Forecasting PM2.5 Air Pollution in Beijing, China

Sarah Mansfield

Introduction + Background

PM2.5, or atmospheric particulate matter that has a diameter <2.5 micrometers, is a major contributor to air pollution in China. Repeated exposure to high concentrations of these particles leads to adverse health risks such as heart disease and lung cancer, and therefore poses a major health risk to the general public - in 2019, China and India alone accounted for 58% of worldwide deaths attributed to PM2.5 pollution.

In more recent years, however, China has made concerted efforts to address these concerns - the first comprehensive five-year plan to improve air quality was implemented between 2013 and 2017, and subsequent plans have been made in order to continue improving air pollution. Between 2010 and 2019, outdoor PM2.5 levels in China decreased by approximately 30%, largely due to actions taken such as a shift from coal to gas in residential and industrial sectors, and a reduction in industrial emissions.

This report seeks to examine whether we can forecast future PM2.5 concentrations based on previous values of PM2.5 and other factors such as air pollutant concentrations (O3, NO2, SO2, CO). By doing so, we hope to be able to get a clearer picture around certain variables that may contribute more to air pollution, as well as get a general idea of the future trajectory of PM2.5 pollution in China. As a result, this project makes use of a data set detailing daily PM2.5 concentration in Beijing, China from 2014 up until March 2022 in order to provide an up-to-date analysis of recent and future trends.

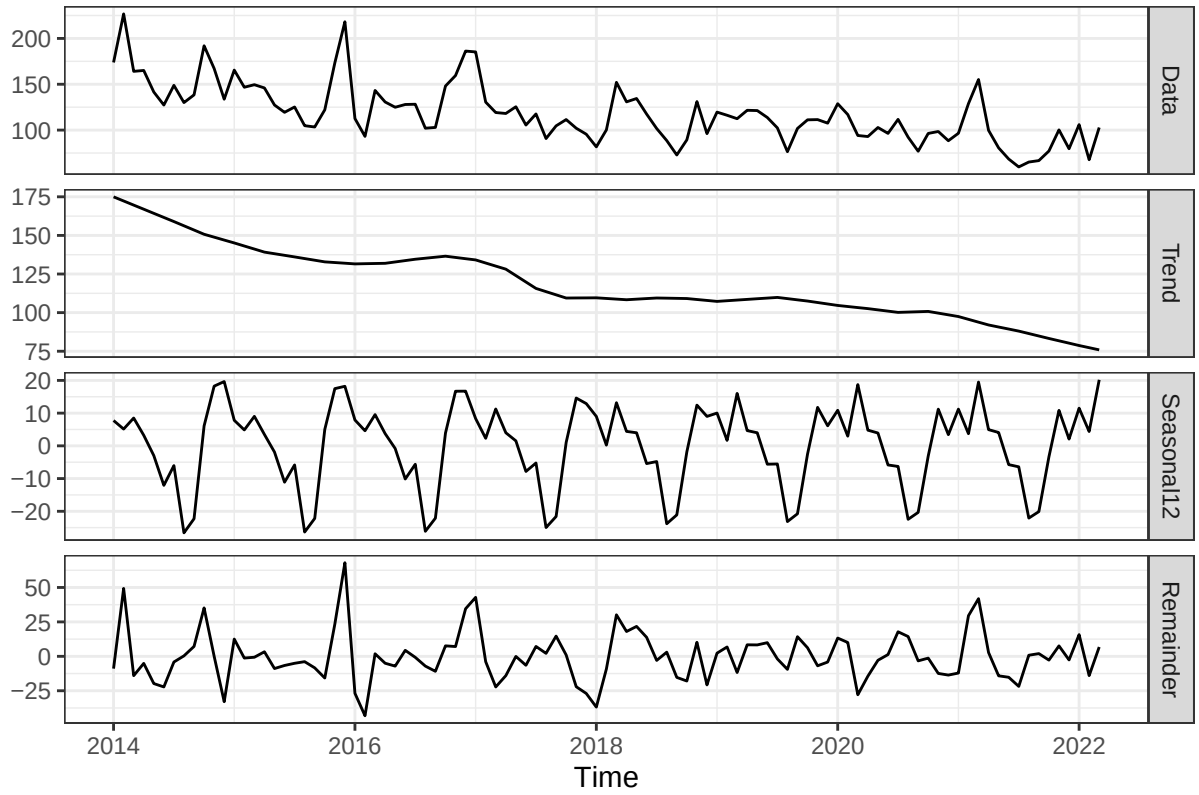
Data

Data was collected from the Beijing Environmental Protection Monitoring Center, and consists of 3,018 total daily observations (where each row contains the daily PM2.5, O3, SO2, CO, and NO2 concentrations) collected from January 1, 2014 - March 31, 2022. Since the aim of this project is to get a more general sense of future PM2.5 concentrations, the data was aggregated into average monthly data, resulting in 99 monthly observations. The time series object was created using this average monthly data and setting frequency = 12 in order to indicate monthly seasonality. Included below is a table showcasing the first 10 observations in the data set:

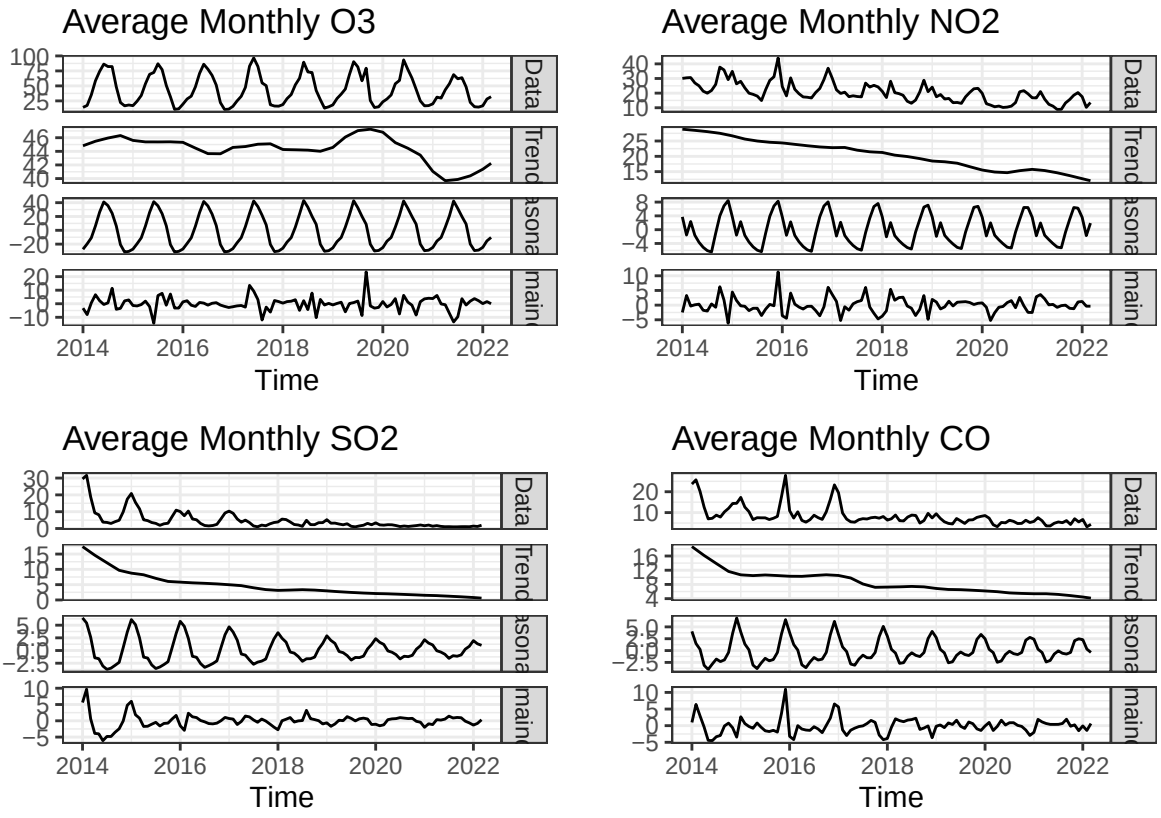
Table 1: Beijing Air Quality Data Set

month	pm25_mean	pm10_mean	o3_mean	no2_mean	so2_mean	co_mean
2014-01-01	173.8000	92.4194	14.1290	30.0645	29.4839	23.6774
2014-02-01	226.7500	111.6429	17.3571	30.4643	31.6429	25.7037
2014-03-01	164.0968	90.7742	35.0000	30.6452	18.8065	20.1613
2014-04-01	165.0667	95.8000	58.5667	26.9667	9.3333	12.5926
2014-05-01	141.5484	90.1290	74.1935	25.0000	8.1935	7.0000
2014-06-01	127.3333	67.8333	86.3333	21.3333	3.8000	7.2667
2014-07-01	148.8065	77.7097	82.2903	19.9677	3.6129	8.7742
2014-08-01	130.0000	70.2581	82.1935	21.8065	3.0323	7.8710
2014-09-01	138.4000	70.3000	48.4000	25.9333	4.0667	10.3667
2014-10-01	192.0000	112.1613	22.2258	37.7742	4.9032	12.0323

Analysis



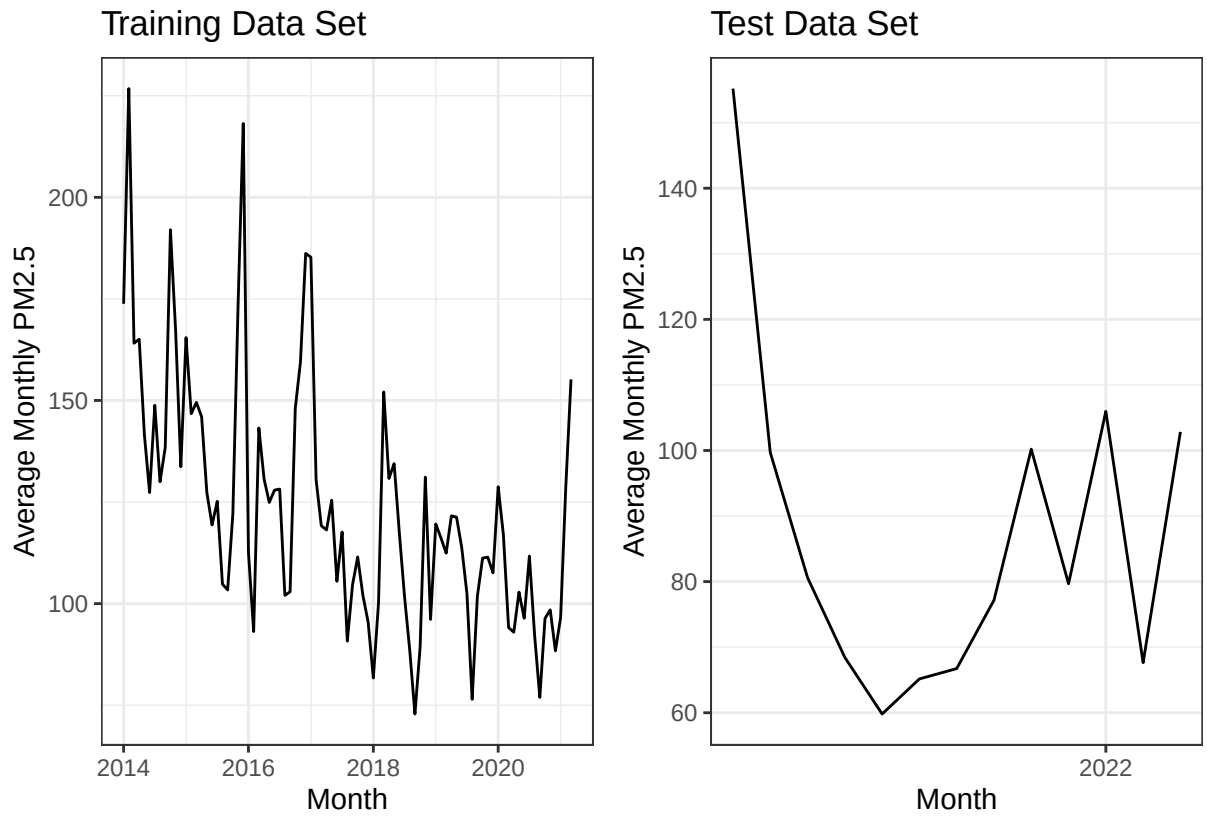
After plotting the decomposed time series object, we can see that there does appear to be a declining trend in PM2.5 over time from 2014-2022, as well as an indication of pretty strong monthly seasonality.



Plotting the decomposed time series plots of our four exogenous air pollutant variables as well (using the mstl() function), we can see that O3 and NO2 have very strong, consistent monthly seasonality, so we'll use these two as exogenous variables.

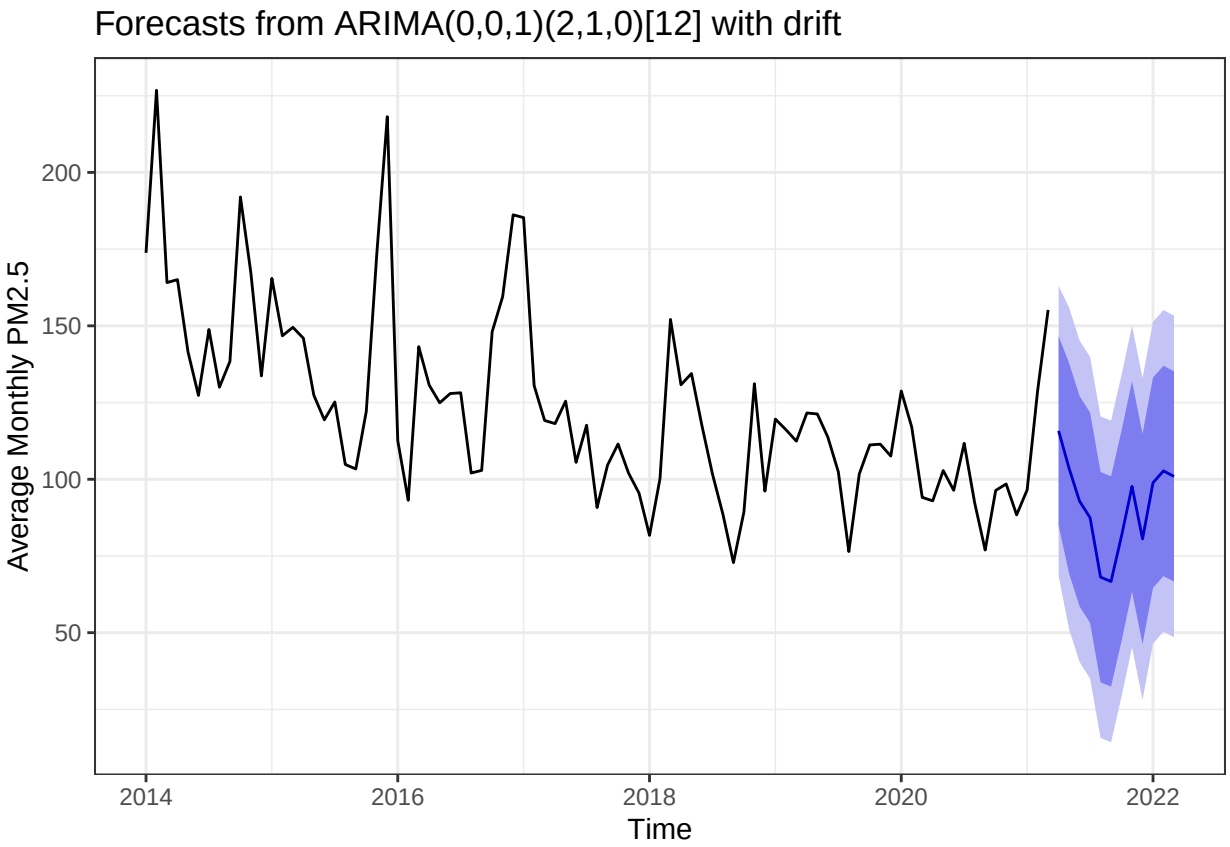
Model Fitting

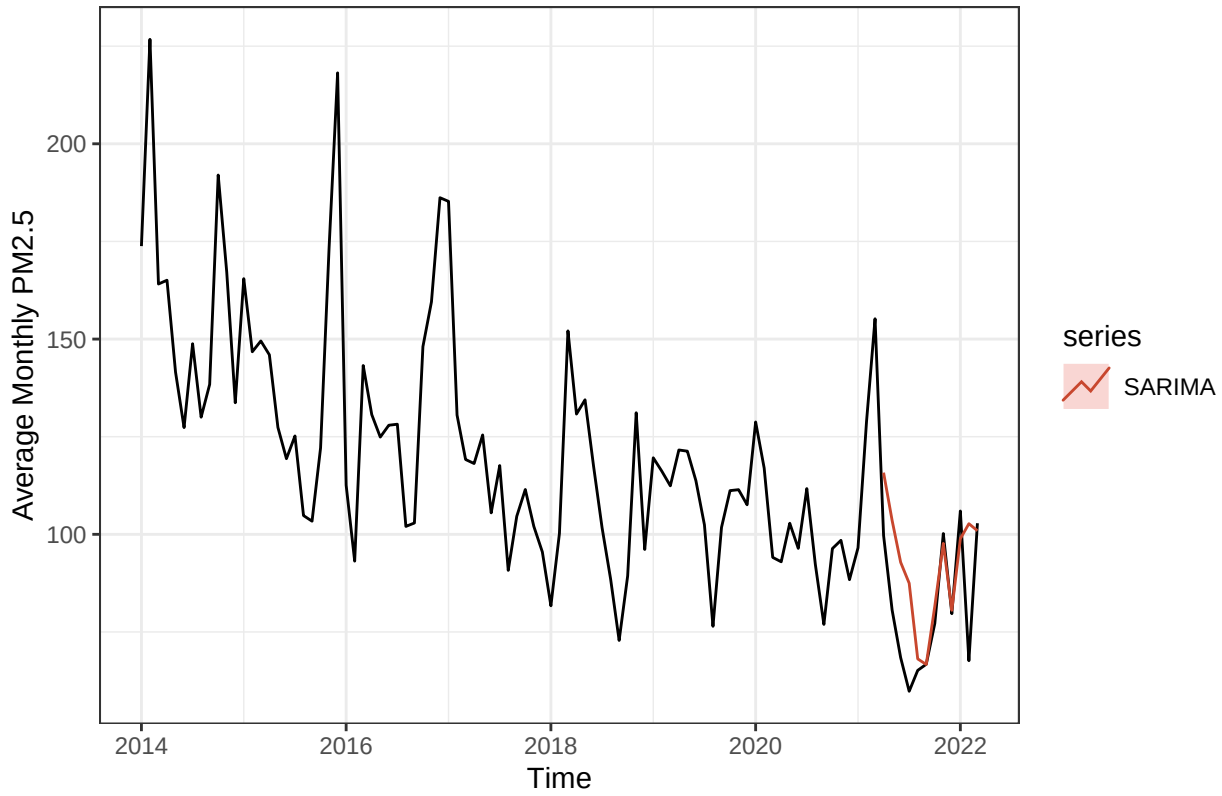
For modeling, the data was split into two subsets - a training set and a test set. The training set was subsetted on all the data from 2014 up until the last 12 months, and the test set was set aside as the last 12 months of data. The time series plots of both sets is pictured below.



Model 1: SARIMA

The first model was a SARIMA model fitted using the auto.arima() function, and was identified as a $SARIMA(0,0,1)(2,1,0)_{12}$ model with drift included.

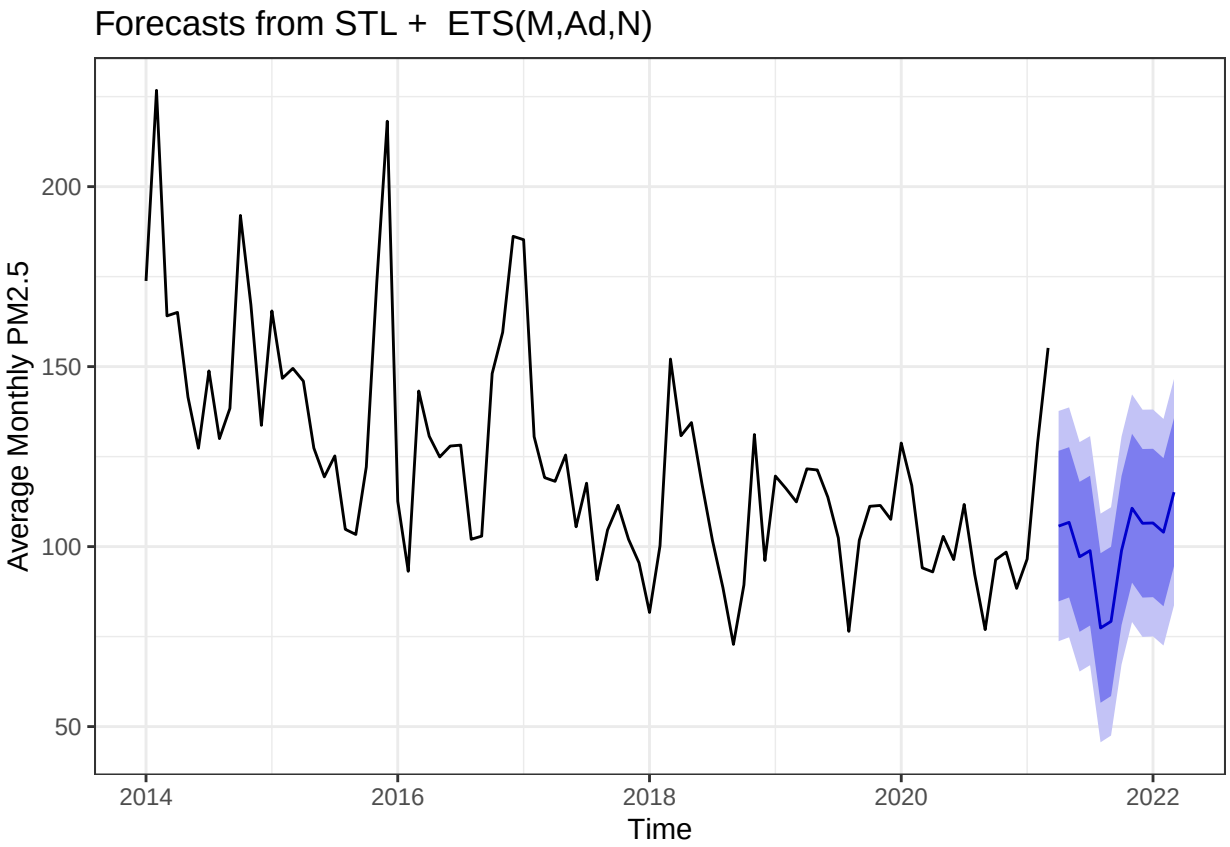

Forecasts from ARIMA(0,0,1)(2,1,0)[12] with drift
Average Monthly PM2.5
200
150
100
50
2014
2016
2018
2020
2022
Time

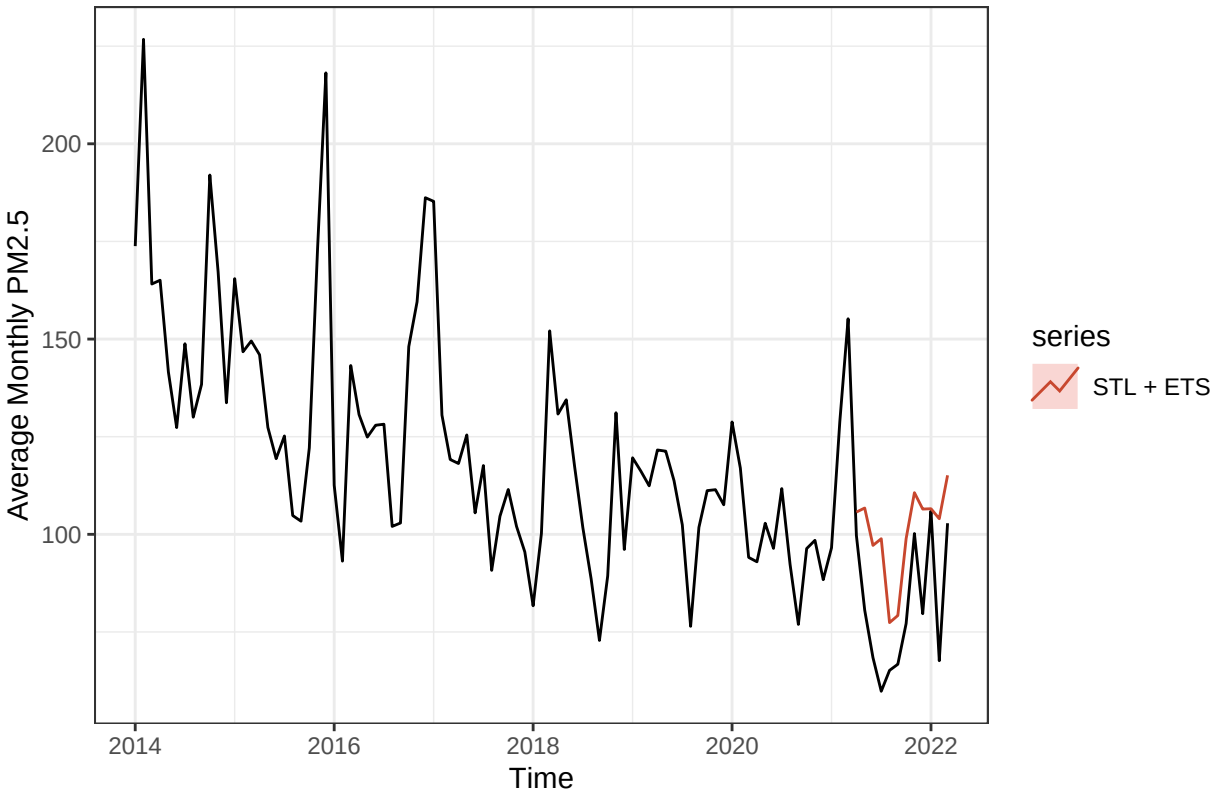


As seen from the forecasts on the test set, the SARIMA model does a pretty good job at approximating average PM2.5 concentration per month, especially from September through January, and all of the forecasts fall well within the confidence bounds.

Model 2: STL + ETS

The second model was an STL + ETS model fitted using the default parameters.

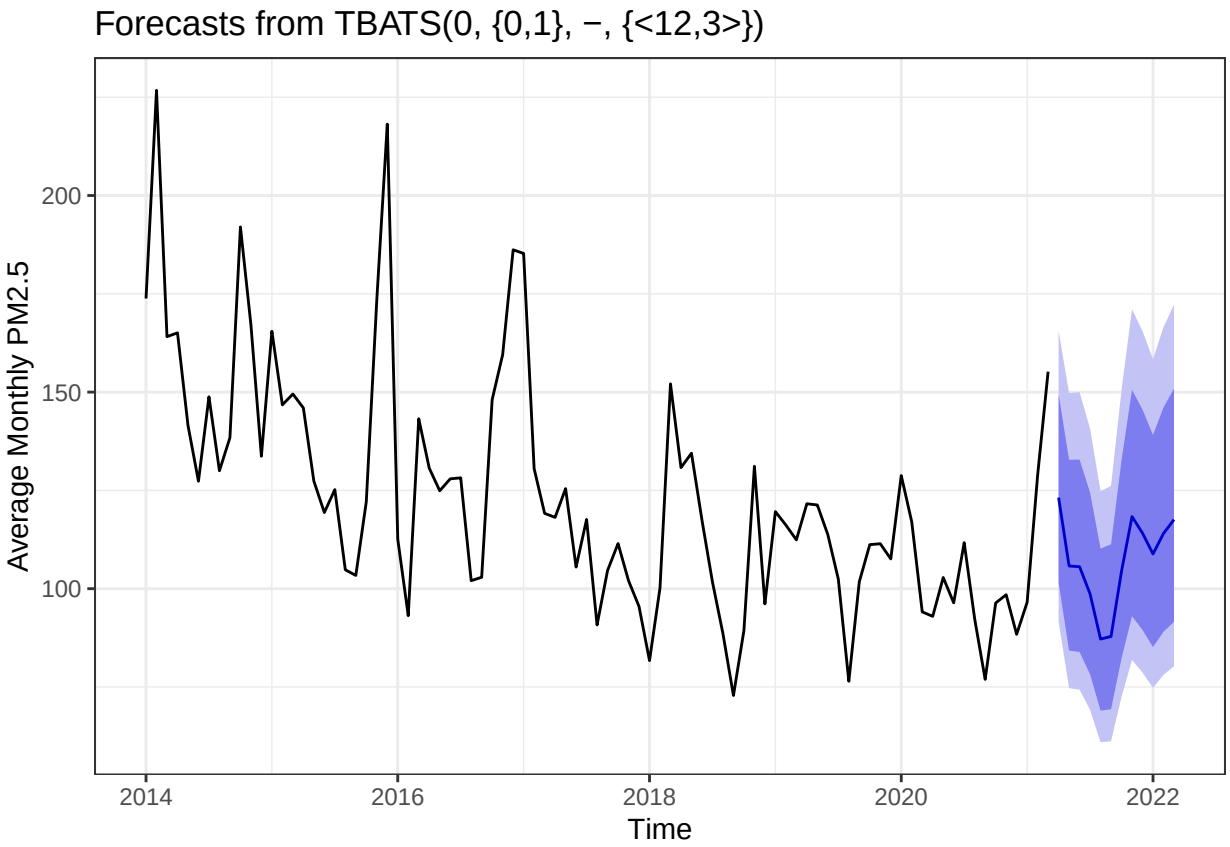

Forecasts from STL + ETS(M,Ad,N)
Average Monthly PM2.5
200
150
100
50
2014
2016
2018
2020
2022
Time

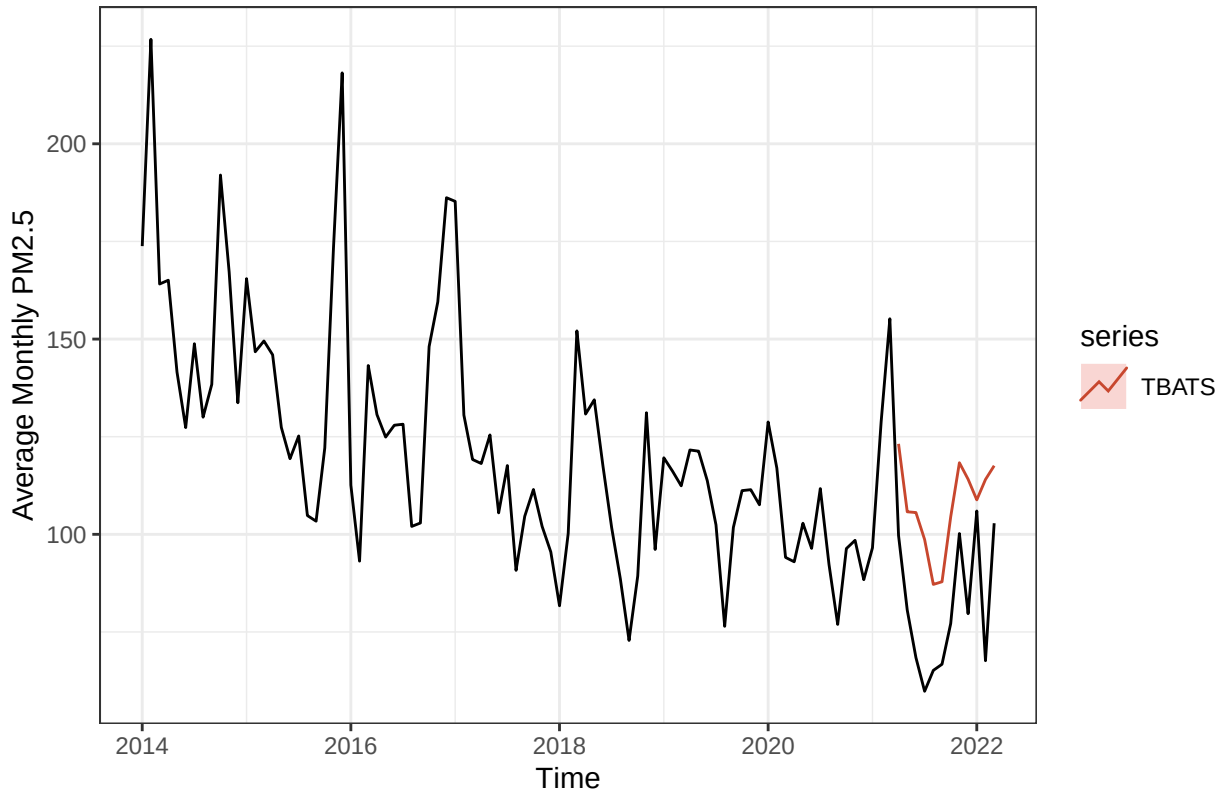


Again, we see that all of the forecasts fall within the confidence bounds; however, despite roughly following the general upward-downward trend in the data, all of the forecasts on the test set are over-approximated when compared to the actual values.

Model 3: TBATS

The third model was a TBATS model fit using the default parameters.

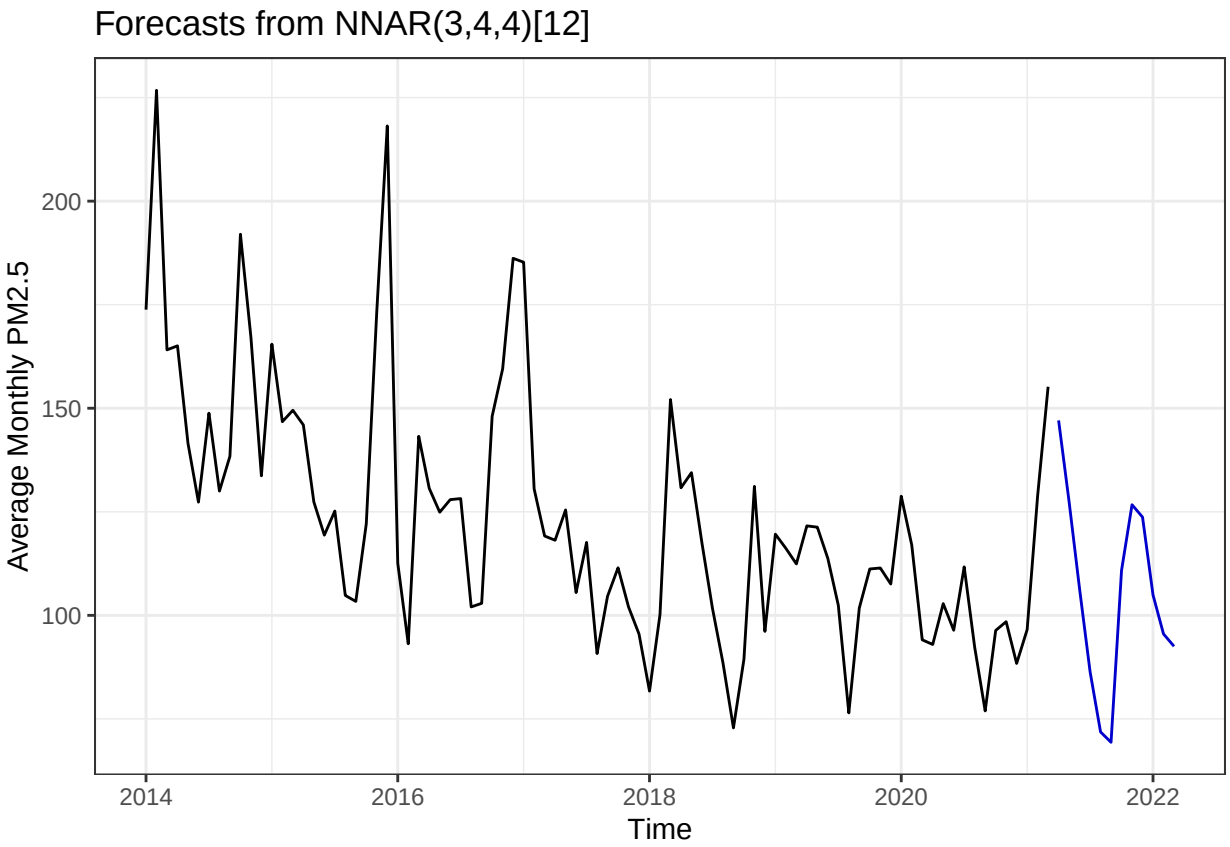

Forecasts from TBATS(0, {0,1}, −, {<12,3>})
Average Monthly PM2.5
200
150
100
2014
2016
2018
2020
2022
Time

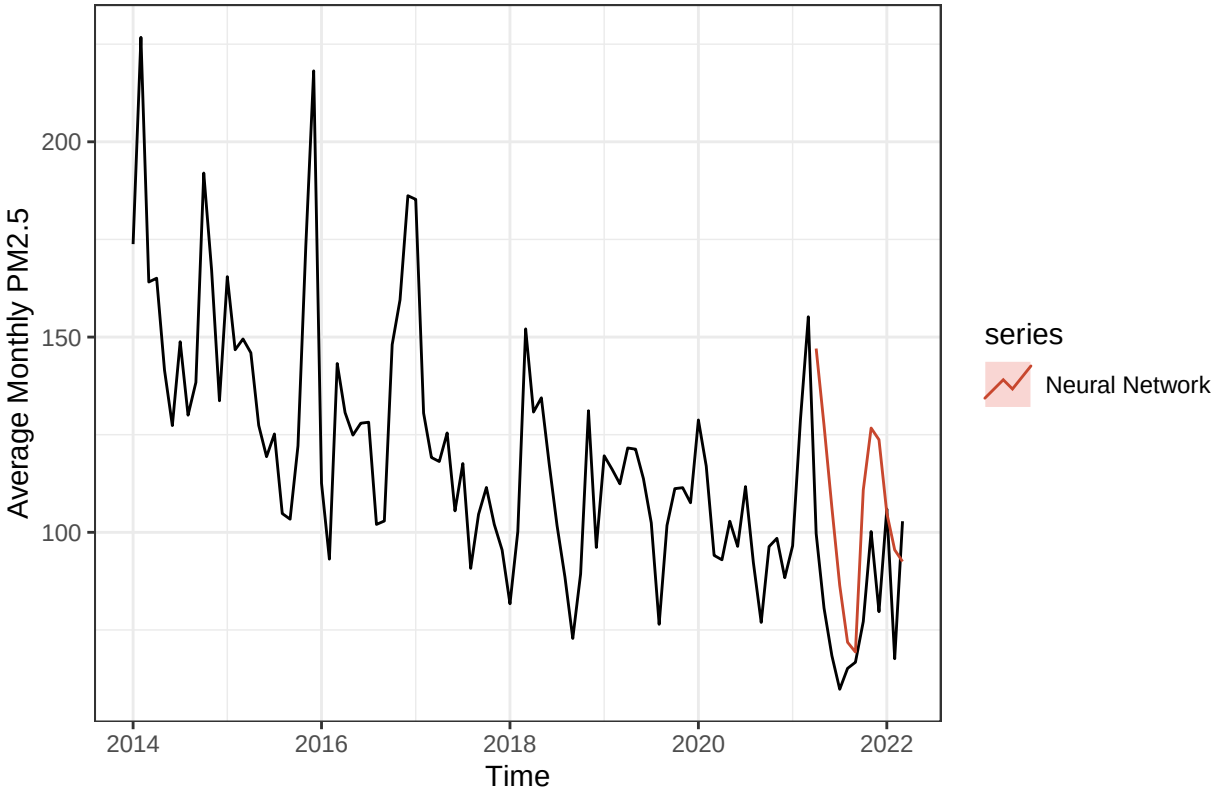


Again, much like the STL + ETS model, the TBATS model tends to over-forecast, and struggles a bit more with anticipating and predicting lower dips in the data.

Model 4: Neural Network

The fourth model was a Neural Network with parameters p=3 and P=4, with these parameters chosen based on fitting different models with different combinations of parameters and comparing accuracy metrics such as RMSE and MAPE.

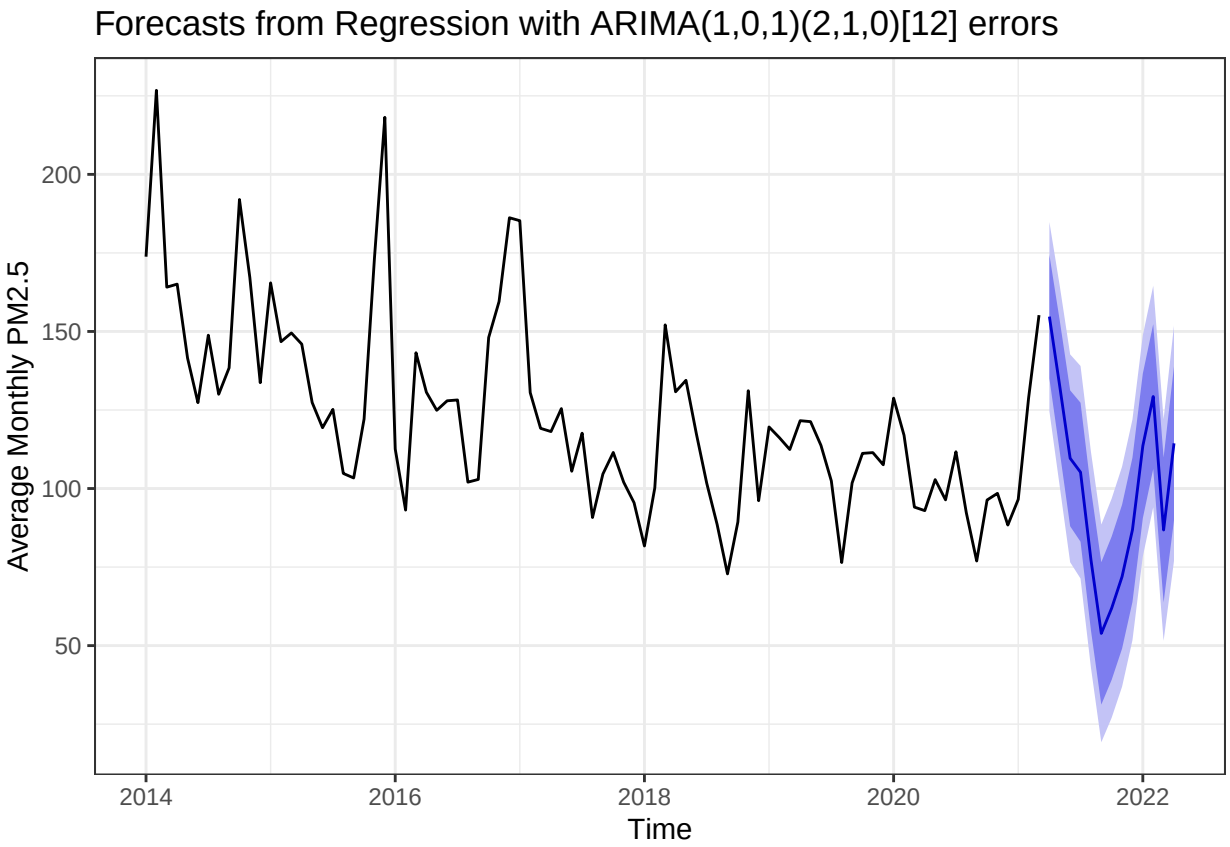

Forecasts from NNAR(3,4,4)[12]
Average Monthly PM2.5
200
150
100
2014
2016
2018
2020
2022
Time

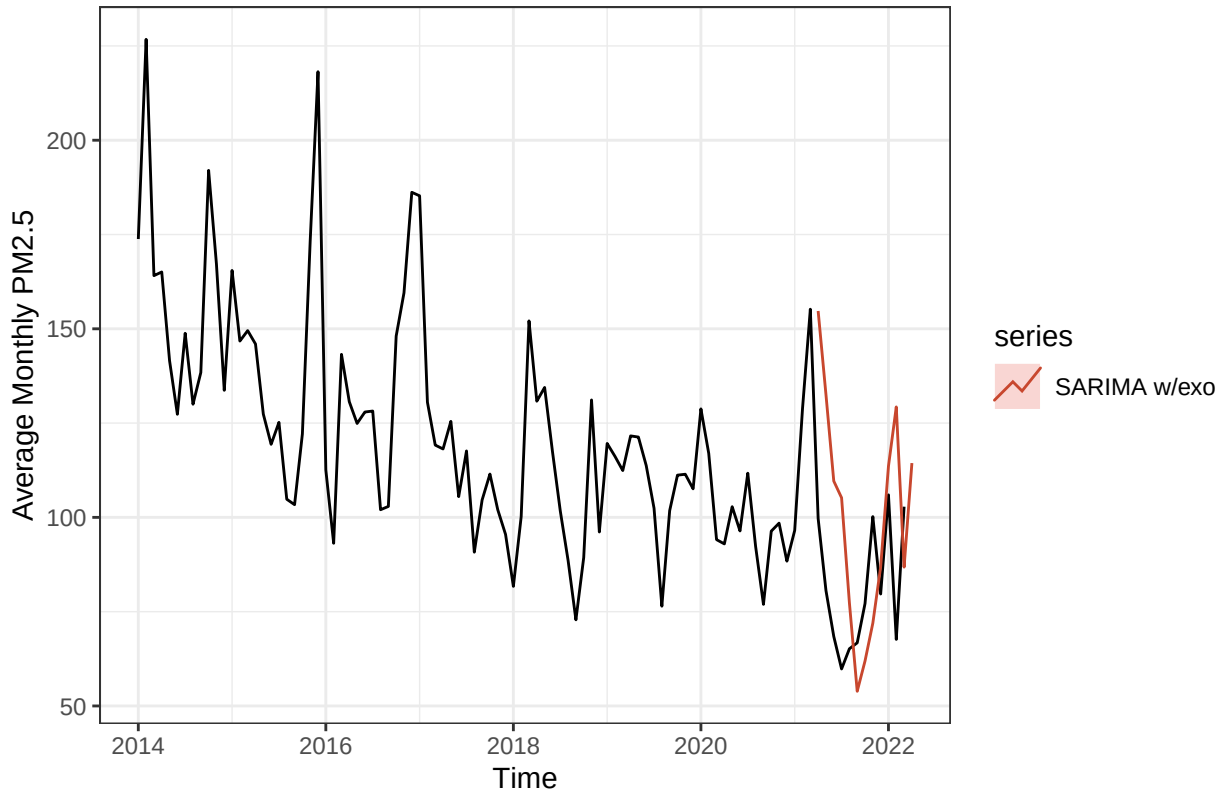


We can see that compared to the previous two models, the neural network does a much better job at predicting harsher dips in the data, though it still tends to over-approximate by a bit much like the other models.

Model 5: SARIMA w/exogenous variables

The fifth model was a SARIMA model fitted using the auto.arima() function as well as our two exogenous variables, and was identified as a $SARIMA(1,0,1)(2,1,0)_{12}$ model.

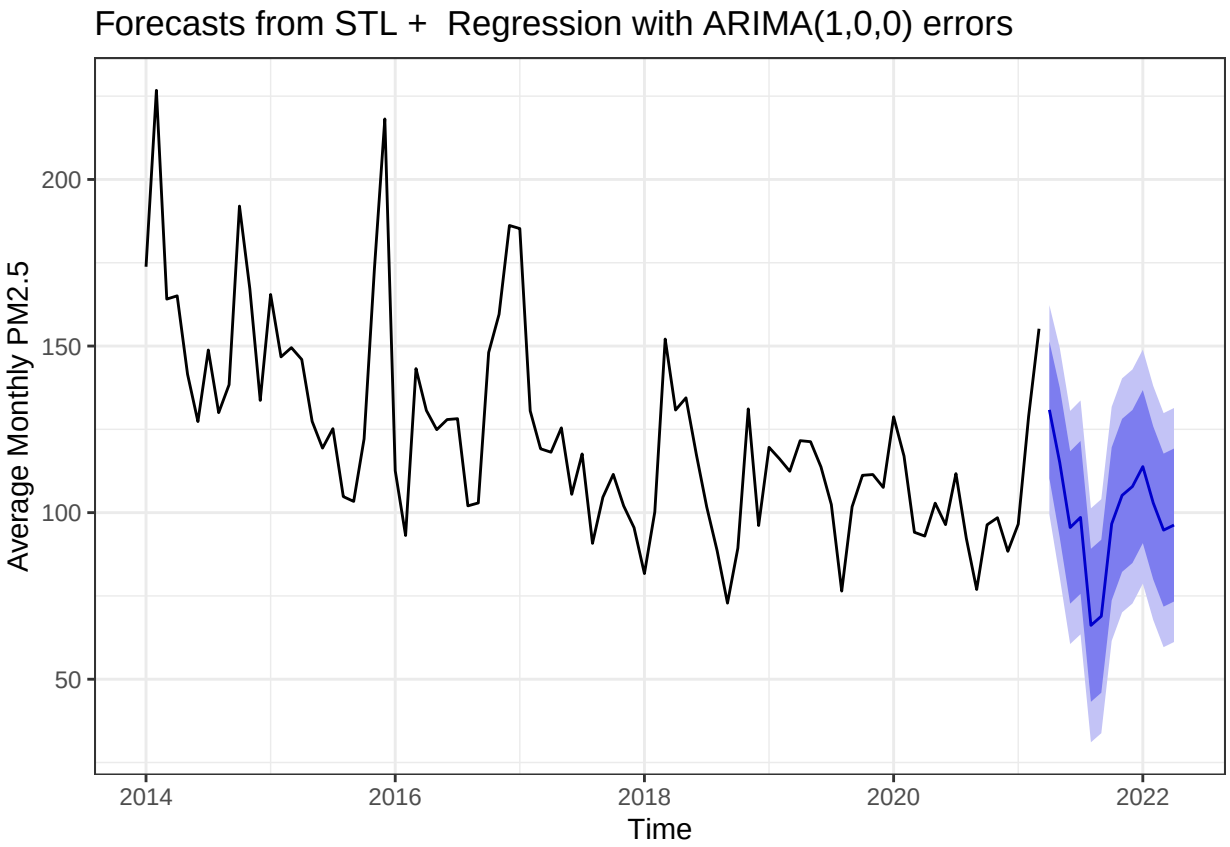

Forecasts from Regression with ARIMA(1,0,1)(2,1,0)[12] errors
Average Monthly PM2.5
200
150
100
50
2014
2016
2018
2020
2022
Time

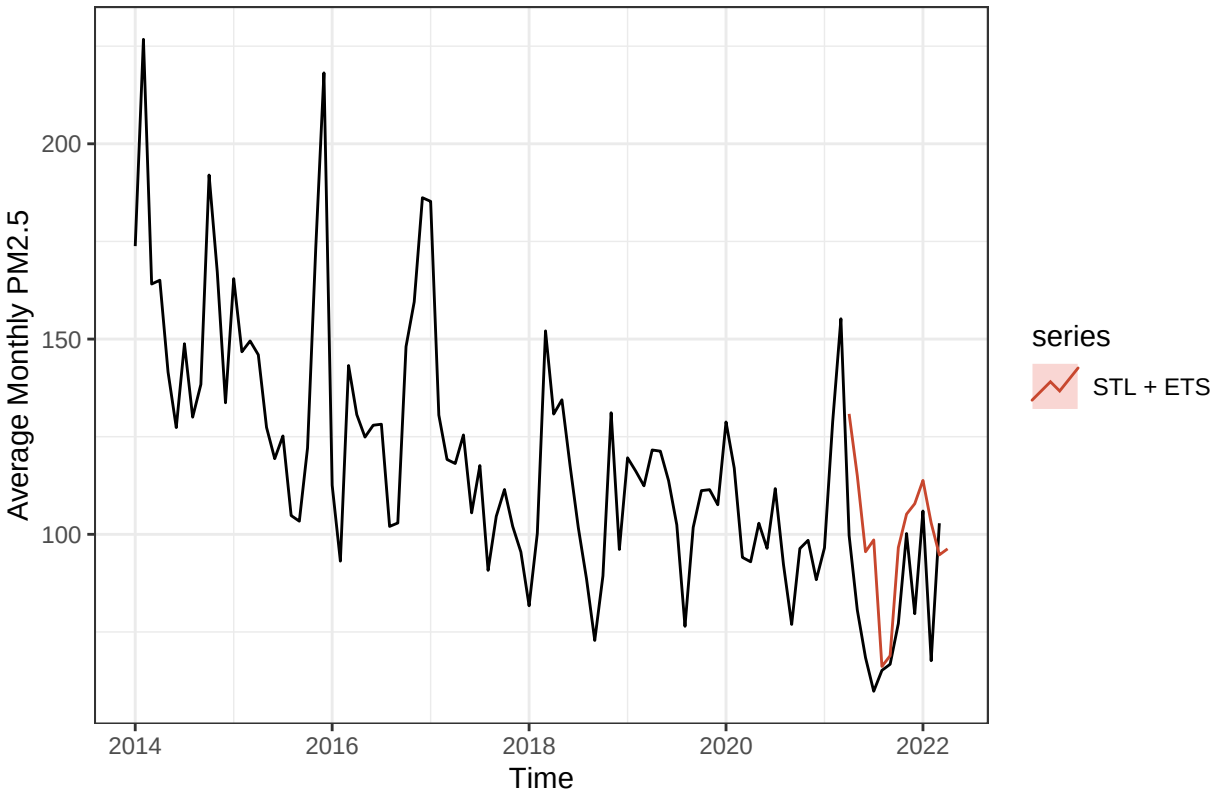


The forecasts from this model were more volatile than the others, showcased by harsher upward and downward spikes. We can see that the predictions here tend to both under and over predict.

Model 6: STL + ETS w/exogenous variables

The sixth model was an STL + ETS model with exogenous variables, trained using the default parameters and by setting method = "arima".

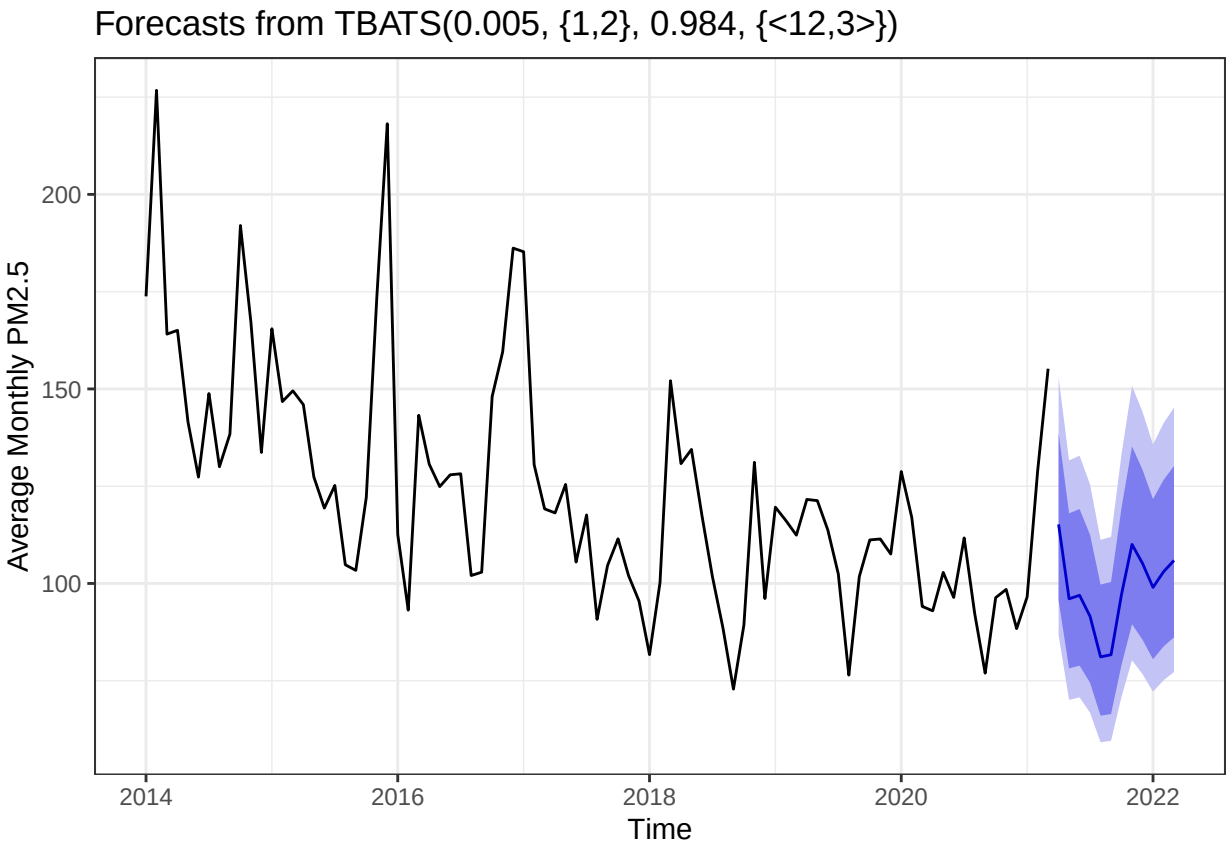

Forecasts from STL + Regression with ARIMA(1,0,0) errors
Average Monthly PM2.5
200
150
100
50
2014
2016
2018
2020
2022
Time

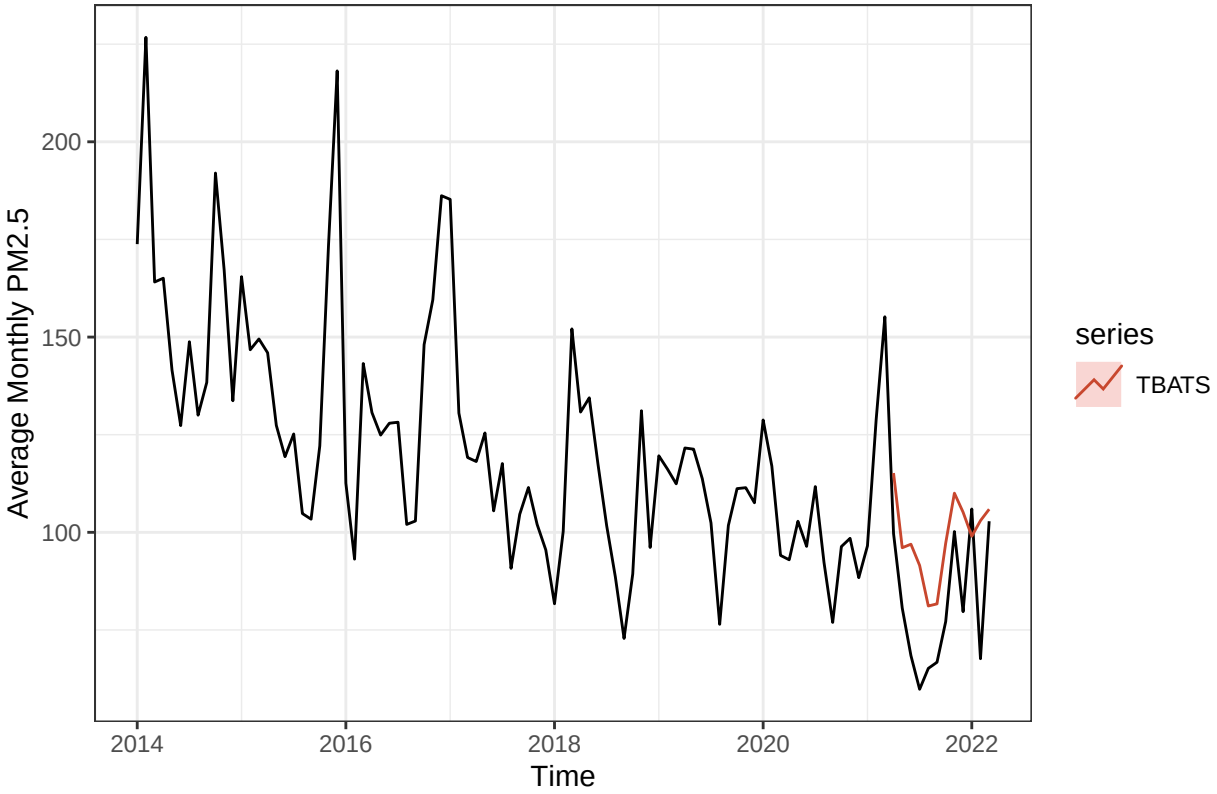


Compared to the STL + ETS model without exogenous variables, this model captured the dip in the data much better, though we note that it still tends to over predict by a bit for some of the more recent months.

Model 7: TBATS w/exogenous variables

The seventh model was a TBATS fit using the exogenous variables and with the default parameters.

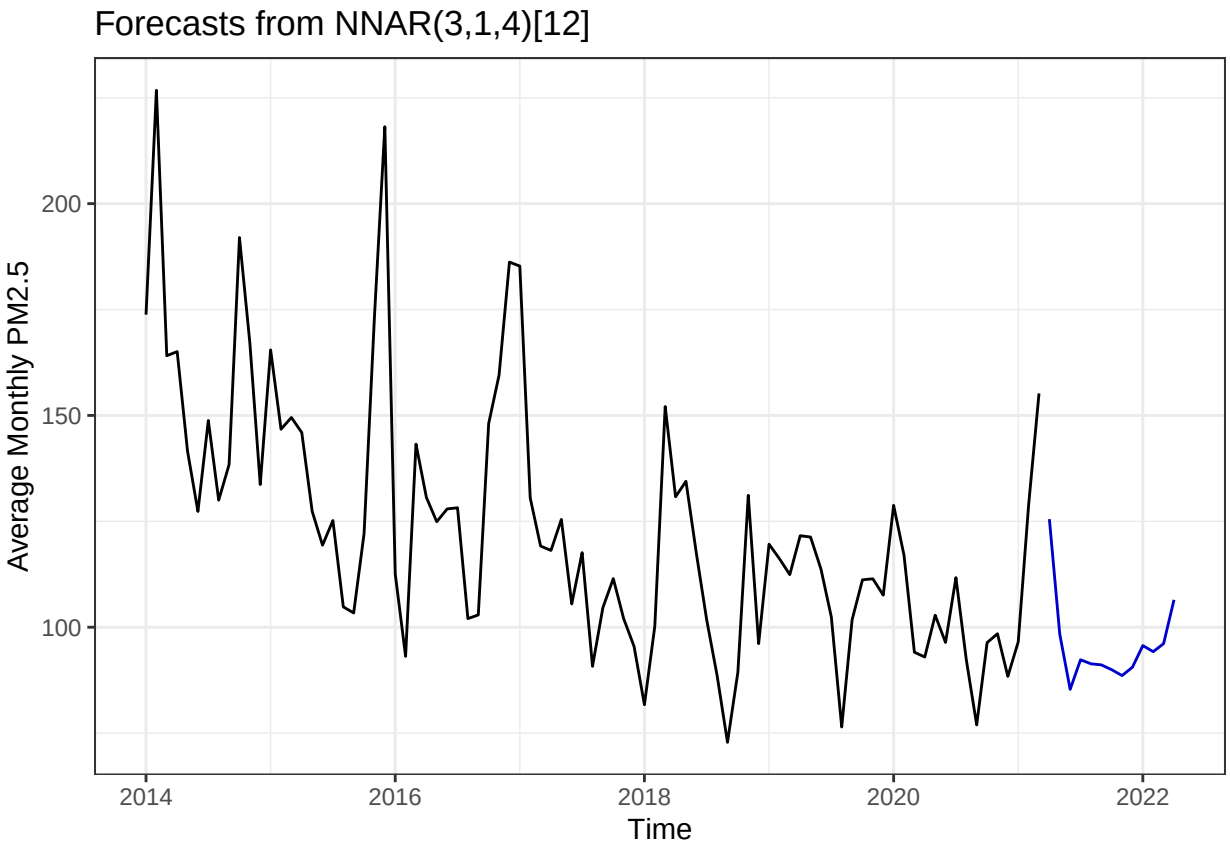

Forecasts from TBATS(0.005, {1,2}, 0.984, {<12,3>})
Average Monthly PM2.5
200
150
100
2014
2016
2018
2020
2022
Time

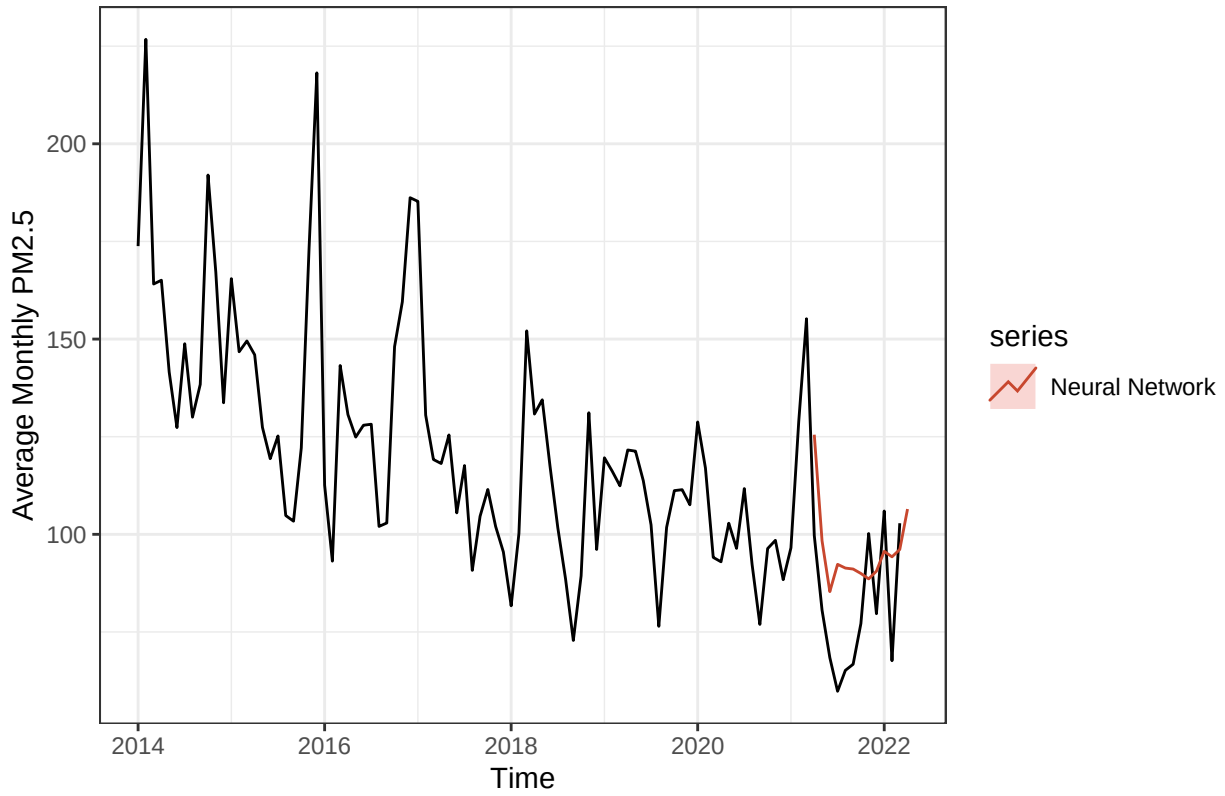


Much like many of the other models, this model again struggles with predicting that dip in the data.

Model 8: Neural Network w/exogenous variables

Lastly, the eighth model was a neural network with exogenous variables and with parameters p=3 and P=1 (chosen using the same methodology as before).

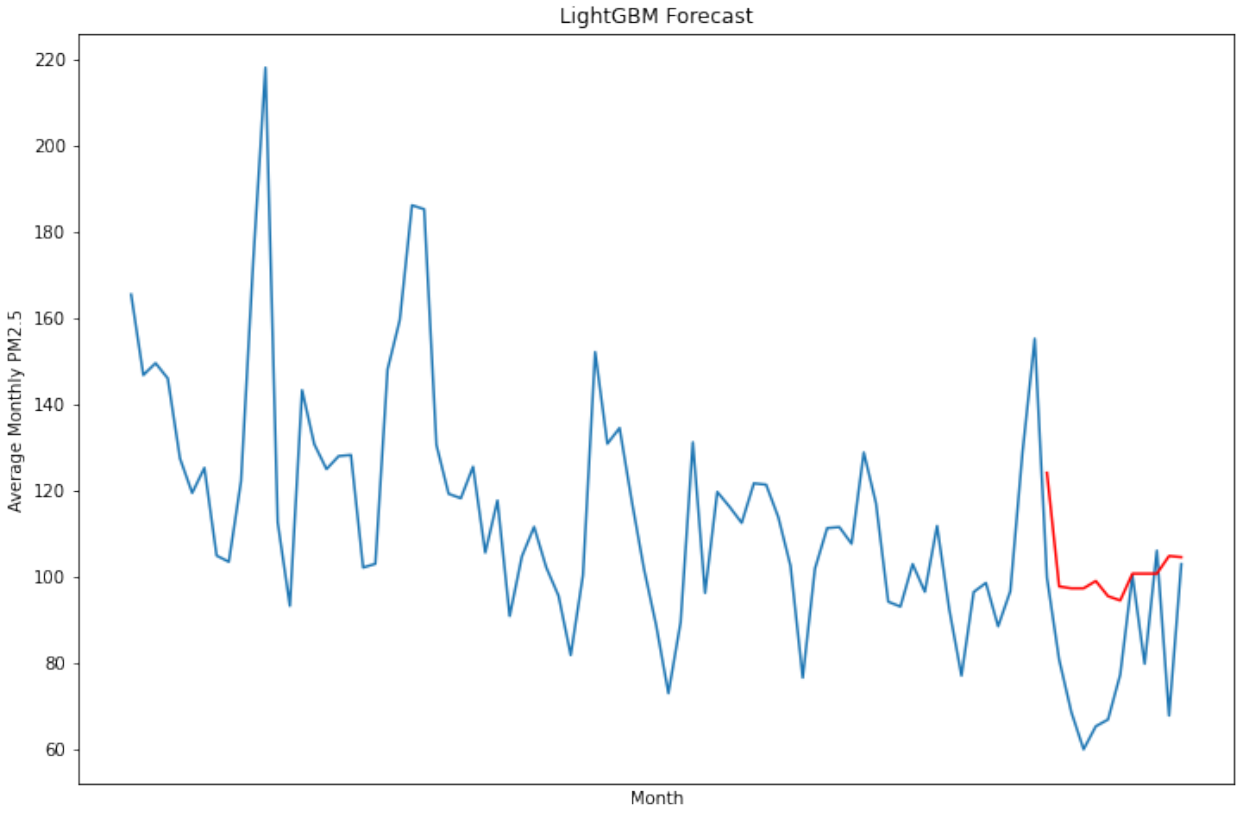

Forecasts from NNAR(3,1,4)[12]
Average Monthly PM2.5
200
150
100
2014
2016
2018
2020
2022
Time

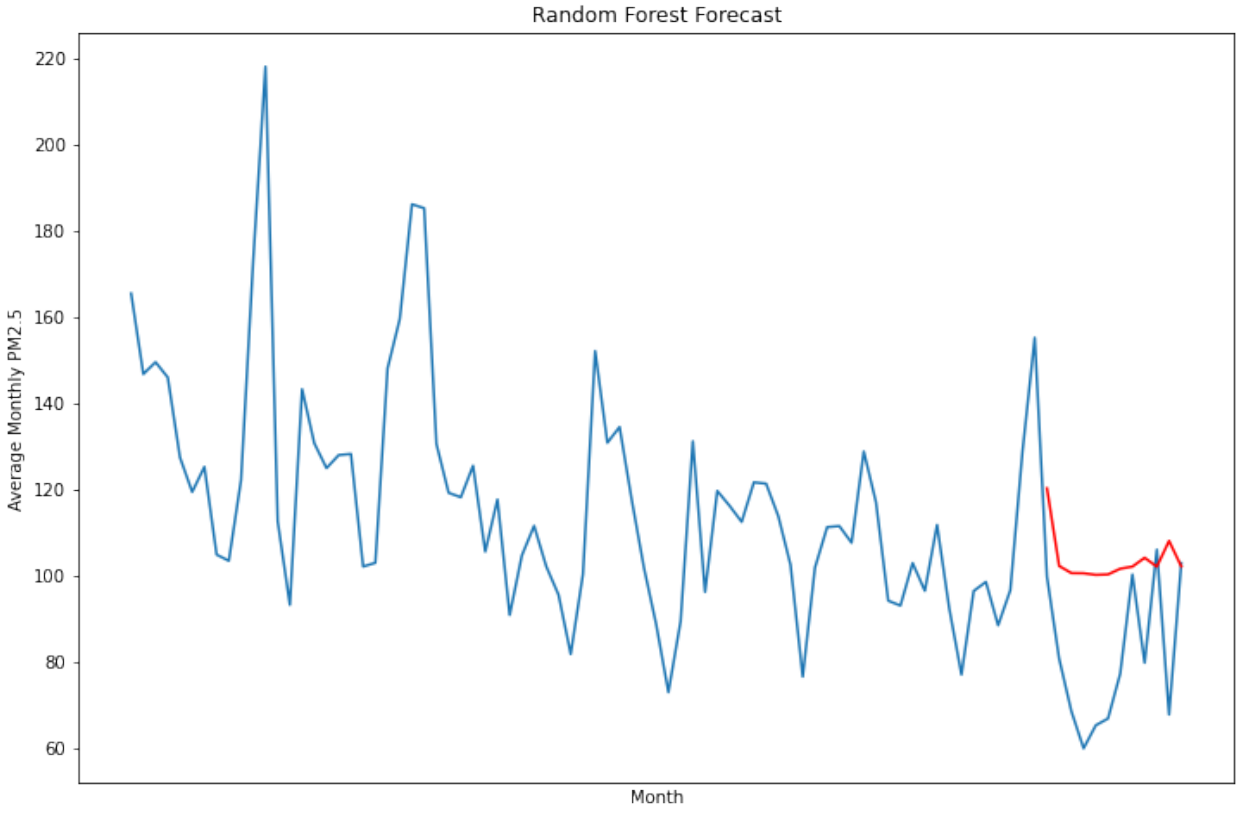


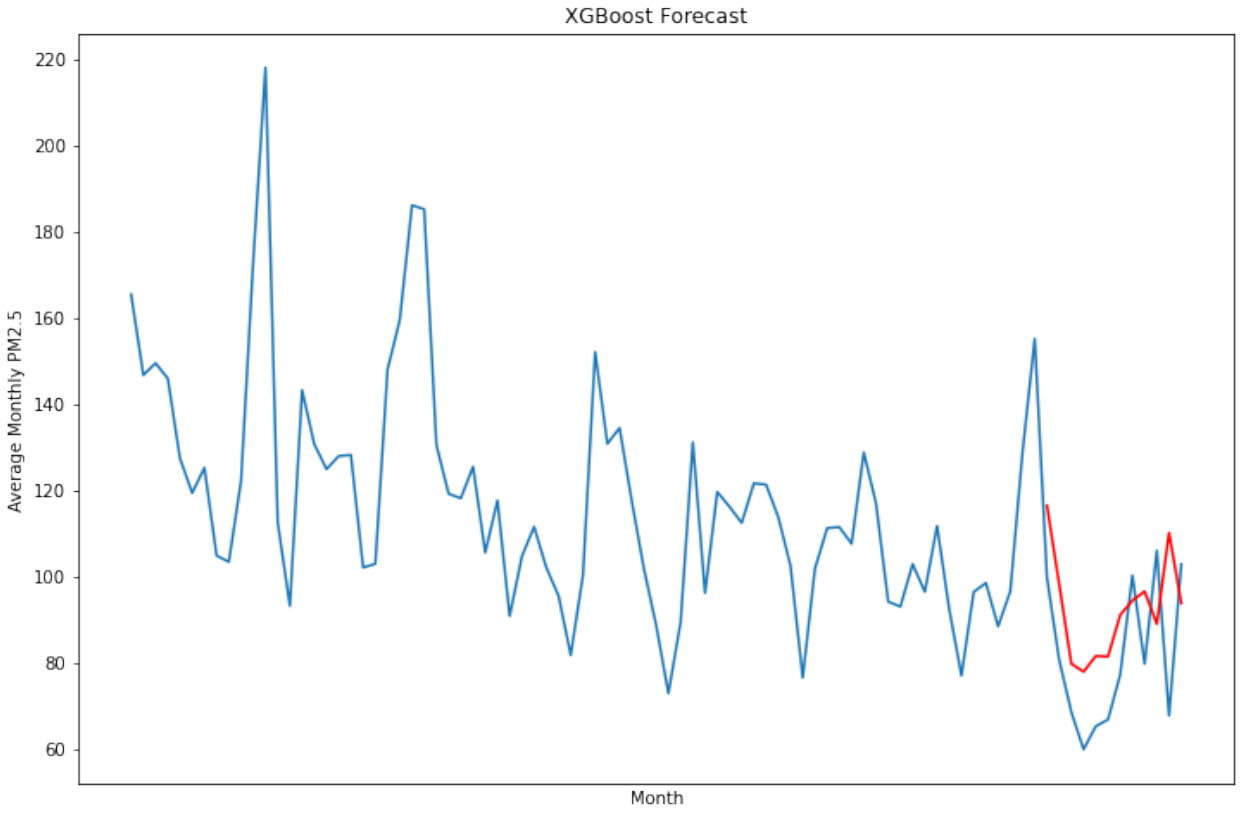
Again, the model doesn't quite capture that dip in the data as well as the previous neural network, but does an okay job at predicting some of the more recent months.

Machine Learning Models

In addition to our more traditional time series models, machine learning models were fit to the data in order to allow for comparison between both methodologies. The chosen machine learning models were Random Forest, LightGBM, and XGBoost - tree-based algorithms, two of which are gradient boosting algorithms (LightGBM, XGBoost). The same two exogenous variables were used (O3, NO2), as well as the past 12 lags (aka months) of data for each data point. Each model was trained in Python using 5-fold cross validation and hyperparameters were tuned using the GridSearchCV function. Forecasts for each model overlaid on the test data is pictured below:

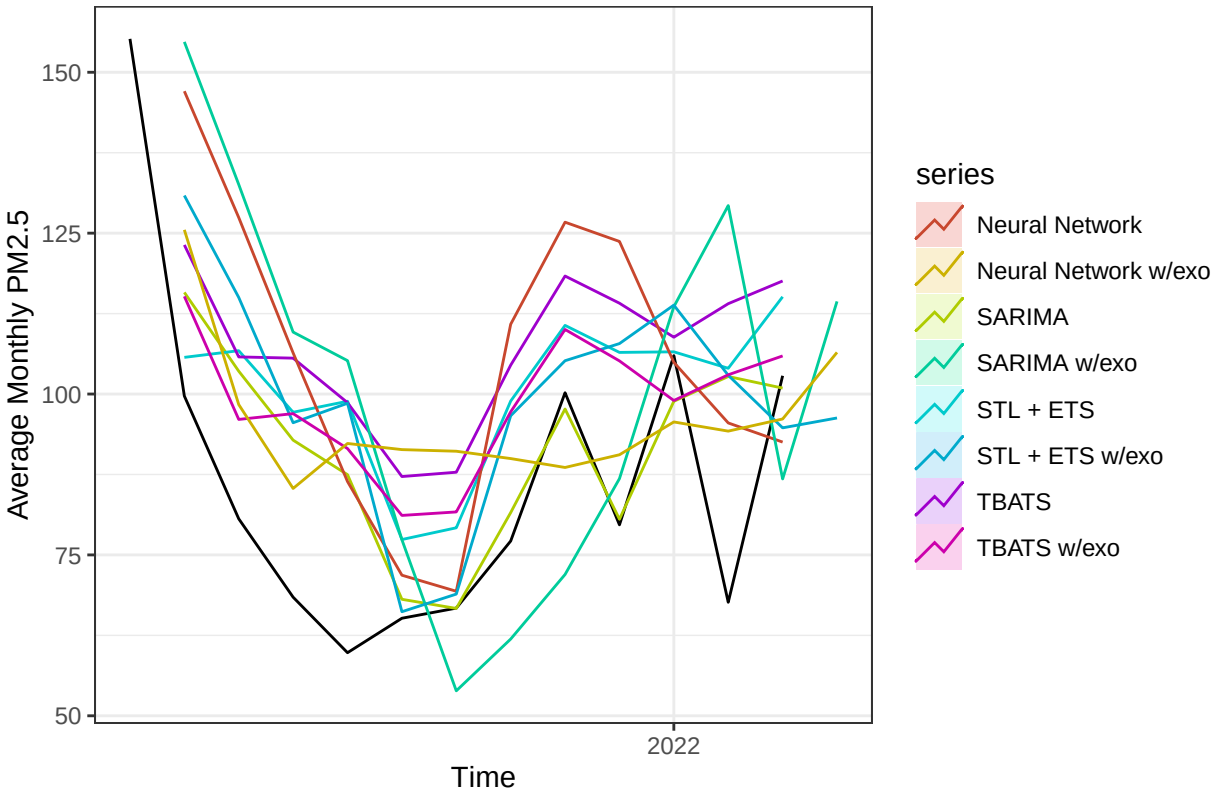

LightGBM Forecast
Average Monthly PM2.5
Month
220
200
180
160
140
120
100
80
60


Random Forest Forecast
Average Monthly PM2.5
Month
220
200
180
160
140
120
100
80
60



Evidently the XGBoost model performed the best out of the three, with the LightGBM and Random Forest models barely capturing that initial dip in the data (much like many of the traditional time series models).

Comparing Models



Overlaying predictions from each of the eight time series models onto the actual data from the past 12 months, we see that the majority of models tend to overestimate, and in particular struggle to capture that initial dip in the data. Again, we note that visually, the SARIMA model seems to do the best job at approximating the test data.

To concretely compare each of the models to one another, we can compute accuracy metrics using the model trained on the training data and evaluated on the test set.

Table 2: Time Series Model Forecast Accuracy for Monthly PM2.5

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
STL+ETS	-19.40562	22.63649	19.40562	-26.80914	26.80914	-0.37869	1.12373
TBATS	-25.97122	28.34812	25.97122	-35.14864	35.14864	-0.56511	1.37628
SARIMA	-10.23515	16.99546	12.16651	-14.87940	16.73838	0.05999	0.76100
NN	-24.06439	30.67891	25.95619	-31.27158	33.10616	0.08012	1.29922
STL+ETS w/exo	-18.51979	24.09067	19.86868	-25.03143	26.34308	-0.15591	1.06926
TBATS w/exo	-17.41038	20.88155	18.57265	-24.39896	25.49578	-0.53588	1.02011
SARIMA w/exo	-17.48508	35.34621	29.54356	-24.91474	38.70926	0.34707	1.48231
NN w/exo	-13.76742	20.16768	18.54008	-20.45327	25.09443	0.02653	1.02907

According to RMSE, the three best models are 1. SARIMA, 2. TBATS w/exogenous variables, and 3. STL+ETS. According to MAPE, the three best models are 1. SARIMA, 2. TBATS w/exogenous variables, and 3. STL+ETS w/exogenous variables. Evidently, in either case SARIMA performed the best, followed by the TBATS w/exogenous variables model. In general, adding exogenous variables did not seem to

significantly help much in terms of model performance - it led to better model performance for the neural network and TBATS models, but worse performance for the other models.

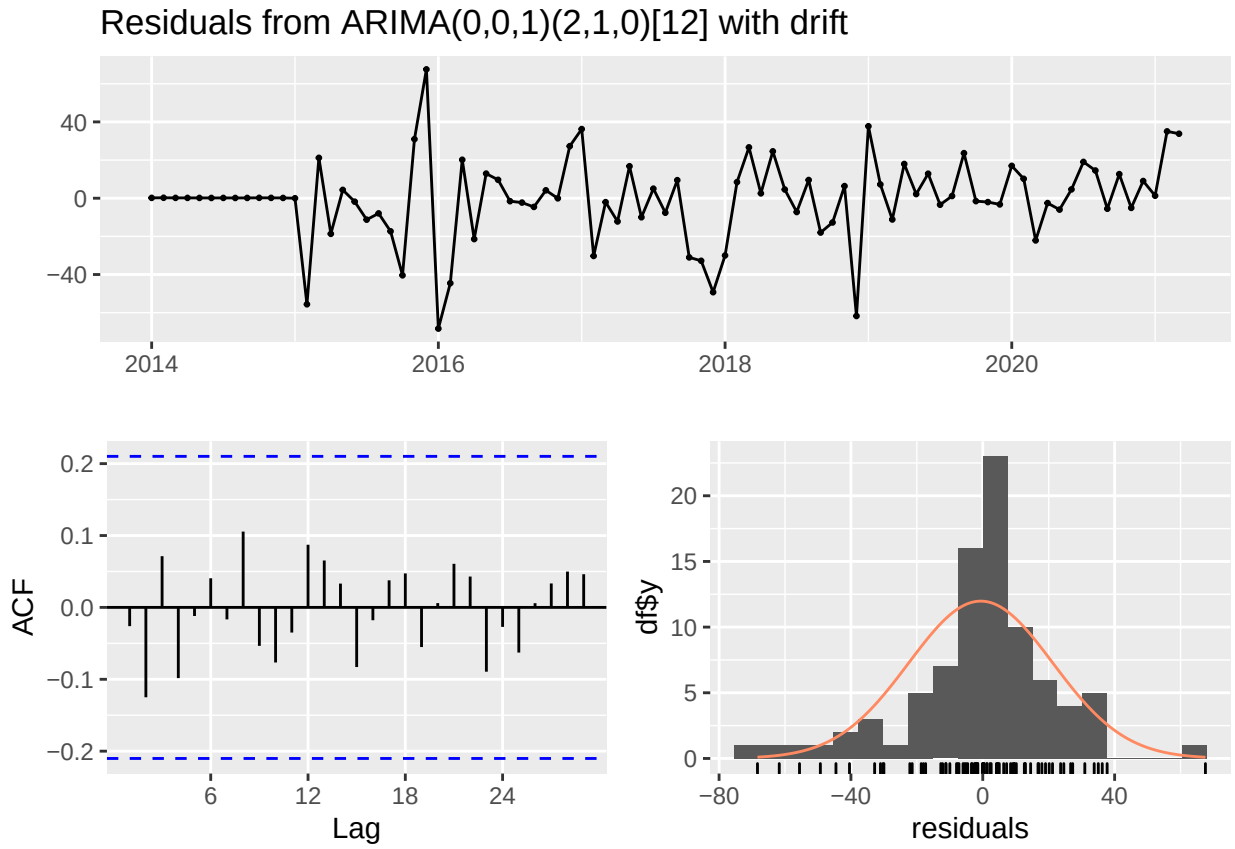
We can also look at the performance of the machine learning models from before.

Table 3: Machine Learning Model Forecast Accuracy for Monthly PM2.5

	RMSE	MAPE
Random Forest	26.99946	32.84750
XGBoost	20.41883	24.36388
LightGBM	23.16184	27.16652

Confirming what we saw before, across both RMSE and MAPE, XGBoost performed the best out of the three ML models, though we note that overall performance of these models ended up being vastly similar to the traditional time series models, with worse performance compared to the SARIMA model.

As a result, we conclude that the SARIMA model does the best job at forecasting the data. To confirm that the model is a good fit, we'll assess its residuals and perform the Ljung-Box test.

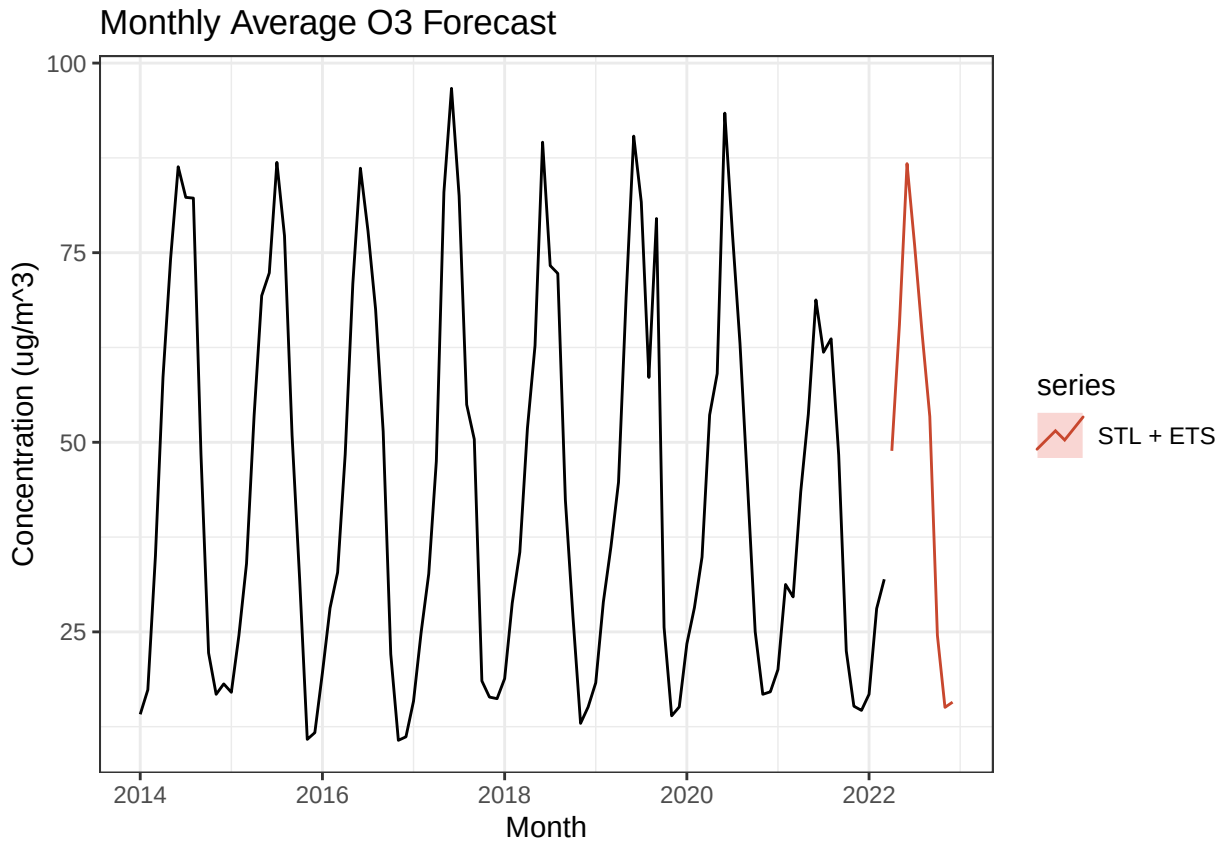


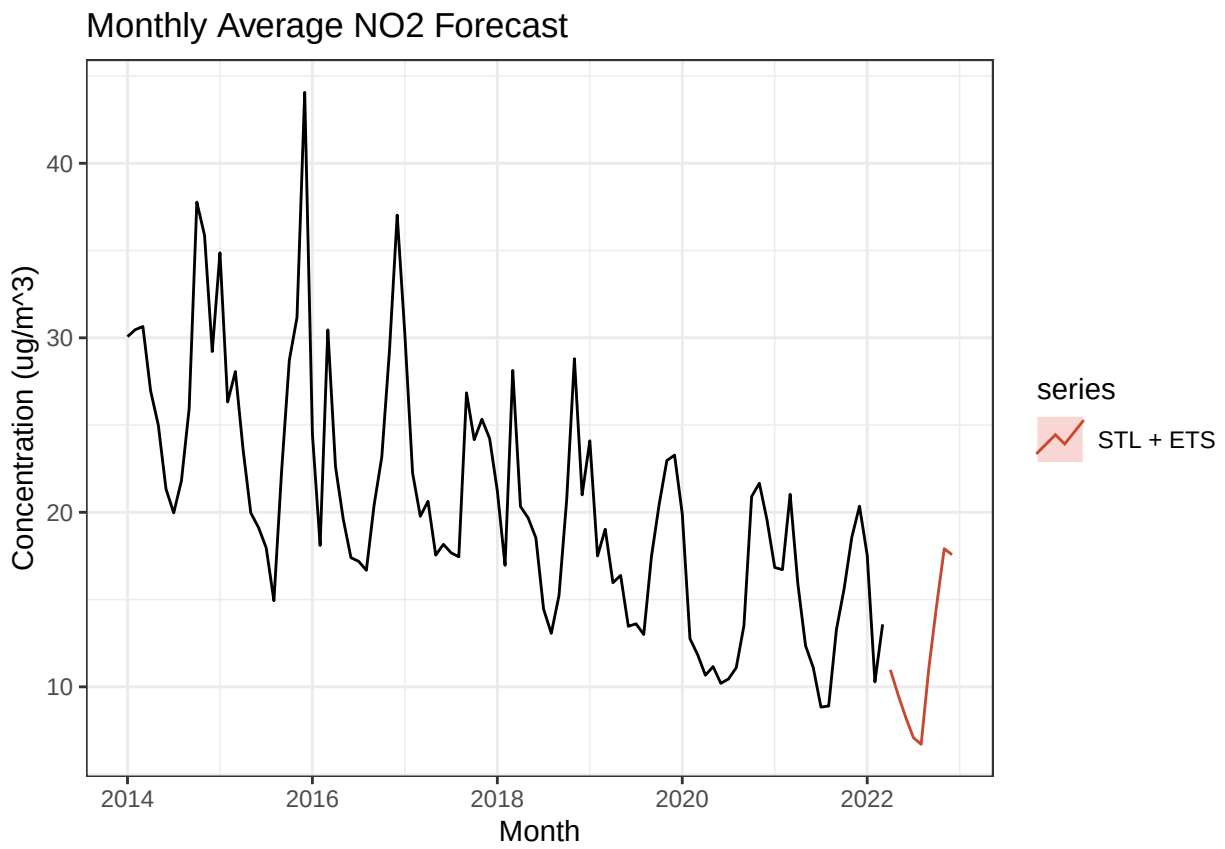

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,0,1)(2,1,0)[12] with drift
## Q* = 7.4212, df = 13, p-value = 0.8792
##
## Model df: 4.   Total lags used: 17
```

Plotting the residual series of the first SARIMA model, we can see that the residual series does seem to look like a white noise series, as we can see from the ACF plot that the majority of ACF values fall within the bounds. Additionally, the Ljung-Box test resulted in a non-significant p-value, which indicates that the model is a good fit for the data.

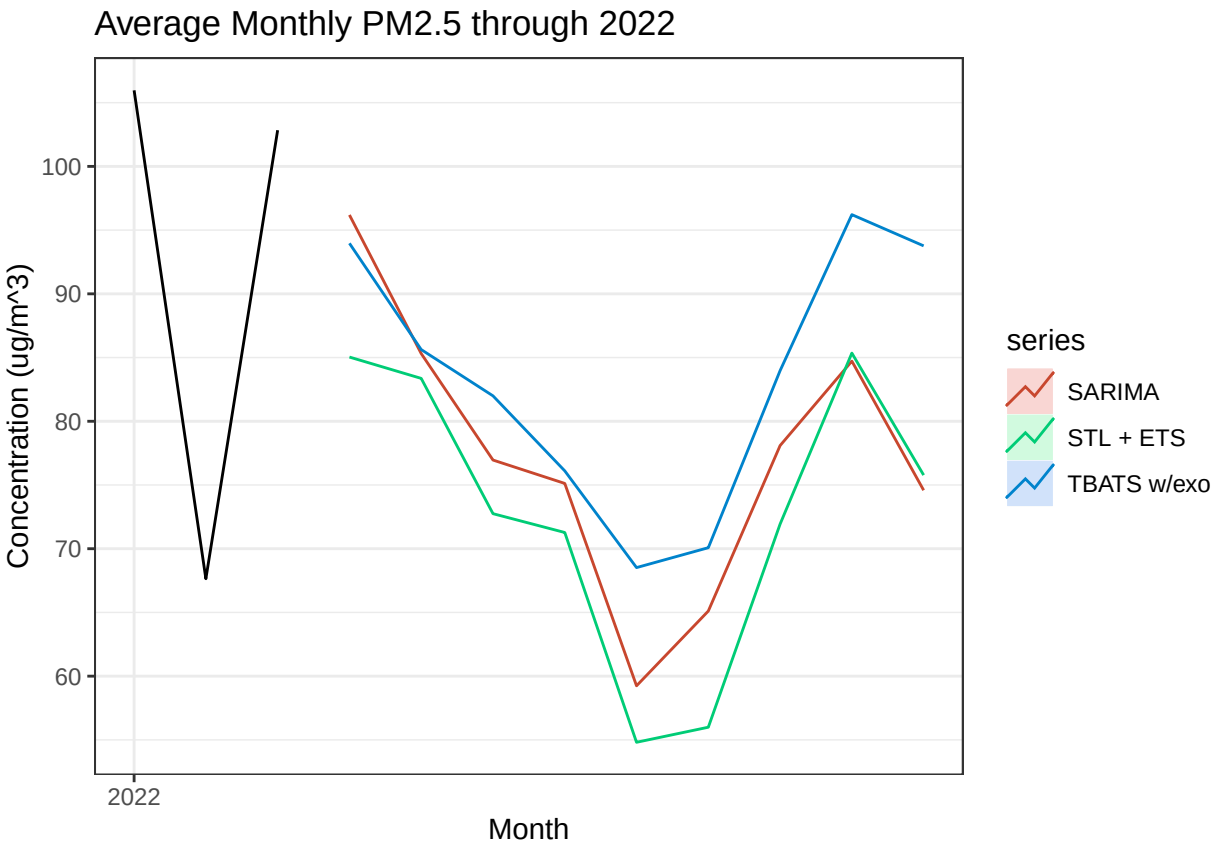
Forecasting through 2022

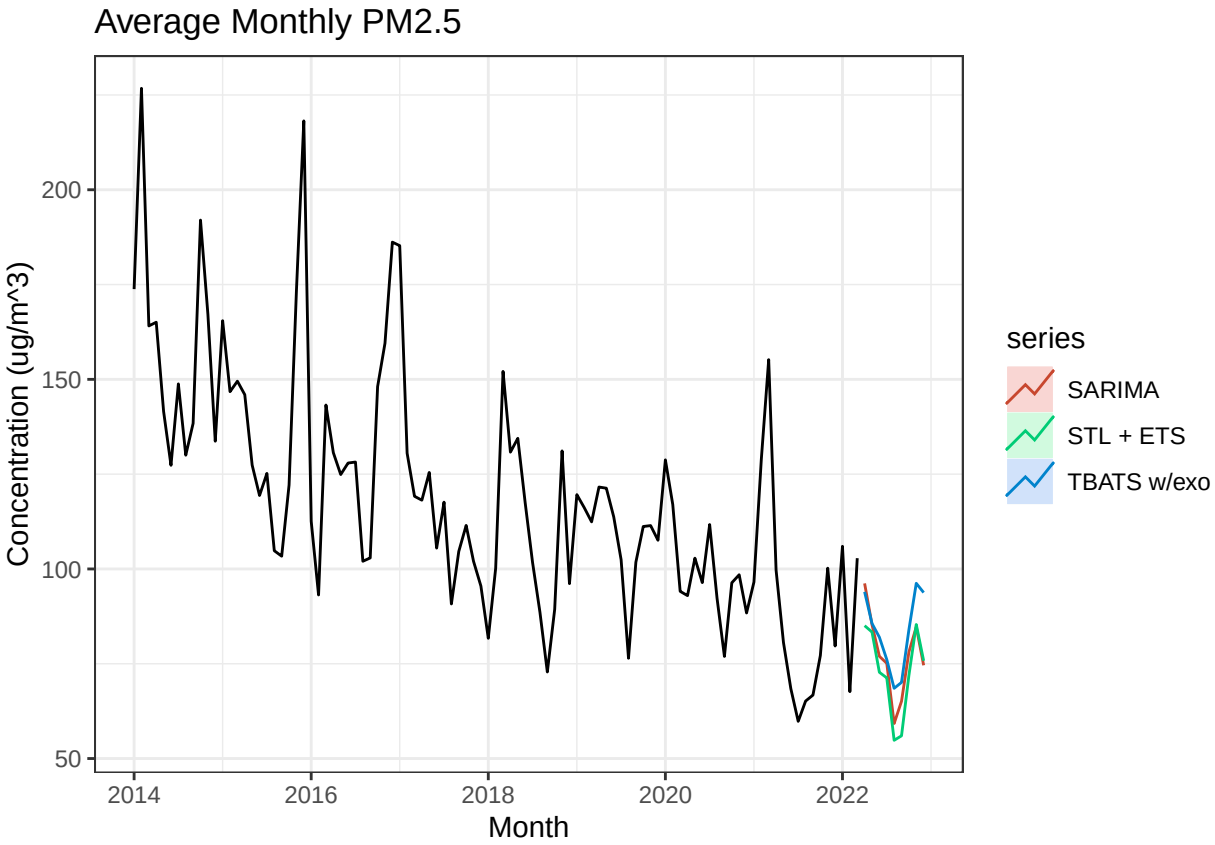
In order to forecast into the future using our models that make use of exogenous variables, we'll need to forecast future values of 03 and NO2. O3 and NO2 values for the next 9 months were forecast using an STL + ETS model, and the plot of their forecasted values is pictured below.





The top three models as identified previously (SARIMA, TBATS w/exogenous variables, and STL + ETS) were fit on the entire data set (January 2014-March 2022), and forecasts were produced for the next 9 months - April 2022 through December 2022.


Average Monthly PM2.5 through 2022
Concentration (ug/m^3)
100
90
80
70
60
2022
Month
series
SARIMA
STL + ETS
TBATS w/exo



Our forecasts through the end of 2022 all follow a similar downward trajectory that seems to line up with the overall trend from previous years. We can see that as it gets closer to the summer months, average monthly PM2.5 concentration decreases to a local minimum, before starting to increase again and reach a peak as winter approaches.

Summary and Conclusions

Overall, we saw that a simple SARIMA model performed best when it came to forecasting and capturing the trend in PM2.5 pollution over time in Beijing. Using exogenous variables did not seem to have much of an effect on overall model performance, though it is important to note that we did not have access to variables such as temperature and rainfall that likely would've provided valuable information to our forecasts and model training process. Our forecasts through the end of 2022 indeed confirm that there is an overall decreasing trend in PM2.5 concentration over time, and the current trajectory expects this trend to continue throughout the end of 2022. We also saw that PM2.5 concentration tends to increase in winter months and decrease in summer months. To improve this model in the future, access to and inclusion of variables such as temperature, rainfall, and other meteorological variables may prove useful towards achieving more accurate PM2.5 forecasts. Data collected from years prior to 2014 may prove useful as well, given that the data set in this project only had 99 observations after aggregation.

There has indeed been a marked performance in air pollution levels in China over the past eight years, with the indication that this downward trend will hopefully continue. However, we do note that the EPA classifies PM2.5 levels at or above 35 ug/m3 during a 24-hour period as unhealthy, and that our forecasts in 2022 are all above 50 each at the minimum. Efforts must continue to be taken in order to maintain decreasing PM2.5 levels, such that a future in which PM2.5 levels are low enough to be in the healthy range is not so distant.