Analysis of Daily PM2.5 Air Pollution Levels in Los Angeles from 2014 through 2019

Lori Kolaczkowski, Chengke Liu, Chen Xi Yang, Véronique Marcotte, Peter Wilson

Introduction

This study employs time series analysis to the examination of an item of environmental concern – daily air pollution (PM2.5 measurements) in the city of Los Angeles, California during a timespan which includes the city's start date of Covid-19 lockdown, March 19, 2020.[1] Initially, we hoped to make a comparison between what PM2.5 levels may have been had the pandemic not occurred, with what PM2.5 levels were actually observed for a segment of time following the city's lockdown. However, limitations in the datasets rendered this goal infeasible, so we decided to move forward with the goal of simply modeling and forecasting daily PM2.5 in Los Angeles. After making adjustments to address missing data, we performed the exploratory data analysis necessary for formulating potential regression models which relate PM2.5 to other relevant variables, and used two different methods to coerce the PM2.5 data to stationarity – detrending using regression with multiple covariates, and differencing. Using the ACF and PACF of the residuals obtained in each method, we then identified potential time series models and selected final models based on model diagnostics and BIC selection criteria. Finally, a selection of different forecasts were made using the final models.

Team members
Each member contributed to model exploration, selection, and forecasting.

Lori Kolaczkowski
I am a local Statistics Master's student with a BS in Marine Biology and 8 years working for NOAA Fisheries Service. Primary project role is writing and theory.
Chengke Liu
I am an online Statistic major student with a Biology background. My primary role in the project is exploratory data analysis and computation.
Chen Xi Yang
I am in my second to last term of the online MS in Statistics. I have a BS in computer science and biology combined major and a MS in experimental medicine (bioinformatics). I currently work as a bioinformatician/biostatistician at the Centre for Heart Lung Innovation of the University of British Columbia. Primary role in the project is programming & theory.
Véronique Marcotte
I'm in the online M.S. in Statistics (and in my last semester yay!) and currently taking this class and the Categorical Data Analysis class. I have a BS in Mathematics and hope to become a Biostatistician after I finish this program. Primary role in the project is programming and preparing presentation slides.
Peter Wilson
I am in my second to last semester in the MS Statistics program. In undergrad, I double majored in Economics and Literature. I spent two years teaching high school math after undergrad. My primary roles in the project are data preparation & research.

Our Data

The air pollution time series data to be analyzed will be in terms of PM 2.5, a parameter often measured when evaluating air pollution levels. Particulate Matter (PM) 2.5 refers to the mass of pollutants “in the air that are two and one half microns or less in width” in a cubic meter of air.[2] Major sources of PM 2.5 include “fireplaces, car engines, and coal- or natural gas–fired power plants.”.[3] Our PM 2.5 data comes from the *World Air Quality Index* at aqicn.org, which in turn relies on data from the United States Environmental Protection Agency.[4] We first selected the data set for PM 2.5, and other air pollution indices (used as covariates to be discussed shortly), measured in downtown Los Angeles (LA), on North Main Street[5], but due to a long, continuous gap in the data, we chose instead to use the *World Air Quality Index* dataset for PM2.5 measured in Reseda, Los Angeles, a location 20 miles northeast of North Main Street. Though more complete than the previous data set, the data set for Reseda PM2.5 and other air pollution indices still had some daily values missing. We settled on a method which would provide air pollution values from a nearby location, Santa Clarita, whose air pollution time series plots were similar to that of Reseda, but with a lower mean. For each date’s missing value in Reseda, we substituted the same date’s value from Santa Clarita, adding to it the difference in the medians of the two series. We made this adjustment using median rather than mean due to a slight skewness in the data. For missing values which were also missing in Santa Clarita, we used linear interpolation within the Reseda series to fill the missing values. The plot of the complete Reseda PM2.5 time series is shown in Figure 1.

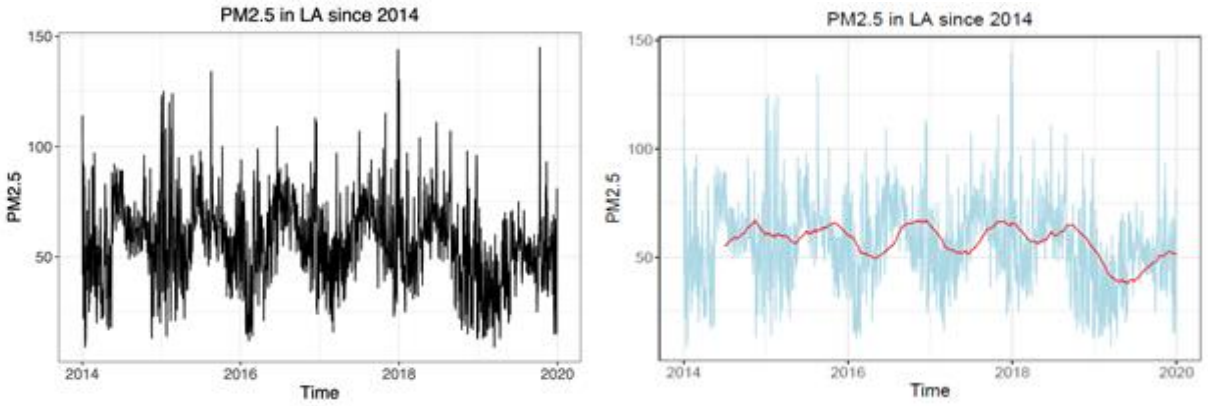

Figure 1. PM2.5 series (on left) and its 6-month moving average in red (on right)

In the interest of having the capability to estimate the mean and autocorrelation functions of the PM2.5 time series data, it is necessary to coerce the series to stationarity, Although a trend was not obvious initially, a 6-month moving average applied to the series, shown in Figure 1, reveals a drop in values at the start of 2019. Furthermore, a non-constant variance is visible in Figure 1, with variability in winter months much higher than that of summer months. And most notably, a strong seasonality is apparent in the clear presence of an annual cycle. We next made attempts to remove these predictable aspects from the data to achieve stationarity.

Addressing Non-constant Variance

We first addressed the non-constant variance. A popular means of stabilizing variance is the log transformation, and we applied this to the PM2.5 series. Since Figure 2 shows no improvement to normal approximation, it does not appear that the log transformation has stabilized the variance. One could say it even slightly worsened the stability of the variance. The reason for this may be that a log transformation is better for correcting a variance that is fanning out over time. Rather than fanning out, the change in variation in the PM2.5 data is seasonal – increasing in winter and decreasing in summer, repeating year after year. Thus, we continued with PM2.5 rather than log(PM2.5).

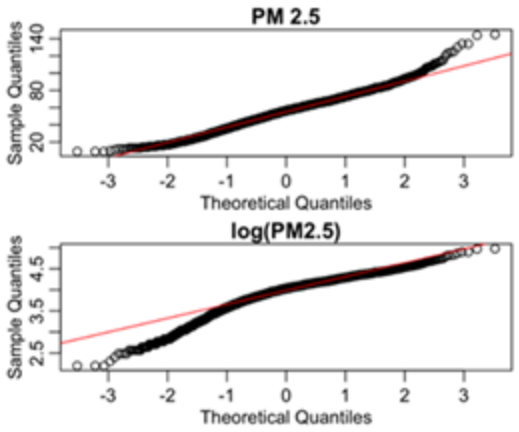

Figure 2. QQ Plots of PM2.5 and log(PM2.5)

To Stationarity: Addressing Trend and Seasonality Via Differencing

First, we applied a 1-day difference to the series, and the resulting series no longer had a seasonal pattern in the mean function, but the variance was still fluctuating in a cyclic fashion. In researching what form of differencing could address this seasonality more effectively, we found that an appropriate method of differencing for this type of series is seasonal differencing (or lag differencing)[1]. Instead of taking the difference between neighboring values, seasonal differencing simply takes the difference between values at the same point in neighboring cycles. In our case, since our daily data shows an annual cycle, we applied a 365-day difference to the original series. The result still showed a strong seasonal component. Upon further reading, we find, “Sometimes it is necessary to take both a seasonal difference and a first difference to obtain stationary data...When both seasonal and first differences are applied, it makes no difference which is done first—the result will be the same.”.

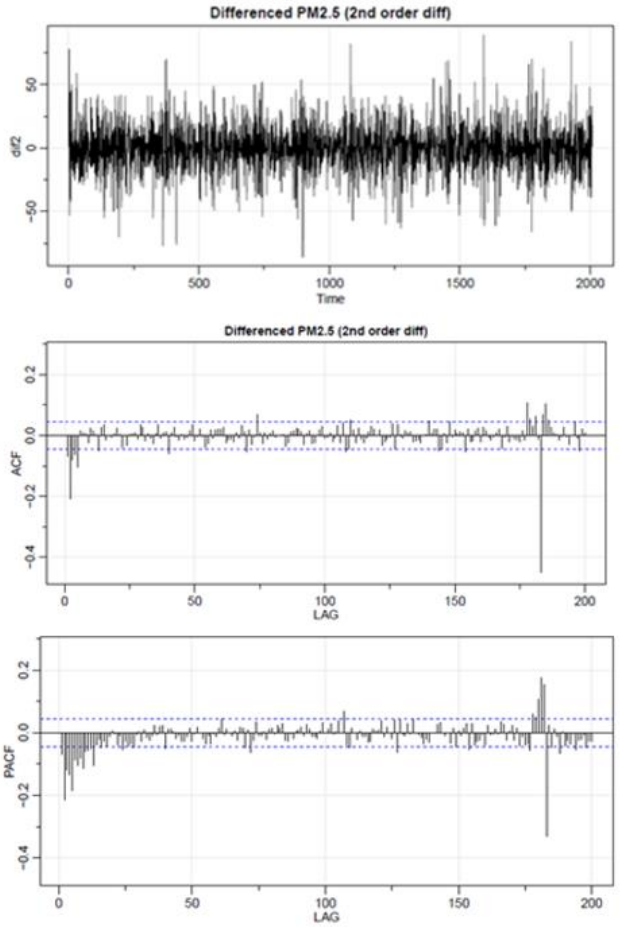

Figure 3. Residuals (top), ACF (middle), and PACF (bottom) of the differenced PM2.5 data

A 1-day difference was applied to the 365-day difference and this improved the plot, but still a fair amount of seasonality was visible. We tried a few different periods for the seasonal difference and, interestingly, a 183-day (half-year) seasonal difference followed by the 1-day difference on the seasonal difference provided a result which appears quite stationary, as shown in the top panel of Figure 3. We suspect this may be due to the fact that, in the original PM2.5 series, variability is much greater in the winter months than in summer months, causing 6-month differencing to be more effective at removing seasonality than annual differencing. The ACF and PACF plots of the second order lag-1 difference on the lag-183 difference on PM2.5 are also shown in Figure 3.

Next, we try an alternative method for coercing the series to stationarity – detrending with regression. In this method, a regression model is formed using some covariates and other terms as predictors for PM2.5. The predicted values from the regression model are then subtracted from the PM2.5 series, leaving its residuals which are checked for stationarity. Before describing this method in detail, we first present our covariates and associated exploratory data analysis.

Our Covariates

We began to explore relationships between PM2.5 and other environmental variables measured daily in Los Angeles which may serve as covariates (predictors) in the regression model used to describe the pattern in the PM2.5 series. These covariates include ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), rain, and maximum air temperature (tempMax). Figure 4 shows each covariate series on the same plot with PM2.5.

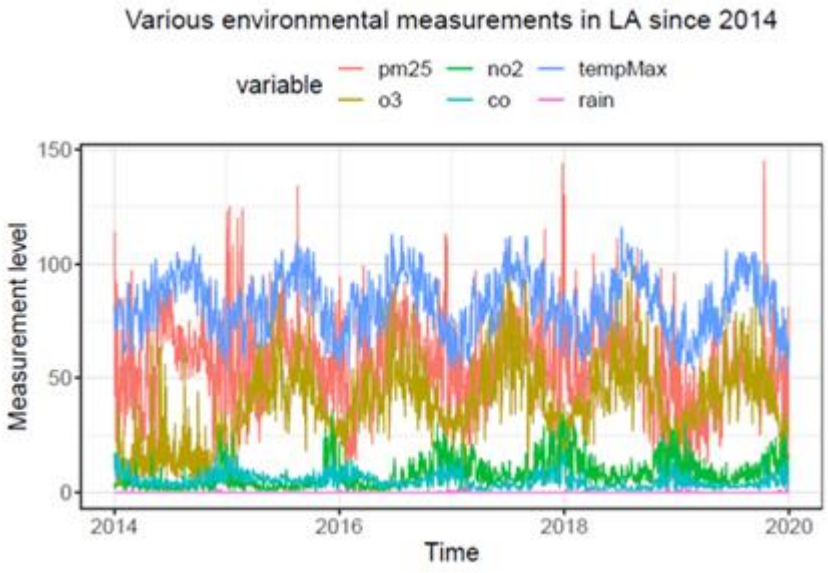

Figure 4. Series of PM2.5 and other continuous environmental variables on same plot

Figure 5 shows the scatterplot matrix used to examine relationships between the potential covariates as well as relationships between each potential covariate and PM2.5. None of the predictors show a particularly strong relationship with PM2.5, but even with mild correlations, we expect the prediction for PM2.5 to improve with inclusion of some of these predictors. With tempMax having the highest correlation with PM2.5, this variable became the primary candidate for the model. For some of the models we would entertain, it was decided O_3 would not be included due to its medium-level correlation with max temperature, and CO would not be included due to its medium-level correlation with NO_2, the variable having a stronger relationship with PM2.5.

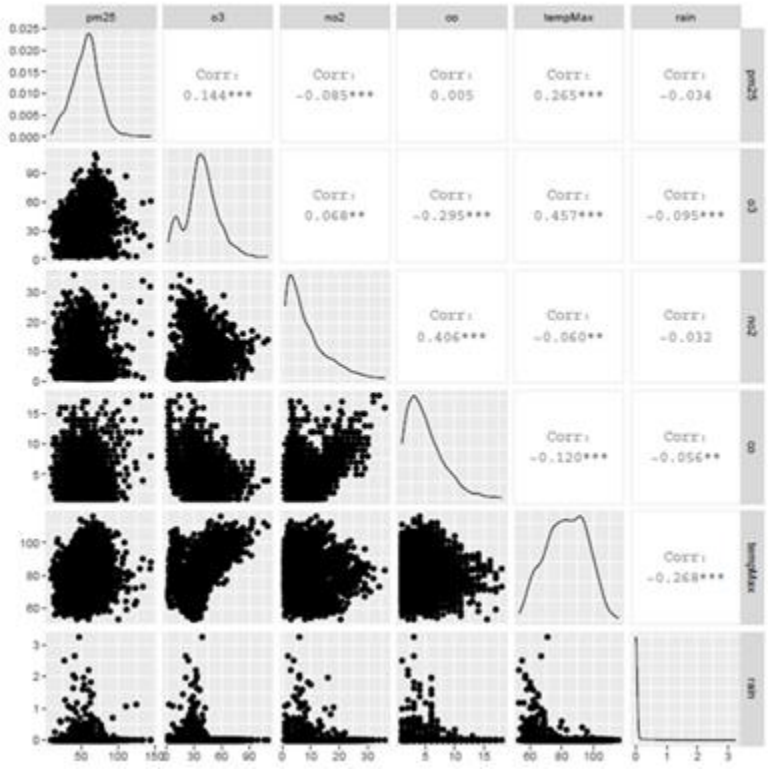

Figure 5. Scatterplot matrix showing relationships between continuous variables

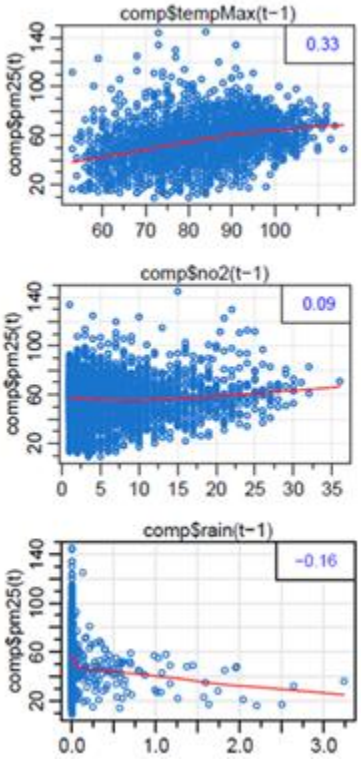

Figure 6. Lag plots for chosen continuous variables

We would also try models including all variables. Lag plots for NO_2, max temperature, and rain were then examined to determine if correlation with PM2.5 might be higher at a certain lag. For all three of those predictors, the correlation is slightly higher at a lag of t-1, which can be seen in comparing Figure 5 and Figure 6. We would try including just these lag-1 predictors in some models, and all continuous predictors at lag-1 in other models. To account for the seasonal nature of our data, we would either include an indicator variable for season or we would include a pair of signal terms, sine and cosine, with a period of 365 days.

To Stationarity: Addressing Trend and Seasonality Via Detrending with Regression

Based on examination of the scatterplot matrix and lag plots, and the seasonal nature of our data, we decided to investigate the following regression models for the pattern in the PM2.5 data.

model	AIC	BIC	R2
1. all predictors + season indicator	18619.25	18681.86	0.1812941
2. NO2, temperature, rain + season indicator	18651.63	18702.86	0.1675851
3. all predictors + sinusoidal terms	18593.25	18650.18	0.1902113
4. NO2, temperature, rain + sinusoidal terms	18637.08	18682.62	0.1723404
5. all lag-1 predictors + season indicator	18262.71	18325.32	0.2988804
6. lag-1 NO2, temperature, rain + season indicator	18392.45	18443.68	0.2547297
7. all lag-1 predictors + sinusoidal terms	18131.25	18188.17	0.3391263
8. lag-1 NO2, temperature, rain + sinusoidal terms	18332.92	18378.45	0.2740540

The following model (7) was chosen, due to lowest AIC, lowest BIC, and highest R^2.

$\hat{P}_t = 63.75 - 0.01_{(0.0008)}\text{trend} + 0.87_{(0.0756)}N_{t-1} + 2.00_{(0.1553)}CO_{t-1} + 0.17_{(0.0254)}O_{t-1} - 0.25_{(0.0433)}T_{t-1} - 2.64_{(1.7461)}R_{t-1} - 16.42_{(0.8318)}Z1 - 5.75_{(0.5476)}Z2$

Where P_t = pm25, trend = time(pm25) , T_{t-1} = lag 1 maxTemp, R_{t-1} = lag 1 rain, N_{t-1} = lag 1 NO_2, CO_{t-1} = lag 1 CO, O_{t-1} = lag 1 O_3, Z1 = cosine signal term with period of 365 days, and Z2 = sine signal term with same period.

Our method for detrending the PM2.5 series was to subtract model 7 seasonality pattern from PM2.5 with the goal of arriving at the residuals of the series, which theoretically should be stationary white noise. The top panel of Figure 7 shows that a reasonable degree of stationarity was achieved in this transformation, since the trend and seasonal pattern are not obviously apparent. The ACF (middle panel) and the PACF (bottom panel) of the residuals of model 7 are also shown in Figure 7.

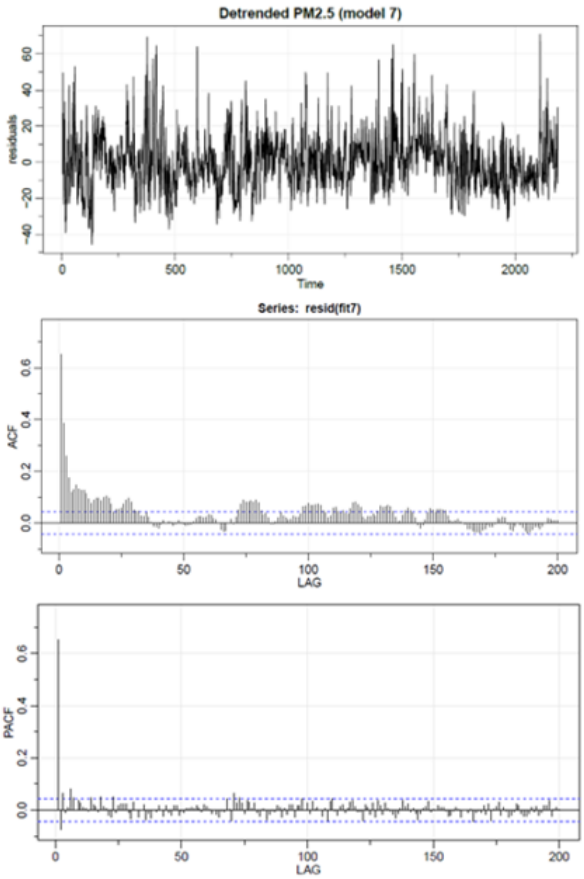

Figure 7. Detrended PM2.5 Using Model 7

Potential Time Series Models: Regression with Autocorrelated Errors

In examining the ACF and PACF of the model 7 residuals, a time series model can be identified and the regression model (model 7) can be fit to the original PM2.5 series again, but with errors following the correlation structure of the selected time series model. This is done by specifying the original PM2.5 series as the first argument, and specifying the regression predictors in the xreg argument of the sarima function in R. The quality of this fit can then be assessed in examining diagnostic plots. This time series modeling approach is known as regression with autocorrelated errors. In the next section we will describe the time series modeling approach using the differenced series – Mixed Seasonal ARIMA. Based on several significant autocorrelations from lag-1 to lag-7 in the PACF of model 7 residuals (see Figure 7), we decided to try fitting an AR(7) model to the stationary residuals obtained from model 7. Since the PACF, showing significant correlations at lag-1, lag-2, and lag-3 also suggests an AR(3) as a potential model, we evaluated this fit as well.

Potential Time Series Models: Mixed Seasonal ARIMA

In examining the ACF of the differenced PM2.5 series (see Figure 3), one highly significant correlation is visible in the ACF at lag-183, and significant correlations are present at lags 1 to 5, with virtually all other lags showing correlations that are 0. This is indicative of a seasonal component which follows MA(1) with s=183, and a nonseasonal component which follows either MA(4) or MA(5). Thus, the following two mixed seasonal ARIMA models will be fit to the differenced PM2.5 series and evaluated for goodness of fit: ARIMA(0,1,4)x(0,1,1)$_{183}$ and ARIMA(0,1,5)x(0,1,1)$_{183}$.

Choosing the Best Time Series Model

Mixed Seasonal ARIMA Method

The Ljung-Box test for the ARIMA(0,1,4)x(0,1,1)$_{183}$, shown in Figure 8, shows a fairly large portion of significant p-values, indicating a fair amount of correlation in the residuals. The Ljung-Box p-values for the ARIMA(0,1,5)x(0,1,1)$_{183}$, also shown in Figure 8, are a bit larger and fewer dip below the significance level of .05. BIC favors the ARIMA(0,1,5)x(0,1,1)$_{183}$ over the ARIMA(0,1,4)x(0,1,1)$_{183}$ since their values are 7.470291 and 7.476103, respectively. Since the diagnostics and the BIC were more attractive for the ARIMA(0,1,5)x(0,1,1)$_{183}$, it was the chosen model for the mixed seasonal ARIMA approach.

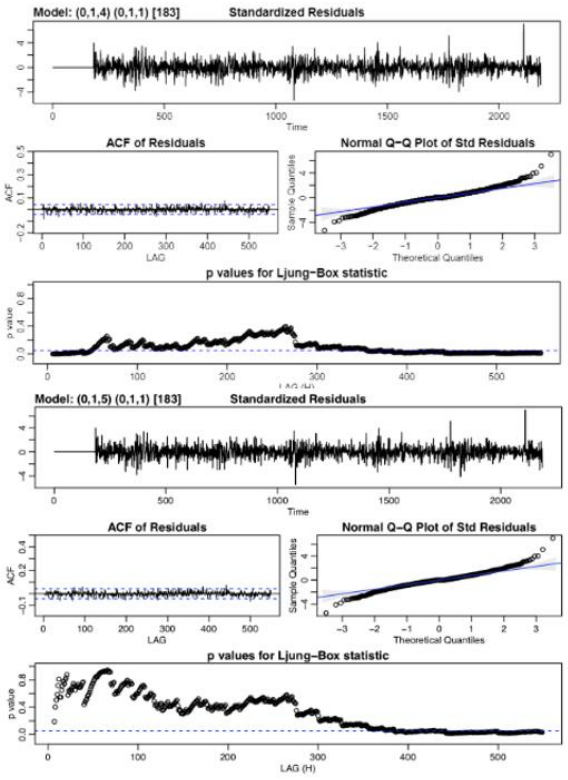

Figure 8. Diagnostic Plots for ARIMA(0,1,4)x(0,1,1)$_{183}$ (top) and ARIMA(0,1,5)x(0,1,1)$_{183}$ (bottom)

Regression with Autocorrelated Errors Method

The diagnostic plots of the AR(7) fit showed no departure from white noise (see Figure 9). However, the diagnostic plots for the AR(3) fit showed too much correlation remaining in the residuals of the fit, according to the Ljung-Box p-values (shown in Figure 9). In pursuit of a decent model with a lower order, we also tried fitting AR models with the following orders: 1, 2, 4, 5, and 6. The diagnostics of all but the AR(6) were not any better than the AR(3), however the diagnostics of the AR(6) (shown in Figure 9) appeared to be acceptable and its BIC, 7.727103, was extremely close to the BIC of the AR(7), 7.728225. Thus, in the interest of reducing the number of model parameters, we chose the AR(6) for the model 7 regression with autocorrelated errors approach. The model is given by the following:

$$\widehat{P}_t = 63.75 - 0.01_{(0.0008)}\text{trend} + 0.87_{(0.0756)}N_{t-1} + 2.00_{(0.1553)}CO_{t-1} + 0.17_{(0.0254)}O_{t-1} - 0.25_{(0.0433)}T_{t-1} - 2.64_{(1.7461)}R_{t-1} - 16.42_{(0.8318)}Z1 - 5.75_{(0.5476)}Z2 + \varepsilon_t$$

where $\epsilon_t = 0.7325_{(.0217)}\epsilon_{t-1} - 0.1353_{(.0267)}\epsilon_{t-2} + 0.071_{(.0267)}\epsilon_{t-3} - 0.0033_{(.0268)}\epsilon_{t-4} + -0.0530_{(.0267)}\epsilon_{t-5} + 0.0861_{(0.0215)}\epsilon_{t-6} + w_t$

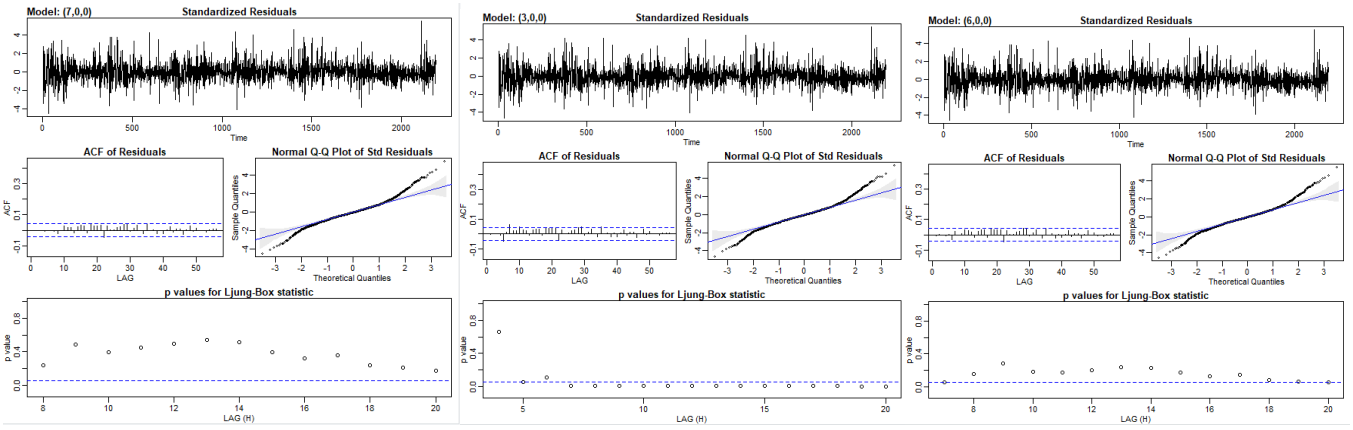

Figure 9. Fit diagnostic plots for AR(7)(left), AR(3)(middle), and AR(6)(right)

Forecasts

First, we computed a 4-step-ahead (4-day) forecast for PM2.5 measurements using our two best models: regression model 7 with AR(6) errors, and the mixed seasonal model ARIMA(0,1,5)x(0,1,1)$_{183}$. We did this by excluding the last 4 days of 2019 from each model, and then proceeded to forecast the excluded days, so that forecasted values and their associated prediction intervals could be compared to actual observed values. This provides a rough estimate of forecast model performance. We then used the better-performing model to obtain a long-term forecast (1 year). Finally, we make a couple of hypothetical forecasts with this model, using a mock covariate with extreme values.

4-Day Forecast: Regression Model 7 with AR(6) Errors

Table 1 shows the four forecasted values, their 95% prediction intervals, and the actual observed value of PM2.5 for that day. All four actual values fall within the forecast interval. Figure 10 shows the plot of the forecast.

Table 1. 4-step-ahead forecast results for model 7 with AR(6) errors

	Forecasted PM2.5	95% P.I. (Lower)	95% P.I. (Upper)	Observed PM2.5
Day 1	25.34241	3.355757	47.32906	15
Day 2	45.37193	18.14055	72.60331	53
Day 3	52.90261	24.27742	81.5278	81
Day 4	53.11451	23.89144	82.33758	81

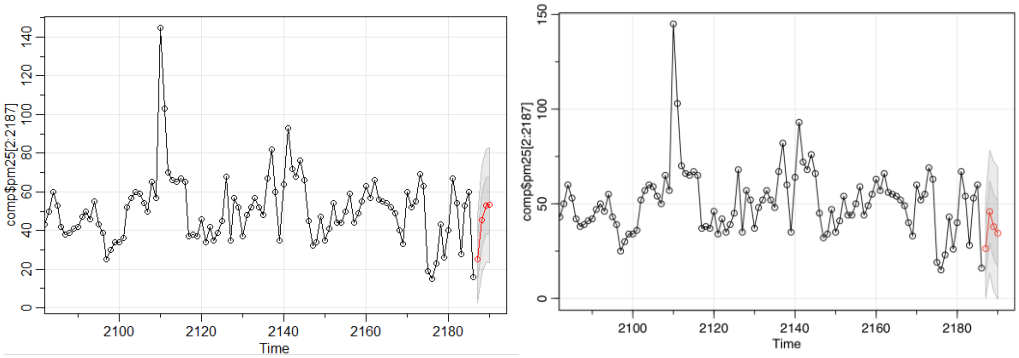

Figure 10. 4-step-ahead Forecast plots for regression model (left) and mixed sarima (right)

4-Day Forecast: Mixed Seasonal ARIMA(0,1,5)x(0,1,1)$_{183}$

Table 2 shows the four forecasted values, their 95% prediction intervals, and the actual observed value of PM2.5 for that day. The last two actual values do not fall within the forecast interval. Figure 10 shows the plot of the forecast.

Table 2. 4-step-ahead forecast results for the mixed seasonal model ARIMA(0,1,5)x(0,1,1)$_{183}$

	Forecasted PM2.5	95% P.I. (Lower)	95% P.I. (Upper)	Observed PM2.5
Day 1	26.32422	0.703439	51.94500	15
Day 2	45.83864	13.93771	77.73957	53
Day 3	37.88331	4.355672	71.41094	81
Day 4	34.46924	0.243895	68.69459	81

Since the forecast prediction intervals of the regression model 7 with AR(6) errors covered twice as many observed values as did the mixed seasonal ARIMA model, we chose the regression model as the better-performing forecast model. Figure 11 shows a 365-day forecast provided by regression model 7 with AR(6) errors. Note that 351 out of the 365 actual PM2.5 values (96%) fell within the 95% prediction interval of the forecasted value for that day, despite the fact that actual values of PM2.5 for the year forecasted were uncharacteristically low with respect to values of previous years. See Figure 11 for plots of forecasted (top) versus actual (bottom) PM2.5. As a side item of interest, we also ran a couple of additional 365-day forecasts using mock covariate values to examine how they might influence PM2.5 levels. In the first scenario, we made all rain values equal to the maximum in our dataset. The result was that PM2.5 was forecasted to be quite a bit lower. In the second scenario, we increased all maximum air temperature readings by 20°F. Interestingly, a hike in temperatures did not result in any noticeable change to the PM2.5 forecasts. Plots of these hypothetical forecasts are shown in Figure 12.

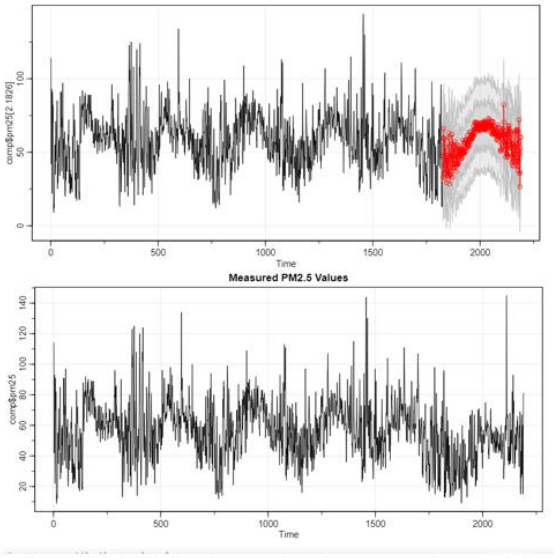

Figure 11. 365-day forecast using regression model 7

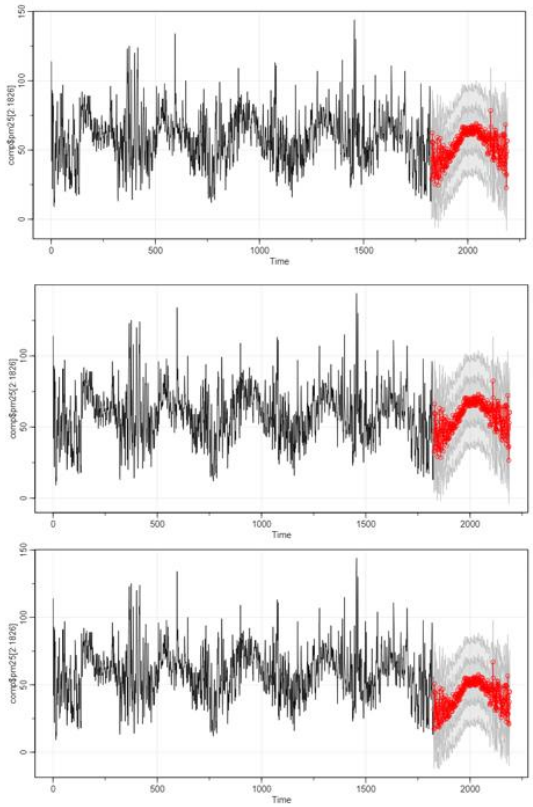

Figure 12. High temperature (top), original (middle) and high rain (bottom)

Conclusion

In trying models with and without covariates, for our PM2.5 series, covariates and our sinusoidal predictors seem to vastly improve the model's forecast performance. Our regression with autocorrelated errors model was even able to produce long-term forecasts consistent with the past pattern.

Some future considerations for possible improvement are the following: If a wind dataset became available for Los Angeles, we would explore how much it might improve regression model 7 and its forecasts, if it were used as a covariate. We'd like to try a SARIMAX modeling approach and see how the results of this type of model would compare. To assess what the limitations of our forecasting model might be in terms of application to predicting air pollution levels in other cities, we might compare our results to that of similar studies.