1. **Introduction**

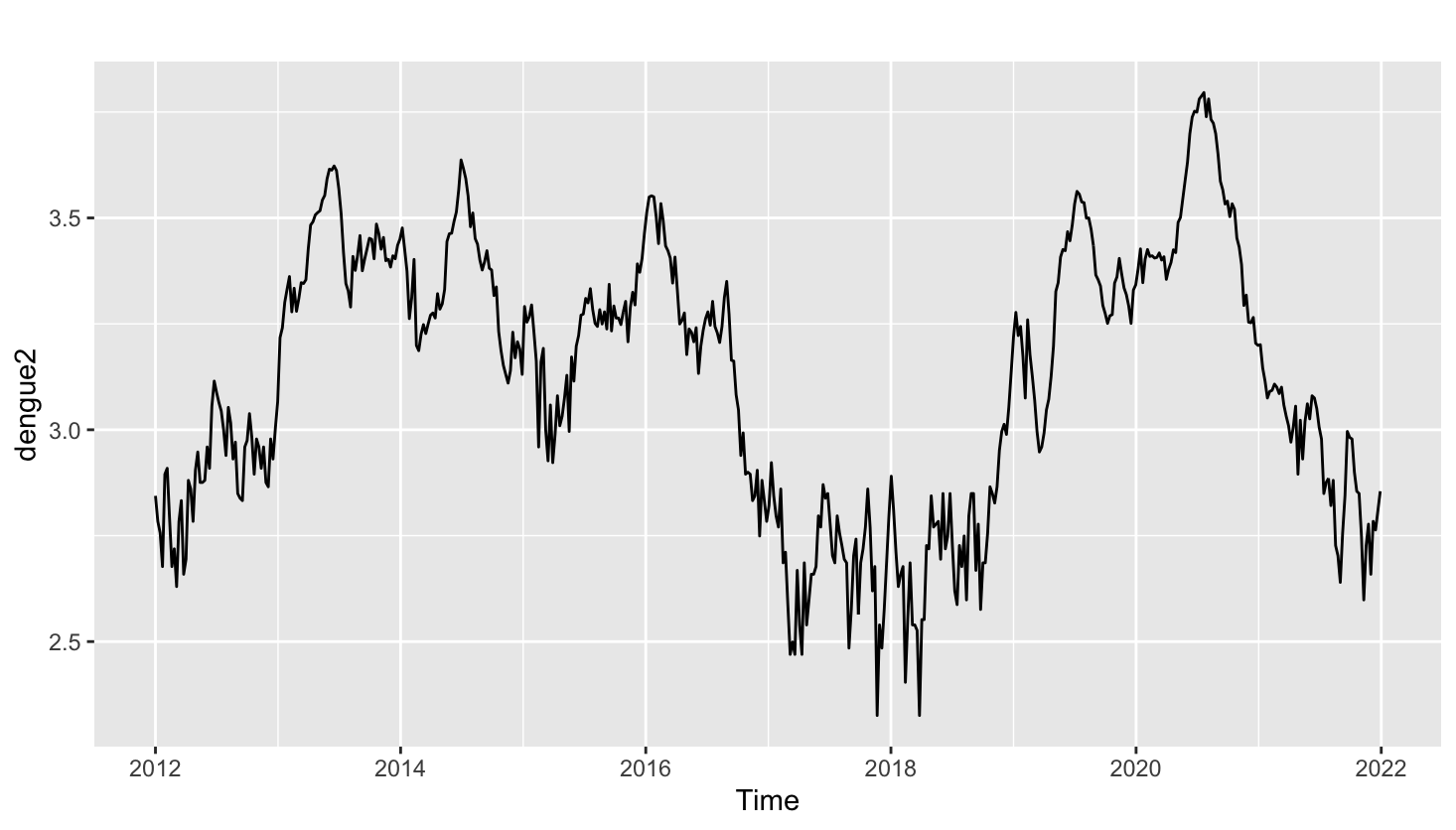
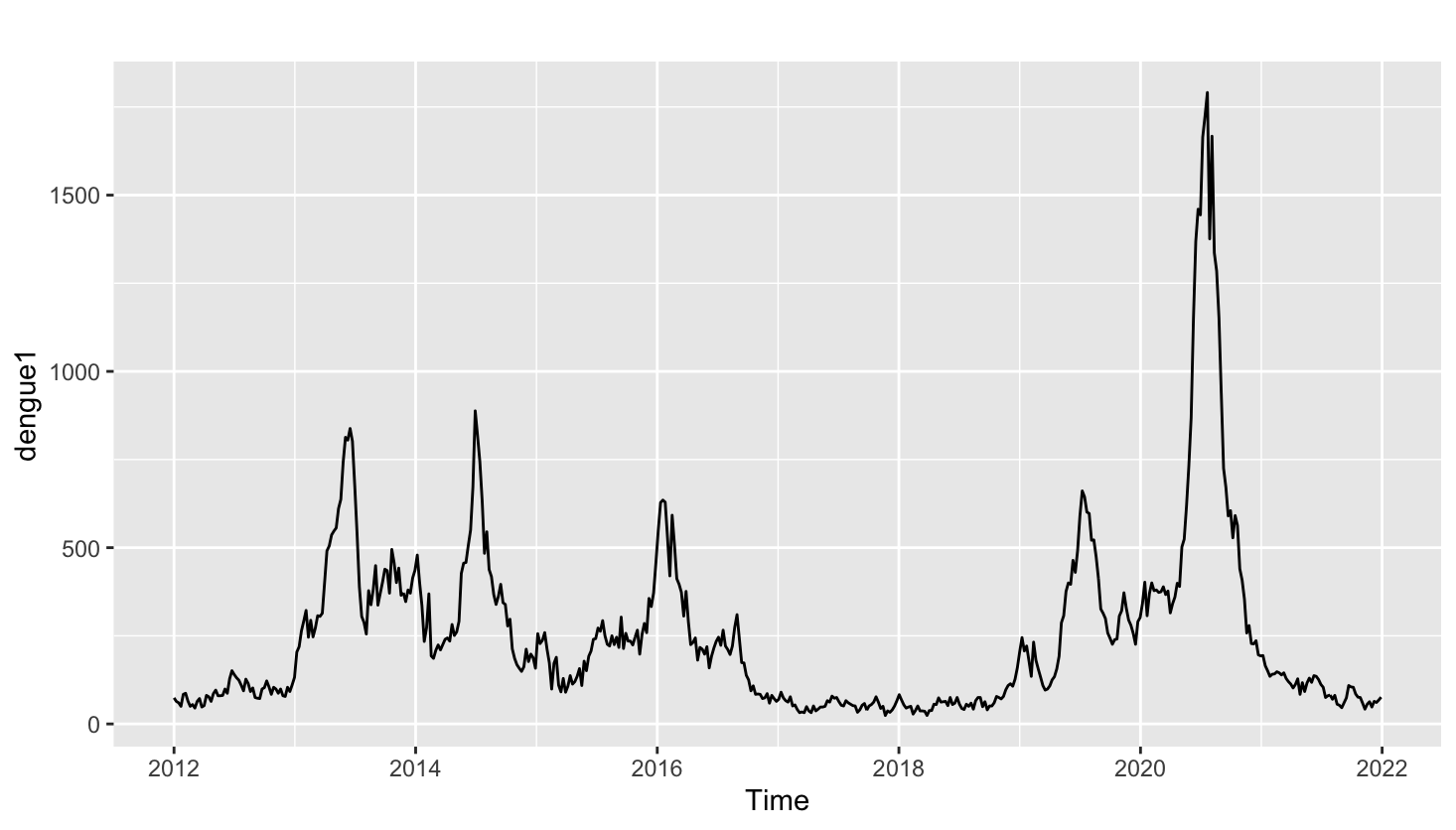
Dengue fever is a viral disease transmitted to humans through the bite of an infected mosquito. It is usually found in tropical and subtropical climates worldwide, most prominently in Latin American and Asian countries. Though it usually leads to having mild symptoms, it is known to develop into acute flu-like illness, which requires medical care and has no specific treatment. In Singapore, as of March 2022, there have been outbreaks and an increase in dengue cases that concern the National Environment Agency, reporting more than 260 cases per week.

Given the preference for tropical climates of the mosquitoes, high temperatures and rainfall precipitation have been identified as key climatic variables that could provide guidance as to when and where the outbreaks could occur. The NEA indicates that the combination of rainfall, temperature, working from home and the surge in the population of the Aedes aegypti mosquito, a species prone to carrying the virus, will all lead to an increase in cases. To prevent the public health hazard, from a policy perspective, the ENA’s myENV app is a powerful tool capable of notifying the population of where there are higher populations of mosquitoes. Thus, adding predictor models could be useful in terms of preparation and prevention of infections. Even further, these predictor models could also indicate where it would be optimal to distribute dengue preparedness packages as well as implement measures to eliminate potential mosquito breeding habitats.

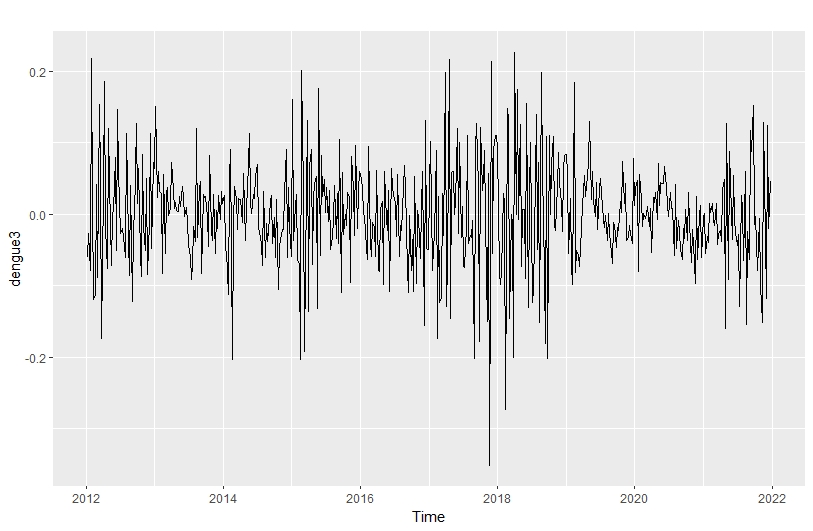
It is for this reason that our hypothesis is that there is a positive correlation between rainfall, temperature and dengue cases with a seasonal pattern in the data. Therefore, we intend to create a time series model that will allow us to visualise and confirm the hypothesis, in addition to being able to forecast areas with high dengue incidence based on precipitation and temperature levels.

1. **Data Processing and Transformation**

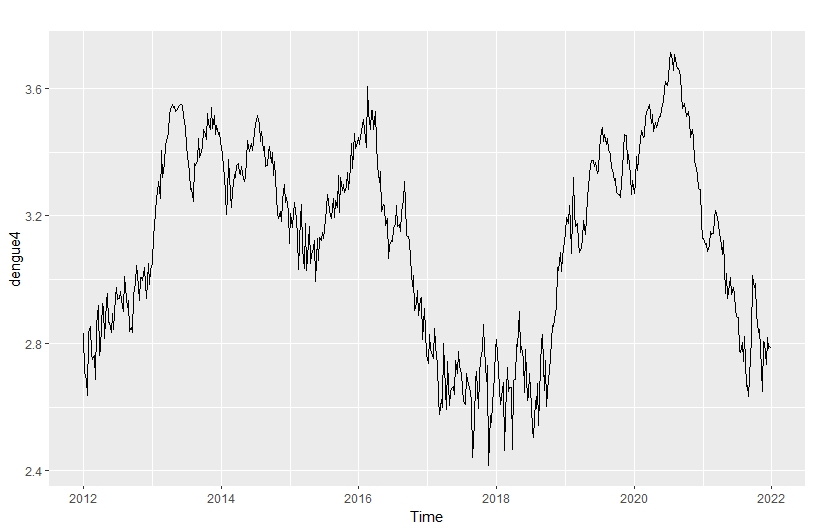
The data for dengue cases is a combination of three different datasets from the Ministry of Health and the Singapore government through data.gov.sg. They include weekly number of Dengue and Dengue Hemorrhagic Fever cases from 2012-2019, weekly infectious diseases from 2020-2021 with a filter for dengue cases, as well as the weekly infectious diseases for 2019, again only including dengue cases. After combining the data and transforming it into a TS object, it was evident that the data was not stationary, therefore, a BoxCox transformation with a lambda value was performed.



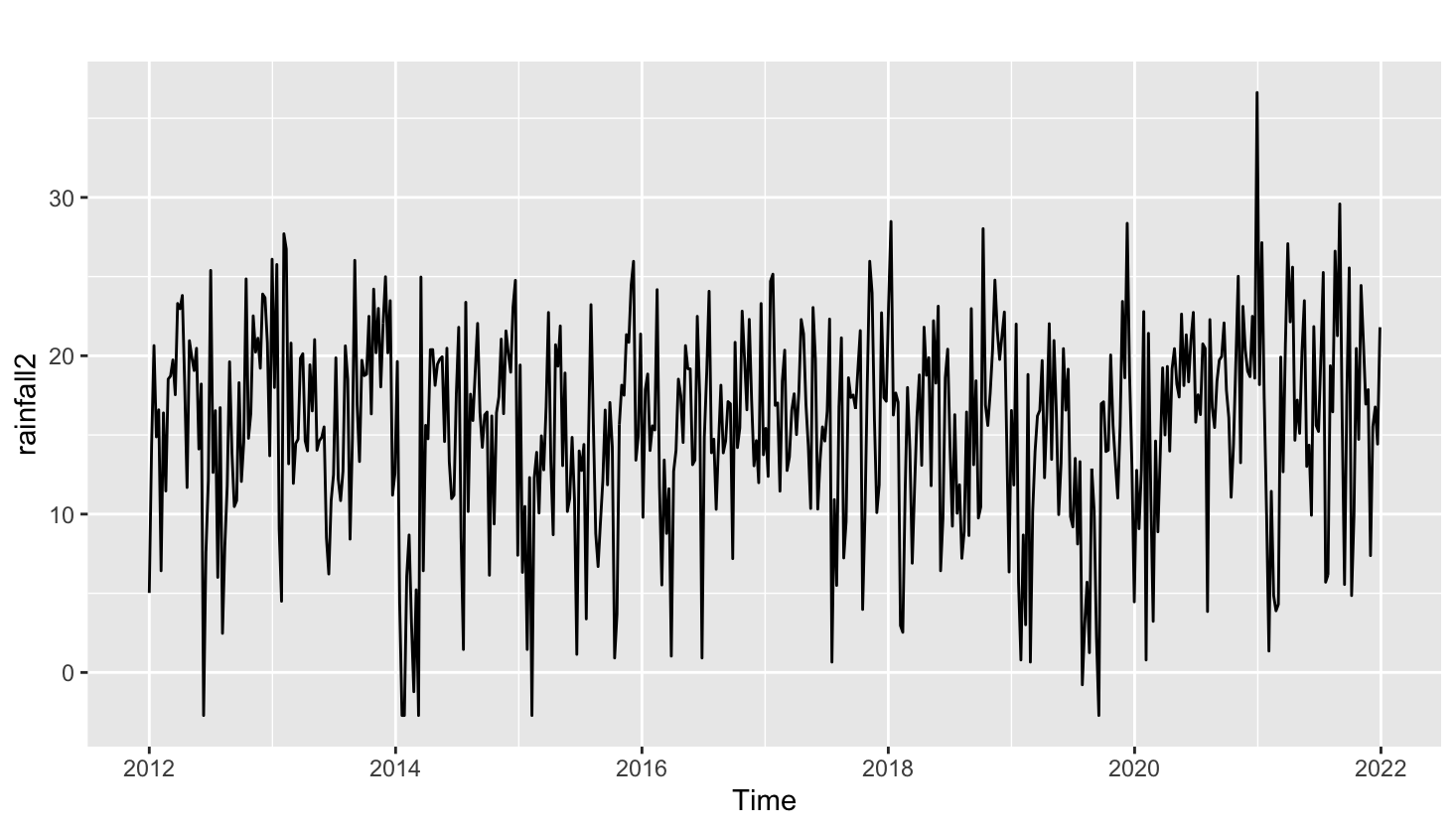
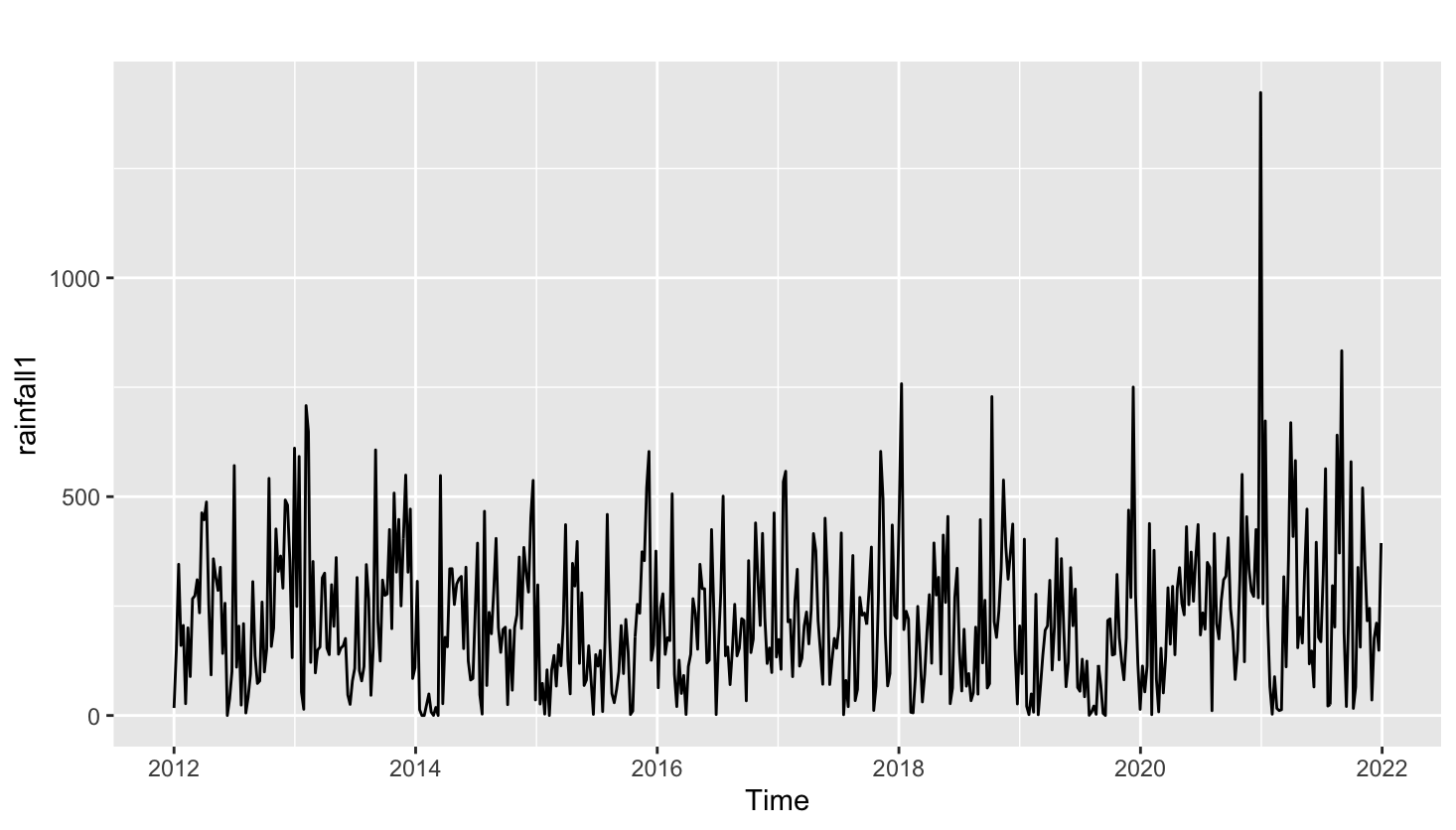
Even further, after computing nsdiffs and ndiffs, the result was that 1 seasonal difference was required, therefore, we applied a first difference to the data using the diff R function. Ultimately, the transformations led to having a stationary time series.



Given that rainfall and temperature have a seasonal nature, it is useful to remove the trend component using the mstl function in order to model using the seasonal component only.

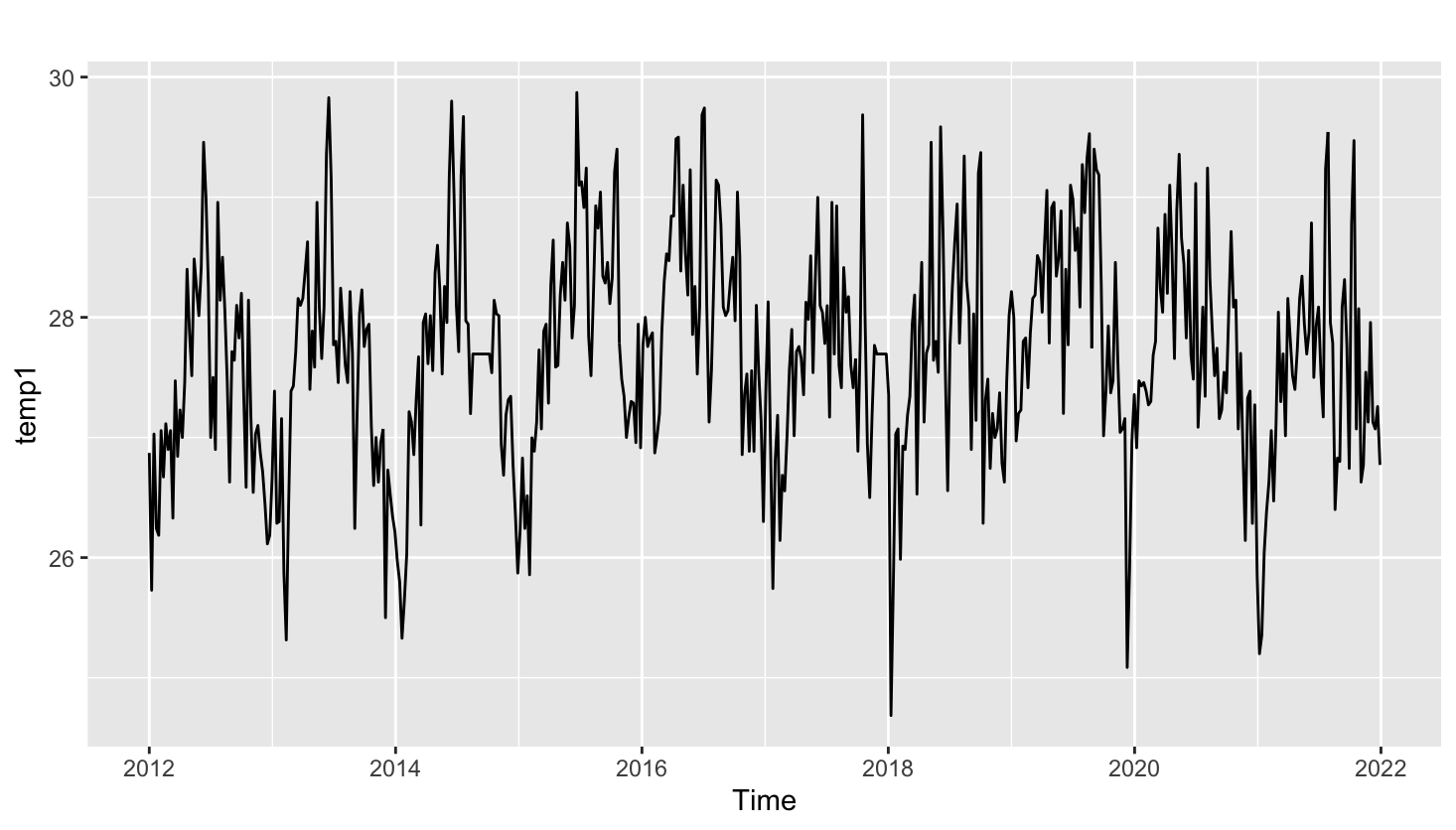


For the data on precipitation, we obtained the information from the Meteorological Service of Singapore. The daily rainfall in mm from January 2012 to January 2022 in the North (Admiralty), South (Marina Barrage), East (Pasir Ris Central), West (Tuas) and Central (Toa Payoh) regions was combined into a single dataset. Given that there were N.A. values in the data, in order to standardise the information, the seasonal naive method was applied in order to replace the N.A. values with the corresponding last year’s information. After the cleanup and plotting it as a TS object, the data was not stationary, so a BoxCox transformation was also applied.



In order to know whether or not differencing was required, the KPSS test was performed as well as nsdiffs and ndiffs, and the result was that no differencing was needed.

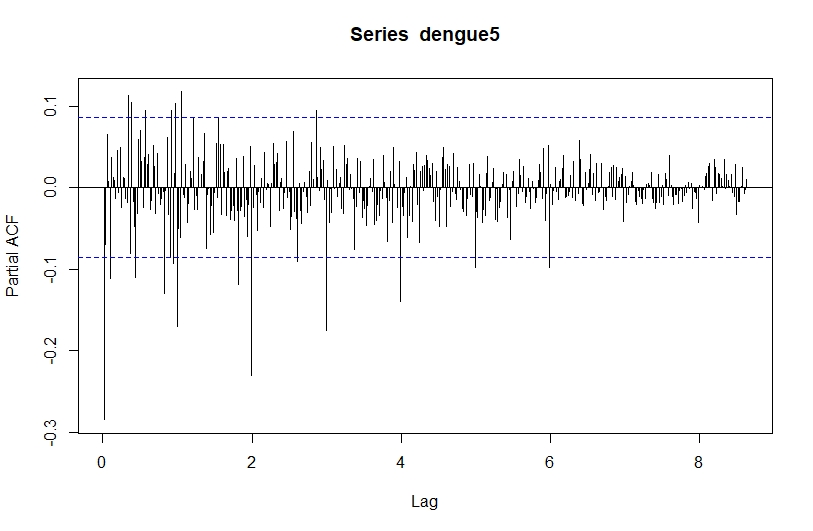
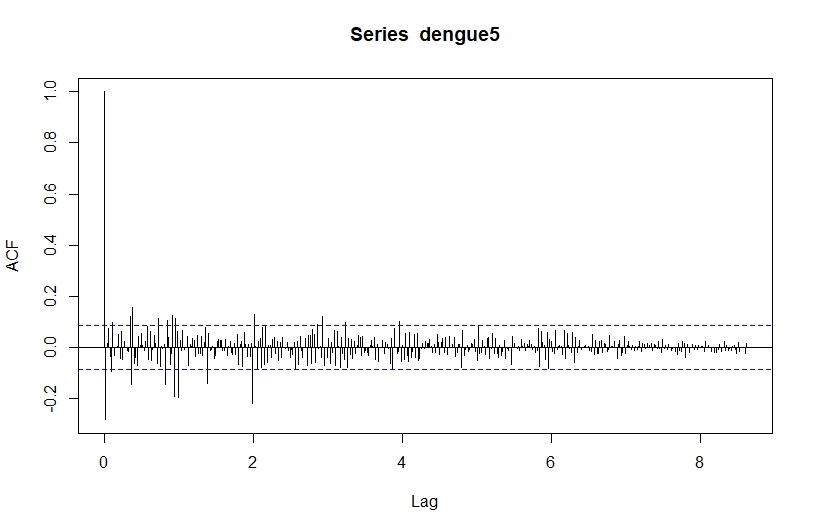
Finally, for the data on temperature, the information was obtained from the same source as precipitation and the same regions were also chosen. Similarly, the N.A. values were replaced with the values from past years using the seasonal naive method. However, the data did not need a BoxCox transformation as there was not much difference between the raw data and the transformed data. Additionally, after running the KPSS test as well as nsdiffs and ndiffs, the results also concluded that no differencing was required.

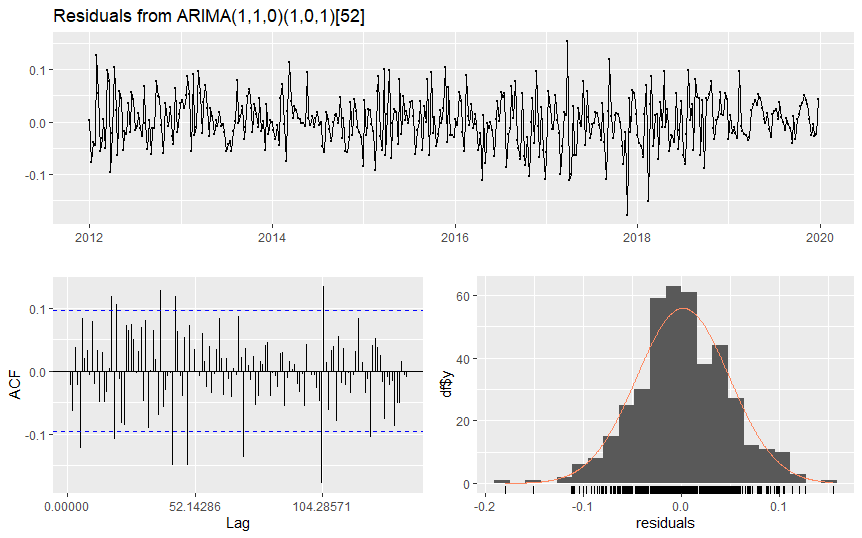


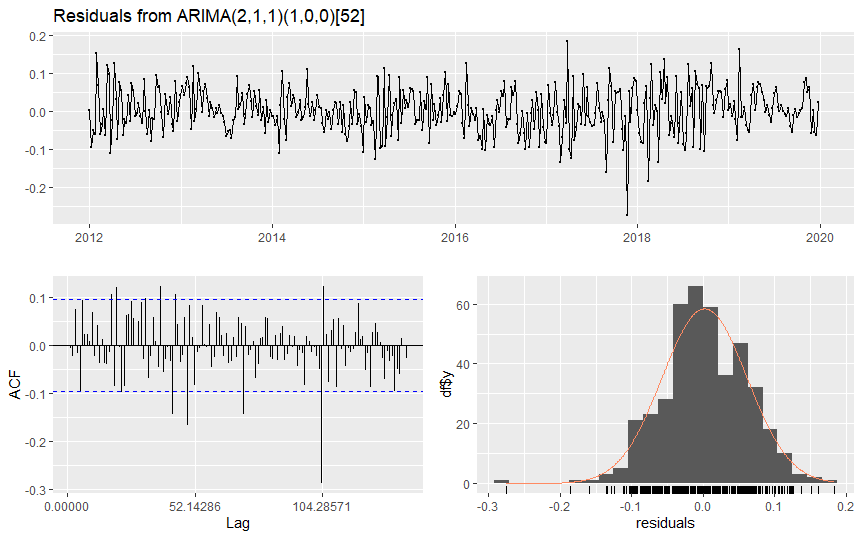
1. **Data Modelling and Methodology**

**3.1 S-ARIMA**

Given the seasonal nature of the data, a seasonal Arima model will be discussed and tested. To do so, an analysis of the ACF and PACF will be conducted. We first extracted the seasonal component using seasadj(mstl(dengue)). Using nsdiffs() and ndiffs() on the seasonal component of Boxcox dengue dataset, a non-seasonal differencing is needed as ndiffs() = 1. After visualizing the ACF and PACF plots, we can see a slow decay over time and exponential decay for the first few lags. The ACF plot also shows non-seasonal spikes after each seasonal lag. Hence this suggests there may be both seasonal AR and non-seasonal AR components. For the PACF plot, we are able to see a gradual decay towards zero hence the presence of MA component possibly seasonal.



The dataset we are using is pre-covid from 2012 to 2019 to train the model as testing will be conducted on 2020 to 2021 dataset. Using auto.arima(), the model is ARIMA(2,1,1)(1,0,0). From the plots, we can deduce that the model is ARIMA(1,1,0)(1,0,1). When comparing these two models, the ARIMA(1,1,0)(1,0,1) model performs better as it has AICc = -1256.05 and RMSE = 0.04720968 while auto.arima has AICc =-1153.7 and RMSE = 0.05934204. Hence with lower RMSE, ARIMA(1,1,0)(1,0,1) performs better. However, the Ljung-Box test shows that ARIMA(1,1,0)(1,0,1) has p-value of 5.919e-05 and ARIMA(2,1,1)(1,0,0) has p-value = 1.279e-05. Both models suggest that there is time-series information left within the residuals. Therefore, both models may not be a good fit for the dengue dataset. Hence, we have to explore ARIMA-X models. The residuals are as shown:

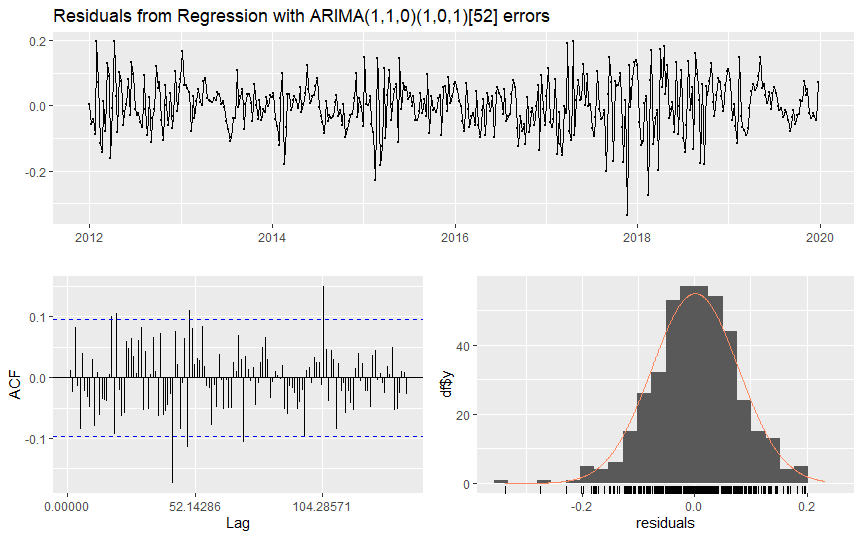


Initially, we tried exploring various combinations of ARIMA models, where we set the seasonal AR component to 1 or 2 and the non-seasonal component to 1 or 2.However, RStudio gives an error for each combination we tried, with the exception of the earlier mentioned ARIMA(1,1,1)(1,0,1)[52] model. Hence we are limited to the auto.arima model as well as the ARIMA(1,1,1)(1,0,1)[52] model.

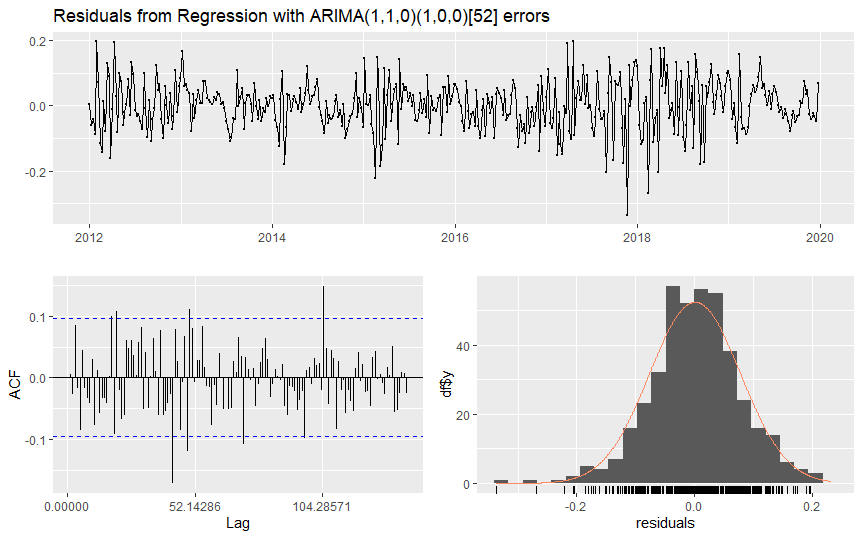
**3.2 ARIMA-X**

To incorporate the exogenous variables such as rainfall and temperature, ARIMA-X model will be used and tested. As previously mentioned, rainfall variable requires Boxcox transformation but no differencing is needed while temperature variable does not require any transformation and no differencing is needed. We will first use auto.arima function to create an ARIMA-X model.

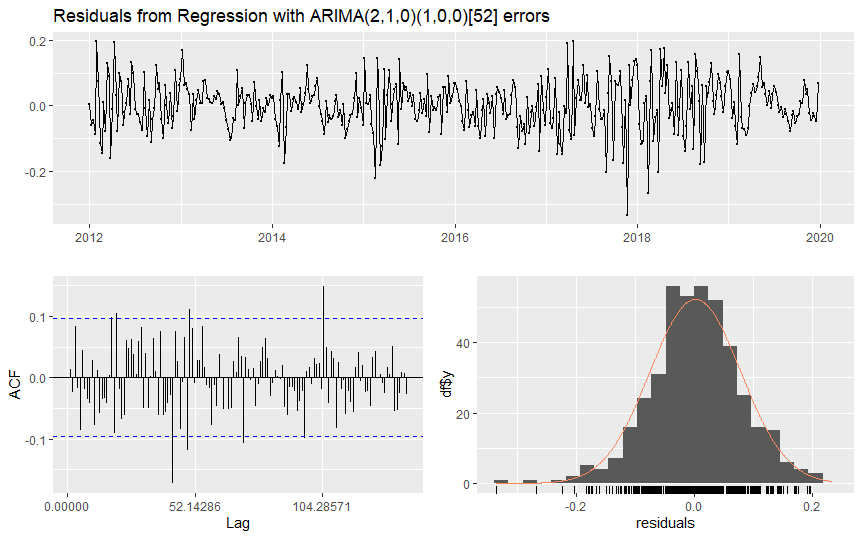
The first model we will be creating is using Boxcox transformed dengue and Boxcox transformed rainfall variable. This results in ARIMA(1,1,0)(1,0,1)[52] model with AICc = -940.29 and RMSE = 0.07701727. The Ljung-Box test shows p-value = 0.000975 hence some time information is left in the time series.



The second model we will be creating is Boxcox transformed dengue and temperature variable. This results in ARIMA(1,1,0)(1,0,0)[52] with AICc = -939.33 and RMSE = 0.07701727. The Ljung-Box test shows a p-value = 0.000975 hence some time information left in the time series.



The last model will be Boxcox transformed dengue, Boxcox transformed rainfall variable and temperature variable. This results in ARIMA(2,1,0)(1,0,0)[52] with AICc = -939.18 and RMSE = 0.07703571. The Ljung-Box test shows a p-value = 0.001022 hence some time information left in the time series.

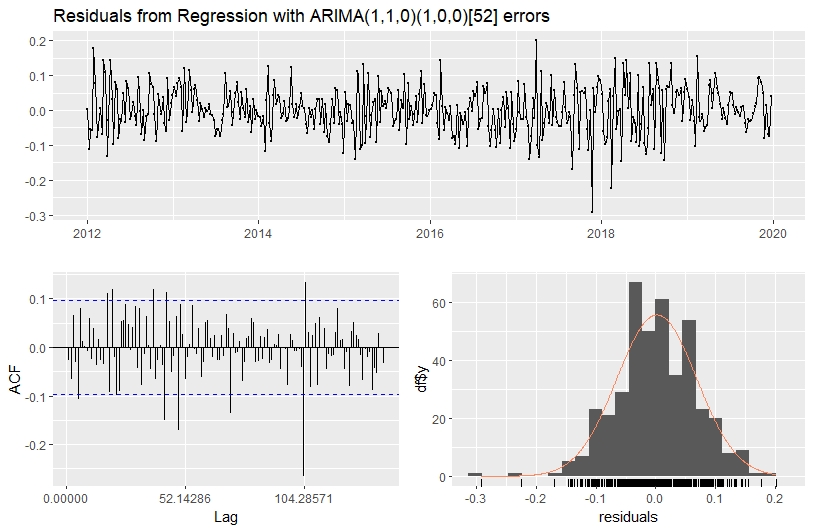


Overall, the last model with rainfall and temperature variable is the best fit among the three models due to the highest AICc value while having similar RMSE and highest p-value. However, it is disappointing that all three models did not pass the Ljung-Box test suggesting that all models are not a good fit. This can be possibly due to other variables that we have not considered such as population density and Covid-19 impact.

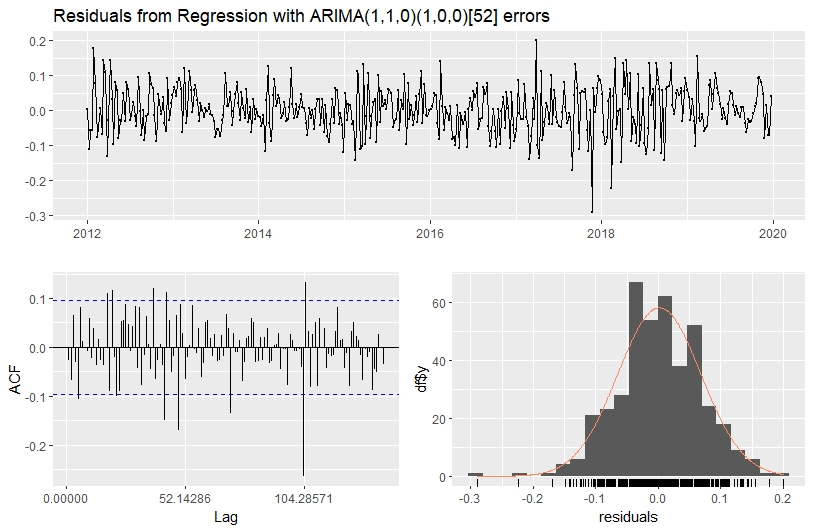
**3.3 SARIMA-X**

For SARIMA-X, we use the seasonal component of the dengue dataset similar to S-ARIMA model but we incorporate exogenous variables such as rainfall and temperature into the model. Similar to ARIMA-X, we will be using auto.arima function to create three models.

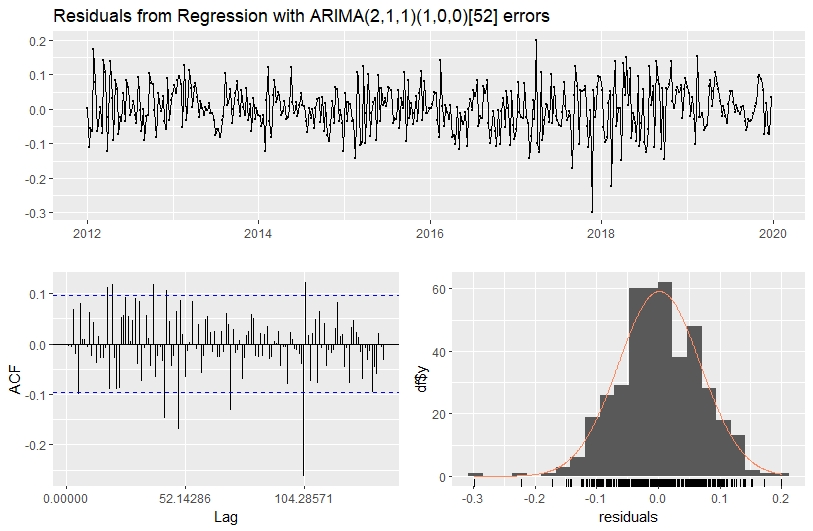
The first model we will be creating is using Boxcox transformed dengue and Boxcox transformed rainfall variable. This results in ARIMA(1,1,0)(1,0,0)[52] model with AICc = -1060.37 and RMSE = 0.0666133. The Ljung-Box test shows p-value = 2.281e-05 hence some time information is left in the time series.



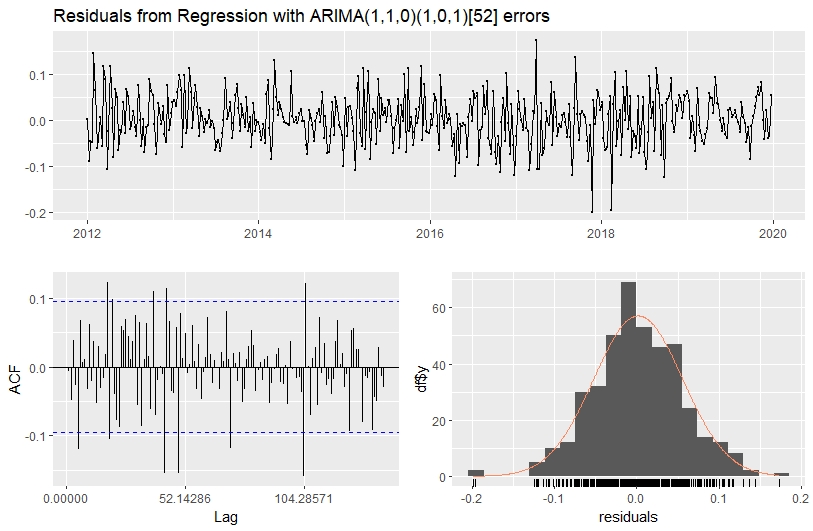
The second model we will be creating is Boxcox transformed dengue and temperature variable. This results in ARIMA(1,1,0)(1,0,0)[52] with AICc = -1060.3 and RMSE = 0.06661923. The Ljung-Box test shows a p-value = 2.104e-05 hence some time information left in the time series.



The third model will be Boxcox transformed dengue, Boxcox transformed rainfall variable and temperature variable. This results in ARIMA(2,1,1)(1,0,0)[52] with AICc = -1056.02 and RMSE = 0.06648509. The Ljung-Box test shows a p-value = 4.594e-05 hence some time information left in the time series.



The last model will be Boxcox transformed dengue, Boxcox transformed rainfall variable and temperature variable but using parameters from S-ARIMA instead of auto.arima function. This results in ARIMA(1,1,0)(1,0,1)[52] with AICc = -1143.76 and RMSE = 0.05401304. The Ljung-Box test shows a p-value = 0.0003318 hence some time information left in the time series.



Overall, the last model performs the best as we use parameters from the S-ARIMA model. However, the results are overall disappointing as all models did not pass the Ljung-Box test implying that the models are not a good fit for the dataset. As earlier mentioned when discussing the limitations of the ARIMA-X models, other variables such as population density should be included to further improve the model. We will now be looking at Simple Feed Forward Neural Network to see if the model is able to perform better than ARIMA models.

**3.4 Feed Forward Neural Network**

Machine learning models have always been debated to perform better than ARIMA models in value forecasting. We used a simple feed forward neural network since it produces relatively accurate forecasts despite being simple to use. It helps to learn the relationship between the independent variables which serve as inputs to the network, and dependent variables that are designed as outputs of the network.

Firstly, since nsdiffs=0, there is no long term trend but there may still be intra-year seasonal patterns, and specified p=12, with 12 input nodes of lag 1 to 11. Number of hidden layers is set to default and P is allowed to range above 1 such that the lowest RMSE is obtained. P=3 obtained the lowest test RMSE, with input lags 12, 24, 36. There are a total of 14 input nodes. Train RMSE of 0.0276 was obtained and test RMSE was 0.251. P-value was 0.701, passing all three levels of significance. We are able to detect the downward pattern in 2021, but failed to detect a rising trend during Covid, which is expected.

Method 2 included passing external regressors like rainfall and temperature using xreg, which might help in predicting the Covid trend. P=2 gave the lowest test RMSE of 0.207 and train RMSE of 0.030. P-value was 0.738 which passed all three levels of significance. By comparing both methods’ out-of-sample performance, method 2 yielded lower RMSE and was a better model. Seems like including external regressors like rainfall and temperature did help the neural network to recognise some kind of Covid trend.

1. **Testing and Accuracy**

Both in-sample and out-of-sample testing was done for all models. The in-sample time period used was weekly data between the weeks of 1 Jan 2019 and 22 Dec 2019. The out-of-sample time period used for testing was weekly data between the weeks of 29 Dec 2019 and 1 Jan 2022.

**In-sample performance:**

|  |  |  |
| --- | --- | --- |
| Model | auto.arima | Manual |
| SARIMA | (2,1,1)(1,0,0) | (1,1,0)(1,0,1) |
| AICc | -1153.7 | -1256.05 |
| p-value | 1.279e-05 | 0.0472096 |
| RMSE | 0.05934204 | 5.919e-05 |

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Rainfall Only | Temperature Only | Both Variables (auto.arima) |
| ARIMAX | (1,1,0)(1,0,1) | (1,1,0)(1,0,0) | (2,1,0)(1,0,0) |
| AICc | -940.29 | -939.33 | -939.18 |
| p-value | 0.0007524 | 0.000975 | 0.001022 |
| RMSE | 0.0770174 | 0.07701727 | 0.0770357 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Rainfall Only | Temperature Only | Both Variables  (auto.arima) | Both Variables  (Manual) |
| SARIMAX | (1,1,0)(1,0,0) | (1,1,0)(1,1,0) | (2,1,1)(1,0,0) | (1,1,0)(1,0,1) |
| AICc | -1060.37 | -1060.3 | -1056.02 | -1143.76 |
| p-value | 2,281e-05 | 2.04e-05 | 4.594e-05 | 0.0003318 |
| RMSE | 0.0666133 | 0.06661923 | 0.066485 | 0.054013 |

|  |  |  |
| --- | --- | --- |
| Model | Without External Regressor | With External Regressor (xreg) |
| Feed forward  Neural Network | NNAR(p=12, P=3, k) | NNAR(p=12, P=2, k) |
| Train RMSE | 0.02758349 | 0.03070044 |
| Test RMSE | 0.25096222 | 0.20764560 |
| p-value | 0.7017 | 0.738 |

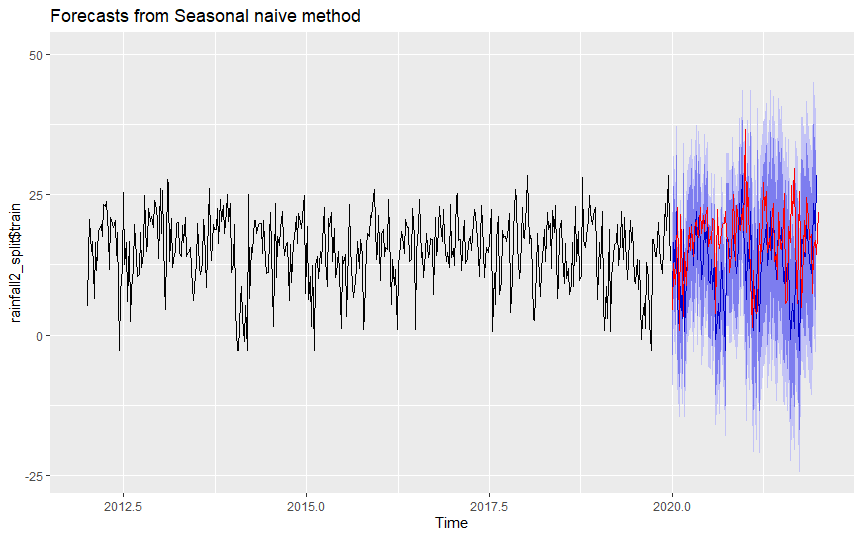
From the above tables, it is quite evident that the Feed Forward Neural Network Models are a much better fit for our in-sample data, with p-values far greater than 0.05, as compared to the S-ARIMA models.

**Out-of-sample Forecasting:**

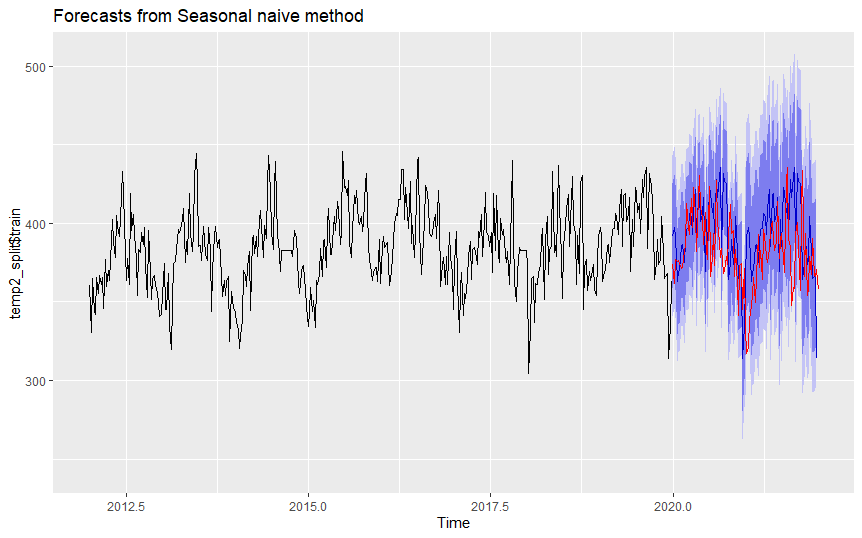
We then utilise our models to perform out-of-sample forecasting. Even though the in-sample forecasts have proven that the S-ARIMA-X models were not a good fit for the data, we will still utilise the S-ARIMA-X Models, as well as the Feed Forward Neural Network Models to forecast for Dengue cases in Singapore. For the S-ARIMA-X models, we have only utilise the S-ARIMA-X models which take into account both rainfall and temperature as xreg variables.

**S-ARIMA-X forecasting(Seasonal Naive Method):**

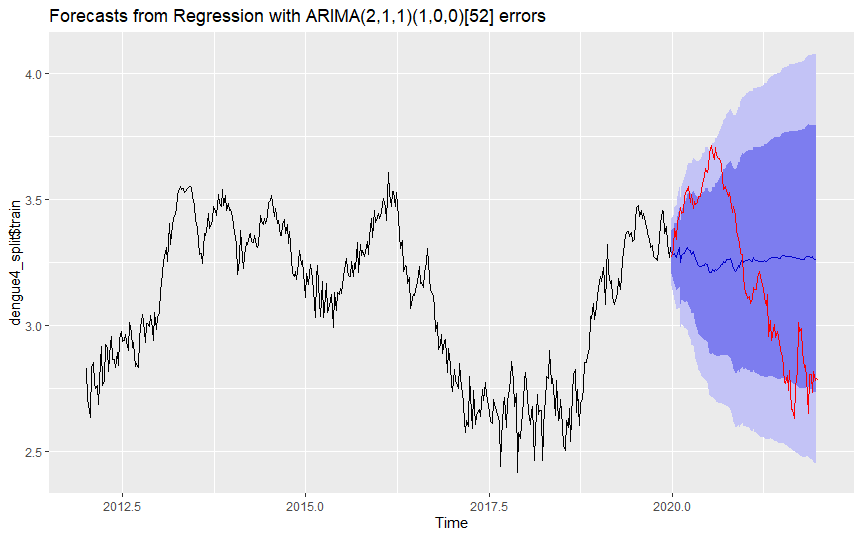
Due to the presence of x-reg variables, it is impossible to forecast future dengue cases directly using the current model. Instead, we have to first forecast for both rainfall and temperature separately in the out-of-sample time period, and then introduce the predicted forecasts for the respective variables as xreg variables when forecasting. We first need to ensure that the time series data for rainfall and temperature are stationary before forecasting, which we have already done prior to the modelling for ARIMA-X. With reference to figures 6 and 7, we can see that the time series for both temperature and rainfall are already stationary, while they also seem to contain some element of seasonal patterns. Hence, the seasonal naive method will be utilised for forecasting temperature and rainfall. The forecasted values in the next few diagrams will be represented by a blue line, while the actual data from our out-of-sample period is represented by the red line. The blue regions surrounding the blue line are the prediction intervals for the forecasts, at a default 95% prediction interval.



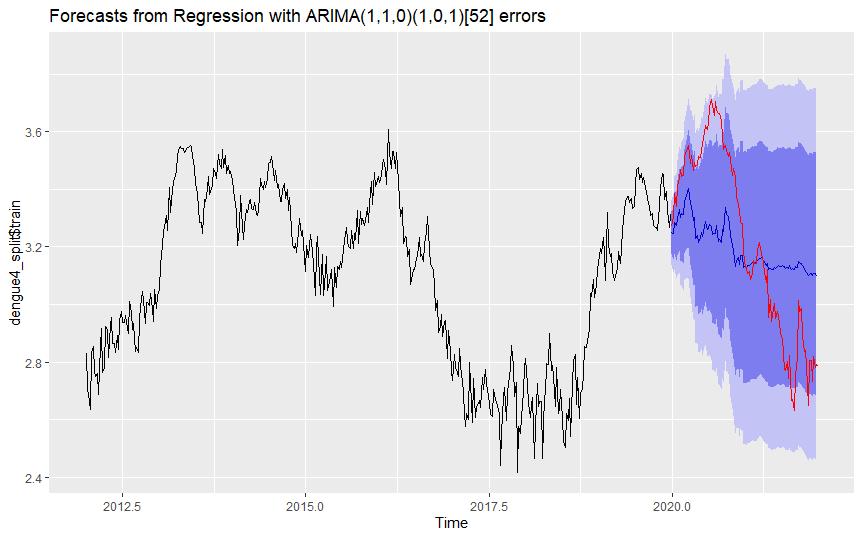
Using the seasonal naive method to forecast for weekly rainfall, we can see that the forecast managed to predict the trend cycles that the actual data followed, while all the actual data points were within the prediction intervals of the forecast. However, we also notice here that the seasonal naive forecast starts to lose its prediction accuracy as the time series progresses further to the present. The RMSEs for this forecast are 7.851179 for the training set, and 9.793153 for the testing set.



Using the seasonal naive method to forecast weekly temperature, we can see from the above diagram that the forecast managed to accurately forecast within the range of the mean temperature data. However, we also noticed that the trends of the forecast were not too accurate when put into comparison with the actual data. The RMSEs for this forecast are 26.21754 for the training set, and 30.74218 for the testing set.



Using the predicted values from the seasonal naive method, we forecasted dengue cases for our 2 S-ARIMA-X models. For the S-ARIMA-X(auto.arima) forecast, we can see that the forecast completely fails to forecast the trend of the actual data set. Furthermore, there were points in the data set which exceeded the prediction interval projection of the forecast. The RMSEs for this forecast are 0.06648509 for the training set, and 0.32726993 for the testing set.

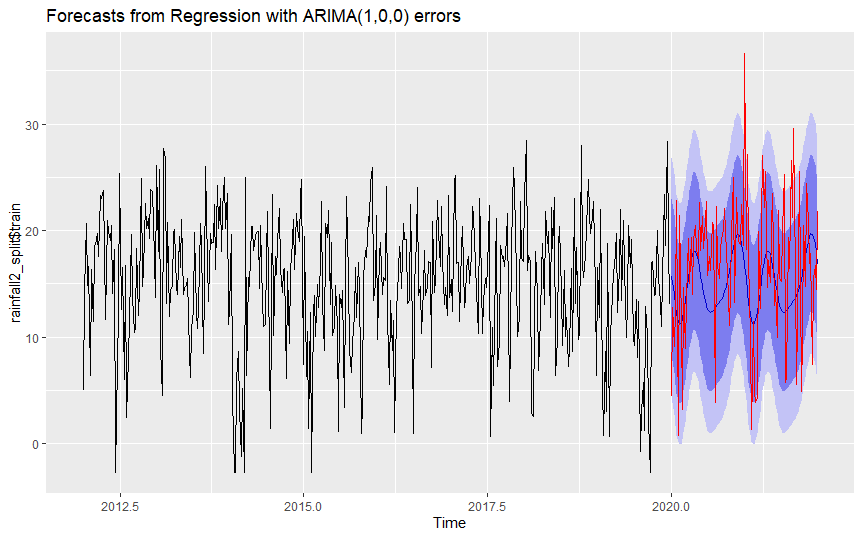


We then attempted to forecast for dengue cases using an S-ARIMA–X model, with the same parameters as our best S-ARIMA model. Visually, we can see that the forecast fared better at capturing a trend component for data as compared to our earlier forecast using the S-ARIMA-X(auto.arima) model. However, the forecast failed to capture the range of variation of the actual data set both in its prediction values, as well as within its prediction intervals. The RMSEs for this forecast are 0.05401304 for the training set, and 0.25657235 for the testing set.

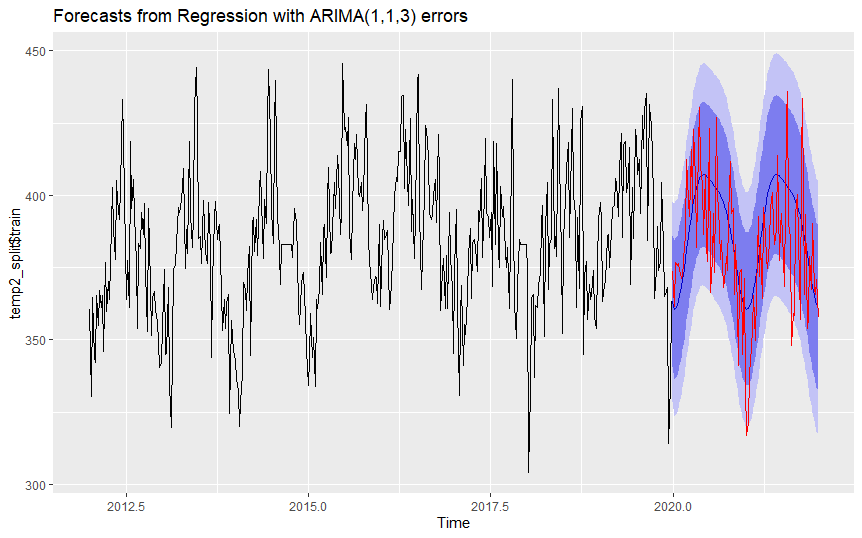
**S-ARIMA-X Forecasting (Dynamic Harmonic Regression Model)**:

One of the issues we faced with our forecasts was that our forecast was based on weekly data, which can cause issues as the exact seasonal period is a non-integer. Rstudio rounds this off to the nearest integer, but this is still an issue as most forecasting methods, including our earlier seasonal naive methods, do not handle such large seasonal periods well. Rob. J Hyndman instead suggests using a dynamic harmonic regression model for forecasting in order to address the seasonality issue. This method involves modelling the seasonal pattern using Fourier terms and allowing ARIMA errors to handle the short-term time-series dynamics.

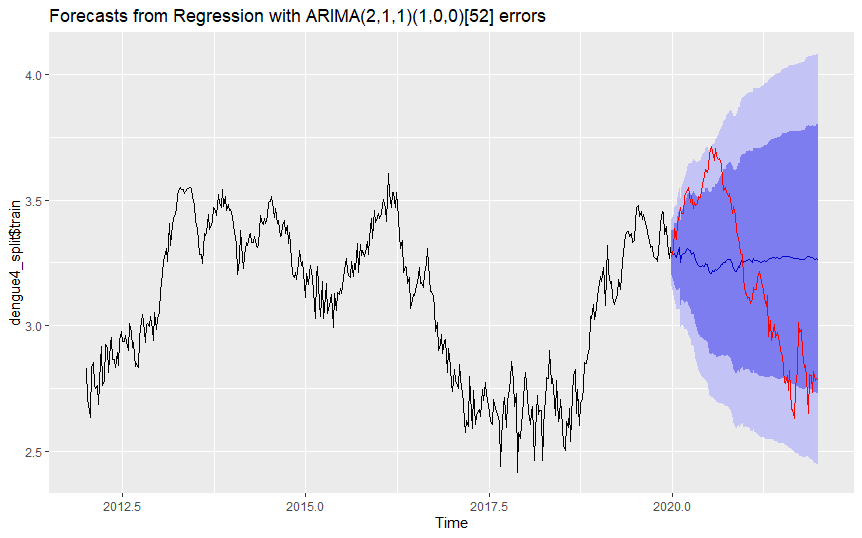
Thus, we utilised the dynamic harmonic regression model to forecast temperature and rainfall for our out-of-sample period, and using those forecasted values, we attempted to forecast for dengue cases alongside our S-ARIMA-X Models. The recommended number of Fourier terms was computed automatically by RStudio for the forecasts.



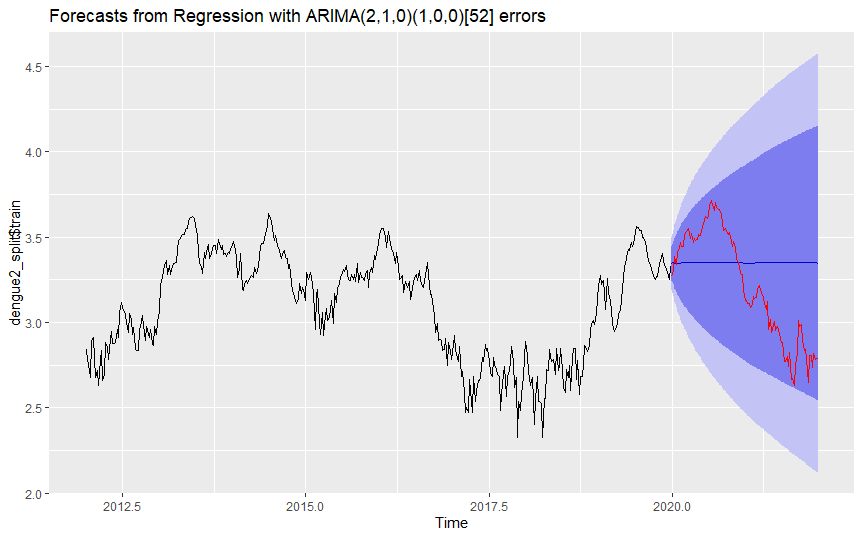
The above forecast showcases the forecast of rainfall using the dynamic harmonic regression model, with the best K terms being determined to be K=3 by RStudio. Visually, we see that the forecasts from the model does not capture the trend cycle of the actual data set accurately. We theorised that this could be because of the frequency of extreme outliers within the actual data set for rainfall, where it is common for certain days to have 0 rainfall and other days to have abnormally high amounts of rainfall, especially during the summer season. The RMSEs for this forecast are 5.638424 for the training set, and 6.477436 for the testing set.



The above forecast showcases the forecast of temperature using the dynamic harmonic regression model, with the best K terms being determined to be K=2 by RStudio. We see that the forecasts managed to accurately capture the seasonal cycles of temperature for the actual test set. However, we can also see that the prediction interval of the forecasts struggled to capture a few of the extreme outliers in the data points. The RMSEs for this forecast are 17.43086 for the training set, and 20.24130 for the testing set.



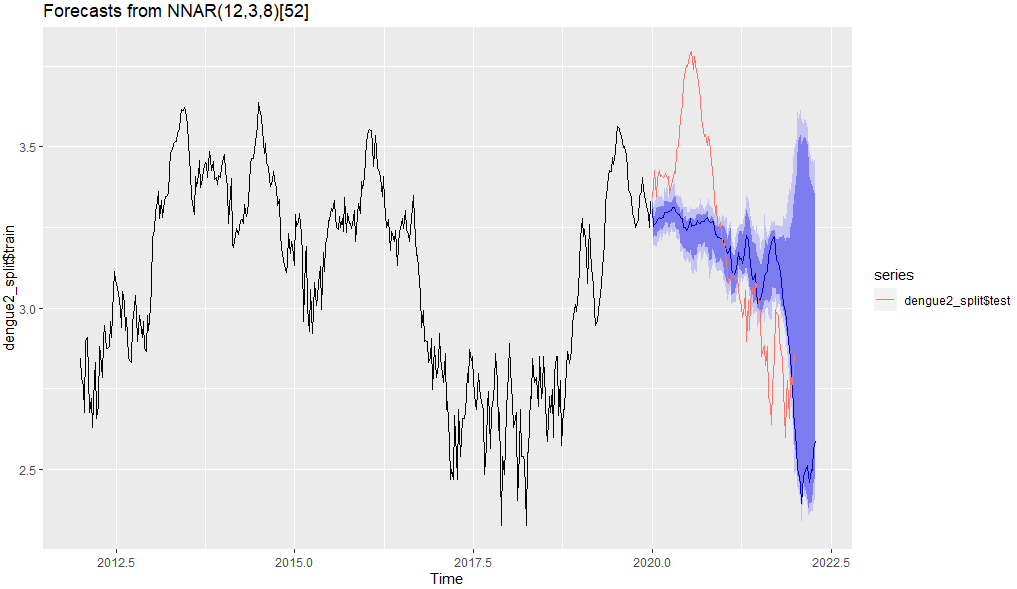
Using the predicted values from the dynamic harmonic regression method, we forecasted dengue cases for our 2 S-ARIMA-X models. For the S-ARIMA-X(auto.arima) forecast, we can see that the forecast again completely fails to forecast the trend of the actual data set, instead, it is trending towards a constant value. Furthermore, there were points in the data set which exceeded the prediction interval projection of the forecast. The RMSEs for this forecast are 0.06648509 for the training set, and 0.32903257 for the testing set.



When forecasting for dengue cases using the other S-ARIMA–X model, we can see that the forecast fared even worse as compared to our earlier forecast using the S-ARIMA-X (auto.arima) mode, with the forecasts completely failing to capture any seasonal or trend components when compared with the test set. However, interestingly, the prediction intervals were wide enough in this model such that all the points of the actual data set were actually captured within the prediction intervals of the forecasts. The RMSEs for this forecast are 0.07703571 for the training set, and 0.34361220 for the testing set.

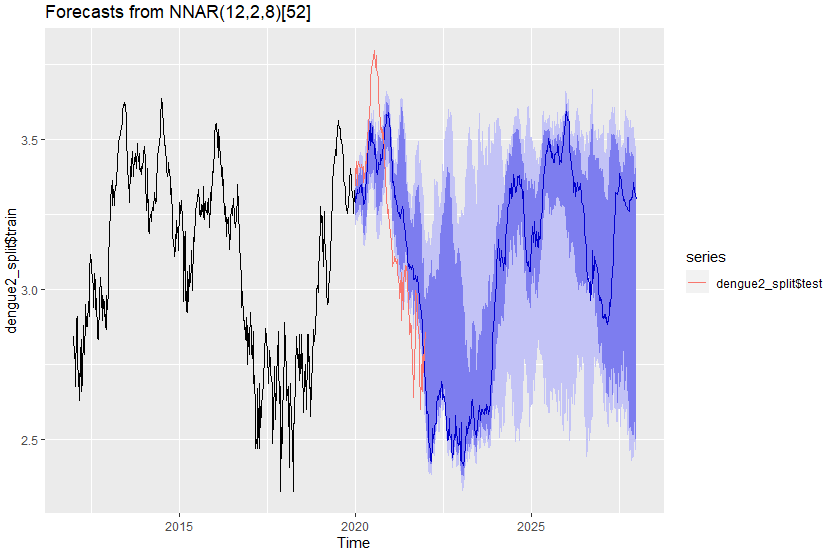
**Forecasting with Forward Feed Neural Networks:**

**Without external regressor**



The above graph shows the forecast from training the neural network on historical dengue data. While the model managed to capture the downward trend in early 2021, all the points of the actual data set were barely captured within the prediction intervals of the forecasts. The model also failed to account for the sharp rising trend during Covid in 2020, which is expected. Hence, external regressors are added to hopefully predict this rising trend.

**With external regressors - rainfall and temperature**



Adding external regressors like rainfall and temperature similar to the ARIMAX models might help to account for the Covid trend, since they are additional covariates added onto lagged values of the response. Interestingly, the model managed to capture the sharp rising trend from Covid in 2020, with quite a decent number of actual data points falling in the prediction intervals. Even though the model still predicted a falling trend in 2021, all the points of the actual data set were barely captured within the prediction intervals of the forecasts. Comparing both forecasts, adding external regressors was better in predicting trends, but both failed to predict values accurately and failed to capture the extreme outliers.

**Out-of-sample performance comparison:**

|  |  |  |
| --- | --- | --- |
| Model | RMSE of Model for Rainfall | RMSE of Model for temperature |
| Seasonal Naive | Training: 7.851179  Test: 9.793153 | Training: 26.21754  Test: 30.74218 |
| Dynamic Harmonic Regression | Training: 5.638424  Test: 6.477436 | Training: 17.43086  Test: 20.24130 |

|  |  |  |
| --- | --- | --- |
| S-ARIMA-X | (2,1,1)(1,0,0)[52]  - auto.arima | (1,1,0)(1,0,1)[52]  - Manual Parameters |
| Seasonal  Naive Method  for predictors | Training: 0.06648509  Test: 0.32726993 | Training: 0.05401304  Test: 0.25657235 |
| Dynamic Harmonic Regression Method  for Predictors | Training: 0.06648509  Test: 0.32903257 | Training: 0.07703571  Test: 0.34361220 |

|  |  |  |
| --- | --- | --- |
| Model | Without External Regressor | With External Regressor (xreg) |
| Feed forward  Neural Network | NNAR(p=12, P=3, k) | NNAR(p=12, P=2, k) |
| Train RMSE | 0.02758349 | 0.03070044 |
| Test RMSE | 0.25096222 | 0.20764560 |

By comparing the RMSEs of the out-of-sample forecasts, we can see that the RMSE for the respective predictors are far lower for the dynamic harmonic regression as compared to the seasonal naive method.

The RMSE for the auto.arima model seems to be the same regardless of which method is being used for predicting temperature and rainfall. Interestingly, the RMSE suggests that the forecasts for dengue using the seasonal naive method predictors for both temperature and rainfall are a better fit as compared to using the dynamic harmonic regression method predictors for both temperature and rainfall for the manual parameters for the S-ARIMA model.

Lastly, as we can see from the earlier analysis of the forecasts for the Feed Forward Neural Network that the forecasts for the Feed Forward Neural Network models were much better at accurately forecasting the number of dengue cases. The lower RMSE for both Neural Network models also reflect this.

1. **Comments and Future Improvements**

As mentioned by Professor Ee, to further improve upon the ARIMA models, it is possible to model after the residuals and check if there are AR or MA components. This will give us a better estimation of the parameters that we may then insert into our models and eliminate the problem of having time-series information left in the residuals.

One of the key variables that affected the accuracy of our model was the Project Wolbachia done by NEA (NEA, n.d.). The project releases male Wolbachia-Aedes mosquitoes to effectively reduce the dengue mosquito population. This can severely affect our results as the number of dengue cases will drop dramatically due to the effect of this project. It may be possible to add a dummy variable to take into factor for this project, allowing us to improve our models' accuracy.

As mentioned previously, other key variables may drastically affect the accuracy of our models. This can be due to the climate change effect increasing transmission intensity of dengue fever as temperature and humidity change (Tran, 2020). The incubation period of mosquitoes will change based on climate factors. Furthermore, climate change can cause minor (albeit really slow) changes in the seasons experienced in a tropical climate like Singapore, which can cause seasons like the monsoon season to occur in different months of the year, as compared to previous years.

Furthermore, Channel News Asia mentioned that the record-high dengue cases can be due to circuit breaker measures as more people stayed home (Cook, 2020). This increases the population density as we have previously mentioned resulting in more dengue cases. Such a dramatic change in measures created an unexpected “shock” in the covid-19 data in the form of a spike, which definitely could not have been captured using the information we had on dengue in earlier time periods. This can have two different effects on our models. Firstly, this dramatically increases the number of dengue cases forecasted by our models as the number of dengue cases during Covid-19 period is much higher. Secondly, such unexpected shocks results in our forecasts for that time period being inaccurate as such shocks will be represented as an outlier in the actual data. In the event of such issues, it would be better to look at the prediction intervals of our model’s forecasts to see whether such unexpected shocks could have been captured within the range of the prediction intervals.

Lastly, other possible factors such as the presence of multiple different seasonal cycles, as well as possible collinearity within the exogenous variables, were possibly present in the data, which could explain why the S-ARIMA-X models were such a poor fit for the data set. Perhaps other complex methods could have been further explored, such as modelling a more complex ARIMA-X model which takes into account for differing seasonality, or other methods such as vector autoregression or even Long Short Term Memory (LSTM) models.

**References**

Wolbachia-Aedes mosquito suppression strategy.

*National Environment Agency, NEA. (n.d.).*

Retrieved April 10, 2022, from https://www.nea.gov.sg/corporate-functions/resources/research/wolbachia-aedes-mosquito-suppression-strategy

*Tran, B.-L., Tseng, W.-C., Chen, C.-C., &amp; Liao, S.-Y. (2020, February 21).*

Estimating the threshold effects of climate on Dengue: A case study of taiwan. International journal of environmental research and public health.

Retrieved April 10, 2022, from

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7068348/#

*Cook, A. (2020, October 1)*. Commentary: Uncovering the factors fueling record-high dengue cases in Singapore. CNA.

Retrieved April 10, 2022, from https://www.channelnewsasia.com/commentary/why-singapore-record-high-dengue-cases-covid-19-2020-coronavirus-687281