuracy(iqfit6)
                   ME     RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set 1.859751 6.327872 3.004992 0.2748274 29.83836 0.3183197 0.2659697
> accuracy(sjfit1)
                       ME     RMSE      MAE  MPE MAPE      MASE      ACF1
Training set 4.601855e-15 13.62031 7.980769 -Inf  Inf 0.2186614 0.1360516
> accuracy(sjfit2)
                    ME     RMSE      MAE  MPE MAPE MASE      ACF1
Training set 0.7310734 72.25535 36.49831 -Inf  Inf    1 0.9649295
> accuracy(sjfit3)
                        ME     RMSE      MAE  MPE MAPE      MASE      ACF1
Training set -9.417438e-13 13.62031 7.980769 -Inf  Inf 0.2186614 0.1360516
> accuracy(sjfit4)
                       ME     RMSE      MAE MPE MAPE      MASE         ACF1
Training set 0.0009169518 13.36861 8.059488 NaN  Inf 0.2208181 -0.001763011
> accuracy(sjfit5)
                     ME     RMSE      MAE  MPE MAPE      MASE        ACF1
Training set -0.0496459 6.199826 4.507531 -Inf  Inf 0.1234997 -0.04468495
> accuracy(sjfit6)
                   ME     RMSE      MAE        MPE     MAPE      MASE      ACF1
Training set 1.865397 11.61147 5.359758 -0.5712228 15.31378 0.1468495 0.4257167
```

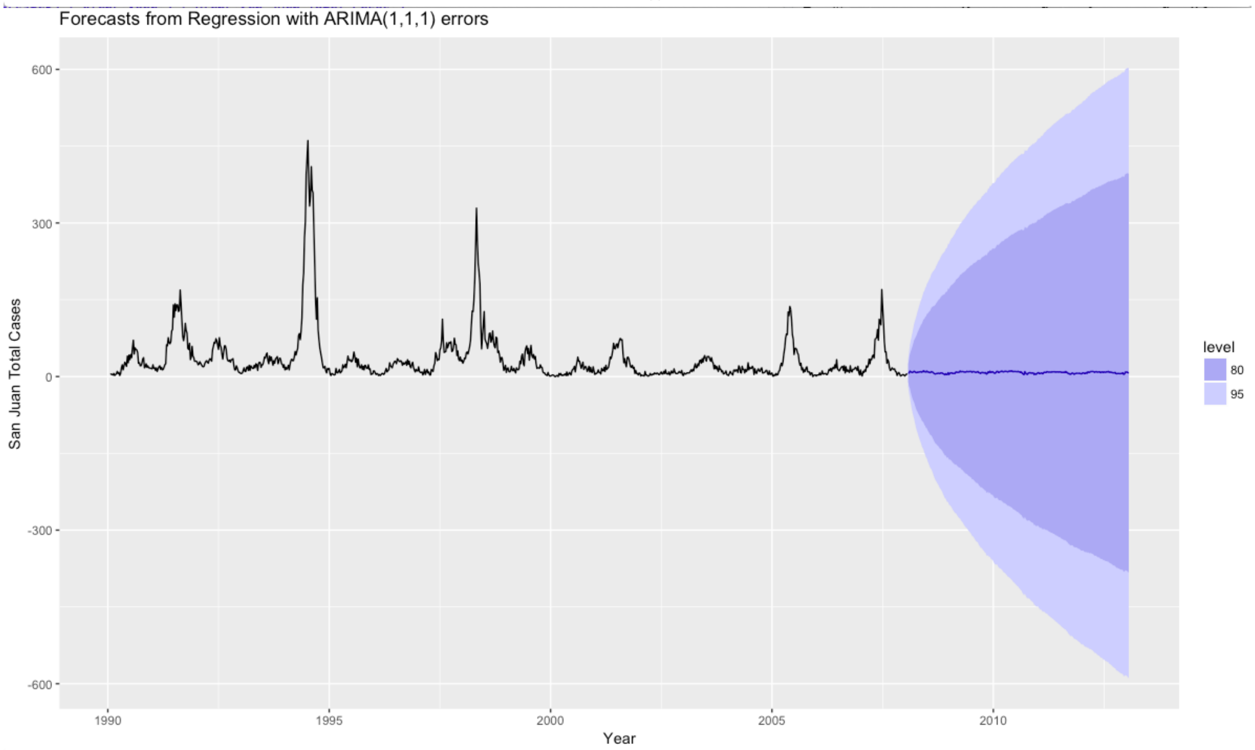
Exhibit 9


Forecasts from Regression with ARIMA(1,0,4) errors
Iquitos Total Cases
120
80
40
0
2000
2005
2010
Year
level
80
95

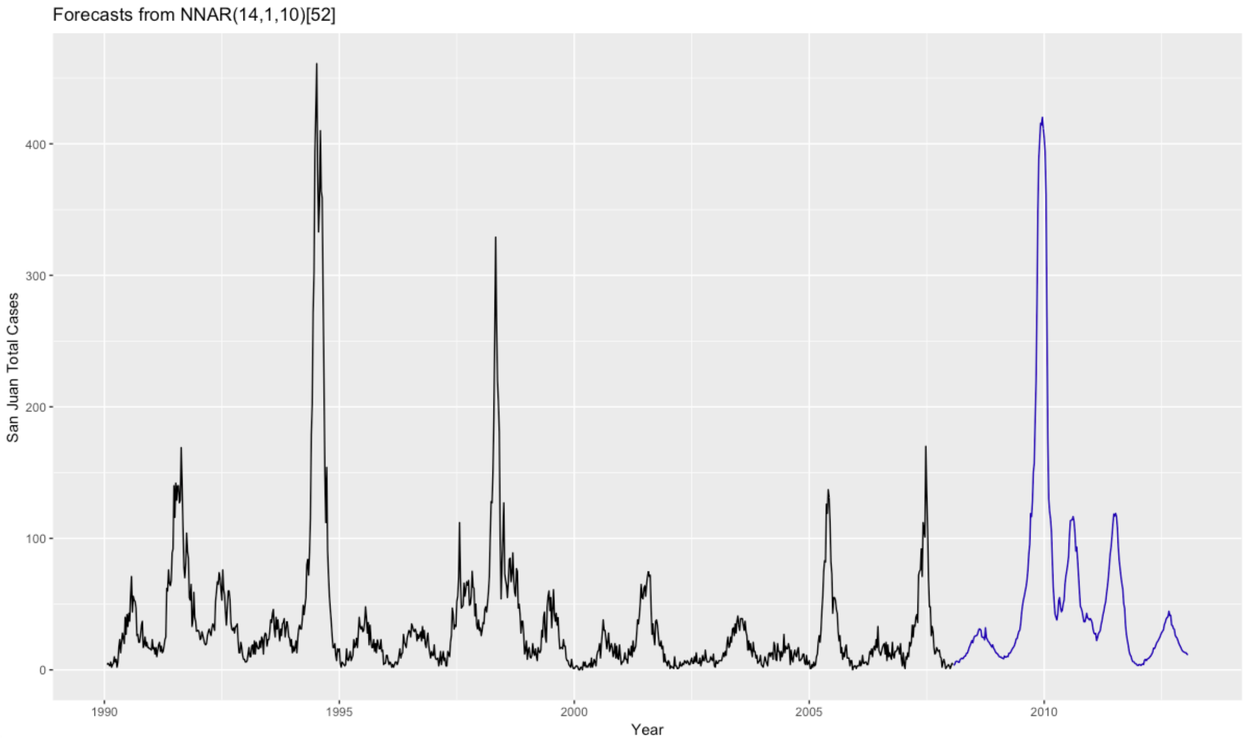

Forecasts from NNAR(5,1,6)[52]
Iquitos Total Cases
120
90
60
30
0
2000
2005
2010
Year

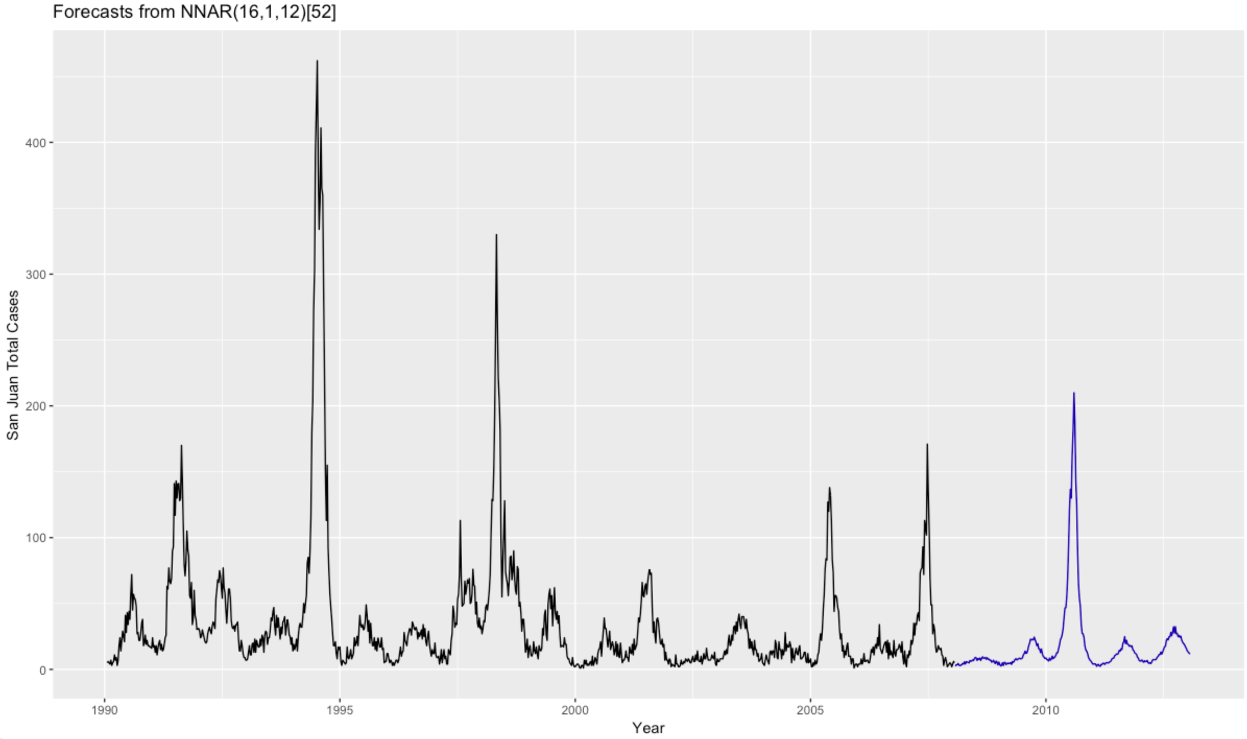


Forecasts from NNAR(6,1,6)[52]



Forecasts from Regression with ARIMA(1,1,1) errors


Forecasts from NNAR(14,1,10)[52]
San Juan Total Cases
400
300
200
100
0
1990
1995
2000
2005
2010
Year


Forecasts from NNAR(16,1,12)[52]
San Juan Total Cases
400
300
200
100
0
1990
1995
2000
2005
2010
Year