Using autoregression time series analysis to predict indoor air quality

630012729

May 16, 2022

1 INTRODUCTION

Air pollution is a major issue in many cities around the world and can have serious health implications for people living in these cities. In 2016, the World Health Organisation (WHO) released the Ambient air pollution report in which they declared air pollution as the 'biggest environmental risk to health' [1]. It is estimated that in Europe, over 500,000 deaths were associated with exposure to air pollution in 2016 and indoor air is likely to contain higher levels of pollution than outdoor air [2]. This of particular concern since the average adult is estimated to spend over 80% of their time indoors [3].

Exposure to poor air quality can lead to many respiratory illnesses as well as exacerbating allergies and cancerous diseases [2]. Furthermore, certain sizes of particulate matter can be linked to specific health issues. Particulate matter less than 2.5 microns in size (PM2.5) are associated to health effects such as pulmonary diseases, cardiovascular disease, diabetes, and adverse birth outcomes [4]. The majority of PM2.5 comes from traffic-related sources however the main indoor sources are cooking and heating in the home or building [5]. While outdoor air pollution can enter the indoor environment through ventilation, the most effective measure to reduce indoor air pollution is to increase ventilation [2].

In the UK, the Air Quality Standards Regulations 2010 state that the safe levels of exposure to PM2.5 must not exceed an annual average of 20 micro-grams/m3 [6]. While air pollution has been improving in the UK [6], an estimated 28,000 deaths in the UK are attributed to air pollution. As such, being able to accurately measure and track the levels of air pollutants in various indoor and outdoor locations is essential to coming up with a strategy to reduce the deaths and health impacts associated with air pollution. This analysis aims to use indoor air pollution data provided by Smartline to build a predictive model that can estimate future daily air pollution levels. The ability to predict potential spikes in air pollution in advanced would allow home owners to take effective action, such as opening windows or using an air filter, to avoid being exposed to large amounts of PN2.5 and other air pollutants.

2 DATA SOURCES

The air pollution data used for the analysis was collected as part of the Smartline project. This was a 5-year research project funded by the European Regional Development Fund to research health and well-being in Cornwall. Air pollution sensors were installed in 238 houses in Cornwall which recorded the PM2.5 levels in the home at a 5 minute intervals from 2018 - 2020. Over this time period, however,, there was faults in the sensors so the sensors didn't always record at regular 5 minute intervals. This meant that each csv file for each house had a different number of observations. To deal with this, the data was re-sampled on a daily level. For this analysis the data from UPRNS 0011 and sensor id 6322070 was used.

Additional weather data was also used for the analysis. This data was obtained from the Centre for Environmental Data Analysis (CEDA) Archive. The CEDA control the UK's national atmospheric and environmental data including long-term weather data. The data used in this analysis was the observed historical weather data for the Cambourne area. More specifically, the hourly data for 2018, 2019, and 2020 was used. This data included hourly temperature, pressure, humidity, and wind data for the area of Cambourne and was in csv format. This data was also resampled to a daily level.

3 DATA ANALYSIS/MODELLING METHODS

For the group project, several analysis methods were chosen. A Fourier Transform of the data was performed to try to identify any cyclical patterns in the data and time-series analysis was used to compare the lag between outdoor air pollution and indoor air pollution. The results of these methods will not be discussed in this paper. Time series regression was also explored during the project to try and predict the next hour of PM2.5 however due to limited time there was not much progress on this model. As such, this paper will explore using time series regression models to predict the next daily level of PM2.5 indoors.

Kang et al. reviewed several machine learning methods for predicting air quality [7]. The authors assessed the appropriateness of artificial neural networks (ANN), decision tree, random forest, deep belief network and support vector machine (SVM)

models to predict various components of air pollution. These models are more complex and would be more effective for larger data with more historical observations. As such, this report started with a more basic autoregression (AR) model and built upon this by adding in moving average and multivariate analysis to form autoregressive integrated moving average (ARIMA) models. ARIMA models have been shown to be a useful analysis method for predicting air quality for both univariate and multivariate scenarios. Kumar and Goyal used a combination of an ARIMA and principle component regression model to predict air pollution levels in Delhi [8] and Abhilash et al. used ARIMA models combined with weather data to predicted air quality in Bengaluru [9].

AR models predict future values of a univariate time series and are regressed using previous values from the time series:

$$y_t = \beta_0 + \beta_1 y_{t-1} ... \beta_k y_{t-k} + \varepsilon_t \tag{1}$$

where β_k is the coefficient for the k^{th} time series value, y_{t-k} is the previous value for $k = 1, 2, 3...l$, and ε_t is the error term. The order of an AR model determines how many previous values to perform the regression on and this is also known as 'lag'. For example, a second order AR model, denoted as AR(2), would calculate the y_t value based on the previous 2 y values.

ARIMA models include the AR component and also have a moving average (MA) component. An MA model uses the error terms of each value to predict a new value and can be defined as:

$$y_t = \mu + \sum_{i=1}^{k} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{2}$$

where μ is the mean of the series, θ_i are the coefficients of the values, and ε_{t-i} are the error terms.

Before any models were created, an initial analysis of the data was conducted to explore any trends in the data and to get an initial idea of what model might be most suitable. This included performing a seasonal decomposition on the data to check for seasonality and trends within the data, and the residuals when these aspects were extracted. Additionally, to determine the order that should be used, autocorrelation and partial autocorrelation plots where created. This provided more information about the relationship between the time series data but also influenced the model selection. An ARIMA model takes the assumption that the data is stationary and hence an Augmented Dickey-Fuller test was used to determine the stationarity of the data. The integrated (I) component of an ARIMA model corresponds to the order of differencing the model requires to ensure it is stationary.

The analysis was conducted using Python 3.9 [10] and the `statsmodels` [11] package to create the models. `statsmodels.tsa.ari` and `statsmodels.tsa.statespace.SARIMAX` methods were used to create the various AR, MA, ARIMA, ARIMAX, SARIMA, and SARIMX models. These functions define the components to be included in the model with parameters where p, d, q are the AR, order of differencing and MA components respectively. The analysis began by creating an AR model, then an MA model, ARIMA model, and ARIMAX model by altering the relevant parameters. This was to assess the performance of each variation of the model to see if including different components improved the performance of the model. From here the effects of including seasonality were then explored by using SARIMA and SARIMAX models. Models with an 'X' component are multivariate models that include the weather data and use this, as well as the AR and MA components to predict future values. Multivariate ARIMAX models can be defined as:

$$y_t = \beta x_t + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{3}$$

where x_t is the other variable included in the model and β is its respective coefficient and p and q terms are the orders of the AR and MA components respectively.

A multivariate analysis using weather was performed since studies suggest that the weather can have a significant influence over indoor air pollution [12][13]. This is due to things like air temperature, wind speed and air pressure affecting the dynamics and flow of pollutants in the air, influencing how long they remain in a given area. It also affects the behaviour of people in their home; colder weather normally results in less ventilation of the home and the burning of heating fuels causing further pollution.

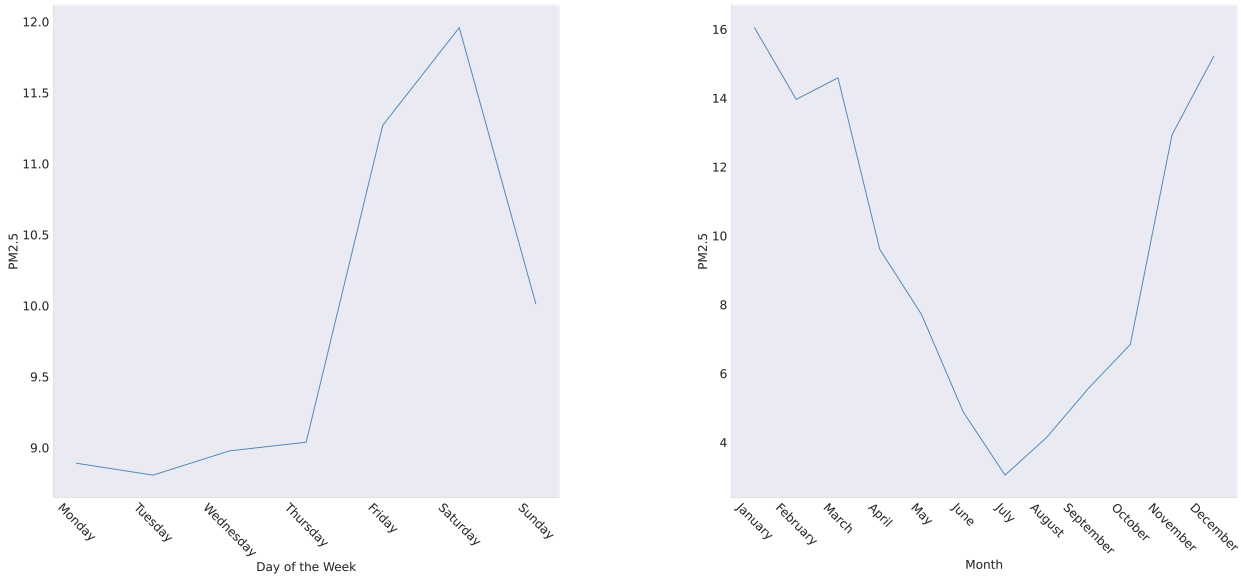
The data was split into a training and testing set with the training data ranging from 1st January 2018 - 21st December 2019, and the test set ranging from 1st January 2020 - 30th June 2020.

The approach taken in this analysis was to start with the simplest model and add in components in a progressive fashion. This allowed for the exploration of the influence each component had on predicting future air quality as well as being able to assess the effectiveness of including covariate terms in the model. Metrics used for comparing the models where Akaike Information Criteria (AIC) for in-sample performance and mean absolute percentage error (MAPE) and root mean squared error (RMSE) for forecasting performance.

4 RESULTS

4.1 Initial exploratory data analysis

The data was aggregated and averaged on the day of the week, and the month of the year to identify any weekly or monthly trends that may be occurring within the series (Figure 1). There appears to be higher levels of PM2.5 present during the weekend reaching a peak on Saturday. This seems to correspond with the working week and the house would be empty most of the time during this period. This could suggest there would be less cooking and cleaning occurring during these days which are key sources of PM2.5. The monthly trend shows a large difference between months corresponding to colder and warming months. A reason for this could be due to the colder weather causing people to ventilate their houses left (i.e. keeping windows and doors shut) and possibly even having using fuels for heating, such as wood burning stoves that would cause an increase in PM2.5 levels.



(a) **(b)**

Figure 1. (a) Weekly trends in the data. The data was averaged for each day of the week. (b) Monthly trends in the data. The data was averaged for each month

To test for stationarity, an augmented Dickey-Fuller (ADF) test was performed resulting in an ADF test statistic of -2.964, Critical Value(10%) of -2.568, and p-value of 0.0384. Since the p-value is less than 0.05 we can reject the null-hypothesis that the data is not stationary and assume that it is. However, since the value is quite close to the critical value, models will also be fitted to the first order difference of the data to compare the effects that differencing might have. Figure 3 shows that the original series has correlation between the points for at least 15 lags however taking the first order reduces this to around 3 lags. The same was carried out to examine the partial autocorrelation plots for each order of differencing (Figure 4)

Figure 1 suggests there is some seasonal trend depending on the time of year. To further investigate this, a seasonal decomposition was performed to get better understanding of these trends. Figure 2 suggests there is a trend that corresponds with the seasons of the year with higher levels of PM2.5 in the colder months (October - April) and lower levels in the warmer months (June - September. The plots also suggest there is a finer level of seasonality occurring on a weekly or bi-weekly period. Again, this could be a repeat of the trend suggested in Figure 1a where there is a peak of PM2.5 over the weekend days.

The relationship between the weather data and the air pollution data was then explored. The trends from the seasonal decomposition of the average daily air temperature and the average daily wind speeds were plotted against the trends in the PM2.5 series (Figure 5). The plots suggest there is a strong relationship between air temperature and PM2.5 levels and a weaker relationship with wind speed. The Pearson correlation between the air temperature and wind speed with PM2.5 with results of -0.898 and 0.631 further suggesting these features may influence air quality levels.

Table 1. Definition of the models and their respective parameters

	Order of Difference	
Model name	0	1
AR	(2, 0, 0)	(5, 1, 0)
MA	(0, 0, 15)	(0, 1, 3)
ARIMA	(2, 0, 15)	(5, 1, 3)
ARIMAX_temp	N/A	(5, 1, 3)
ARIMAX_temp_wind	N/A	(5, 1, 3)
SARIMA*	N/A	(5, 1, 3) (5,1,3,12)
SARIMAX_temp*	N/A	(5, 1, 3) (5,1,3,12)
SARIMAX_temp_wind**	N/A	(3, 1, 1) (3,1,1,12)

Numbers in brackets denote the values used for the (p, d, q) parameters of the ARIMA and SARIMA models where p is the autoregressive order, d is the difference order, and q is the moving average order.* SARIMA models have additional seasonal parameters defined in the second bracket as (P, D, Q, m) where P is the seasonal autoregressive order, D is the seasonal difference order, Q is the seasonal moving average order and m is the length of a single period.

Table 2. Perfomance metrics for each model

	Order of Difference					
	0			**1**		
Model name	**AIC**	**RMSE**	**MAPE**	**AIC**	**RMSE**	**MAPE**
AR	5030	13.25	1.07	5029	11.55	0.73
MA	5001	13.23	1.07	4973	10.61	0.56
ARIMA	4983	11.26	0.66	4969	10.60	0.56
ARIMAX_temp	-	-	-	4935	10.14	0.56
ARIMAX_temp_wind	-	-	-	4926	9.57	0.48
SARIMA	-	-	-	4481	10.70	0.49
SARIMAX_temp	-	-	-	4463	10.02	0.46
SARIMAX_temp_wind	-	-	-	4645	9.68	0.42

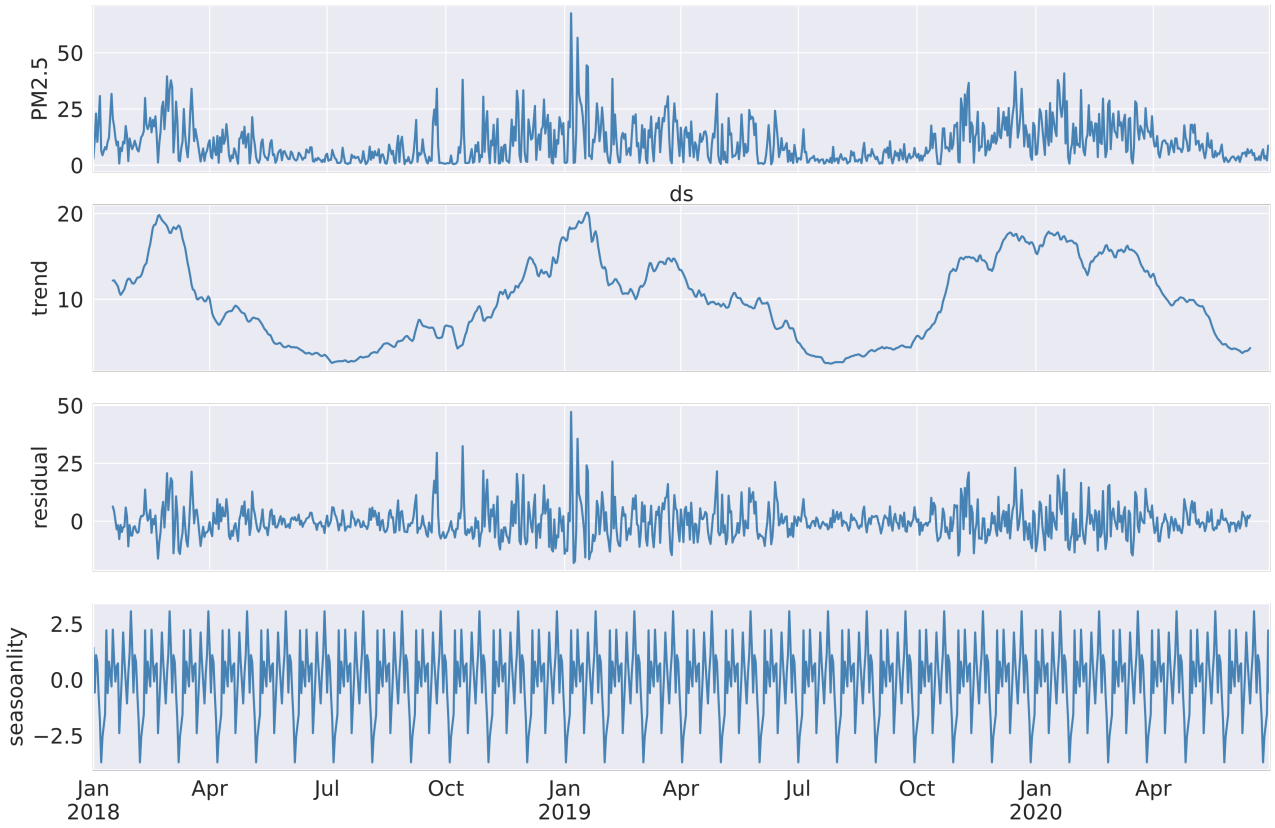


Figure 2. Extracting the trend and seasonality from the data and plotting the residuals

4.2 Model selection

As described in Section 3 a basic AR model was used initially and built upon, adding in various components to increase model performance. An overview of the models and the respective parameters used for each model are given in Table 1. The (p, d, q) and (P, D, Q, m) values were obtained from examining the autocorrelation and partial autocorrelation plots (Figure 3 and Figure 4) to determine what order should be used. Table 2 is a summary of the performance metric of each model where the RMSE and MAPE are measured over a 7 day forecast and compared to the observed values in the testing set. The results indicate that the models do not perform well on this data. It would appear that models using the first order difference of the time series perform better overall and increasing the complexity of the models seems to increase the prediction accuracy. The RMSE performance ranged from 13.25 to 9.57 with ARIMAX_temp_wind performing the best on this metric, however, this is still a relatively high RMSE when compared to the range of the data. Additionally, SARIMAX_temp_wind scored the best in terms of MAPE with a score of 0.42, however, this is still a poor score and means there is an average difference between the predicted value and the actual value of roughly 42%. Looking at the plots of the top three performing models, ARIMAX_temp, ARIMAX_temp_wind, and SARIMAX_temp_wind, (Figure 6, 7, 8) we can see that the prediction also has a large confidence interval range meaning it would not be possible to make a relatively specific prediction on a daily basis. Figure 9 suggests that the ARIMAX model, using temperature as an exogenous value was able to identify the trend and seasonality in the data as it follows the decrease of PM2.5 as it approaches the summer months. The SARIMAX models failed to do this to the same extent.

Vector Autoregression (VAR) is a technique used to calculate predictions based on multiple features that influence each other [14]. VAR was not used in this analysis however it would be interesting to compare the results of the ARIMAX and SARIMAX models to the performance of a VAR model.

5 CONCLUSION

Predicting air quality in the indoor environment can be a challenging task due to the influence of several external factors such as weather and human behaviour. This analysis has used autoregressive time series analysis to attempt to extract the patterns in the data to accurately predict the daily average PM2.5 levels. The results indicate that, for this data, an autoregressive model does

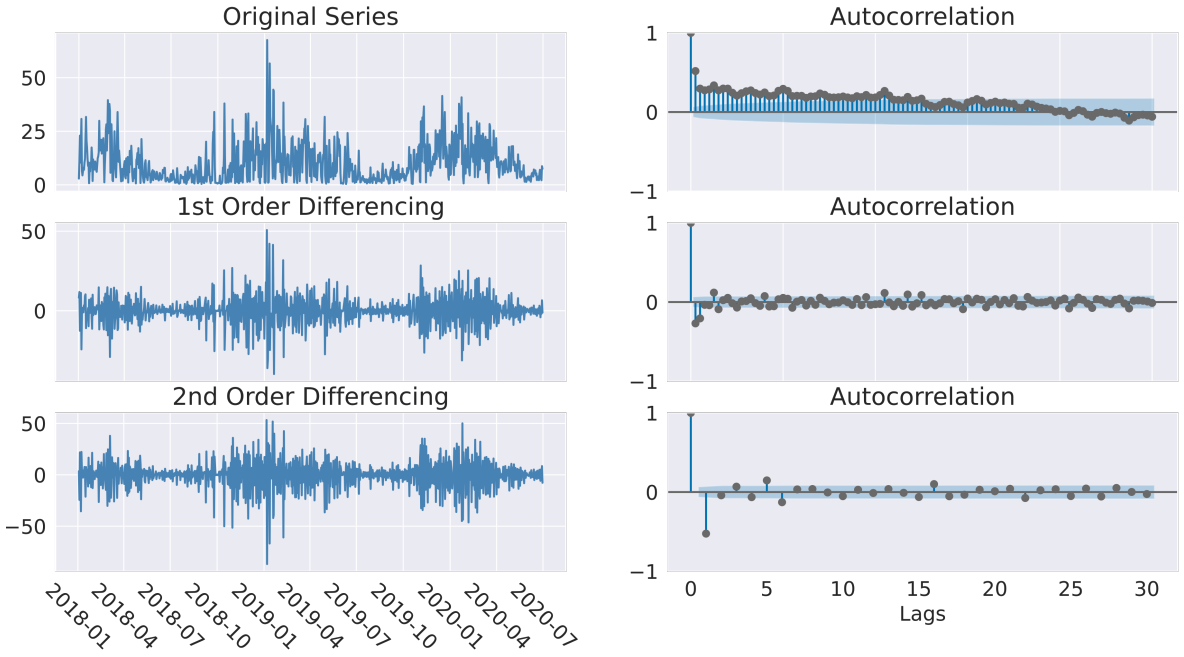


Figure 3. Autocorrelation plots for the original series, 1st order difference, and second order difference. Taking the difference of the series reduces the amount of lags that are correlated to the current step. These were used to determine the order of the moving average term for the models

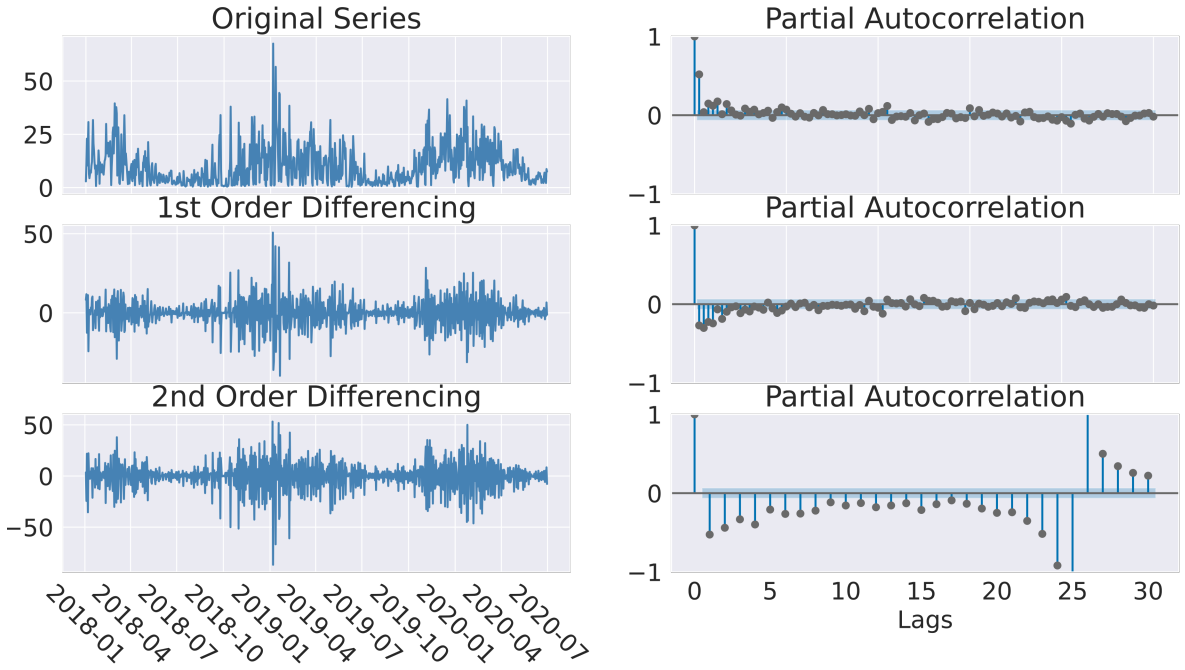
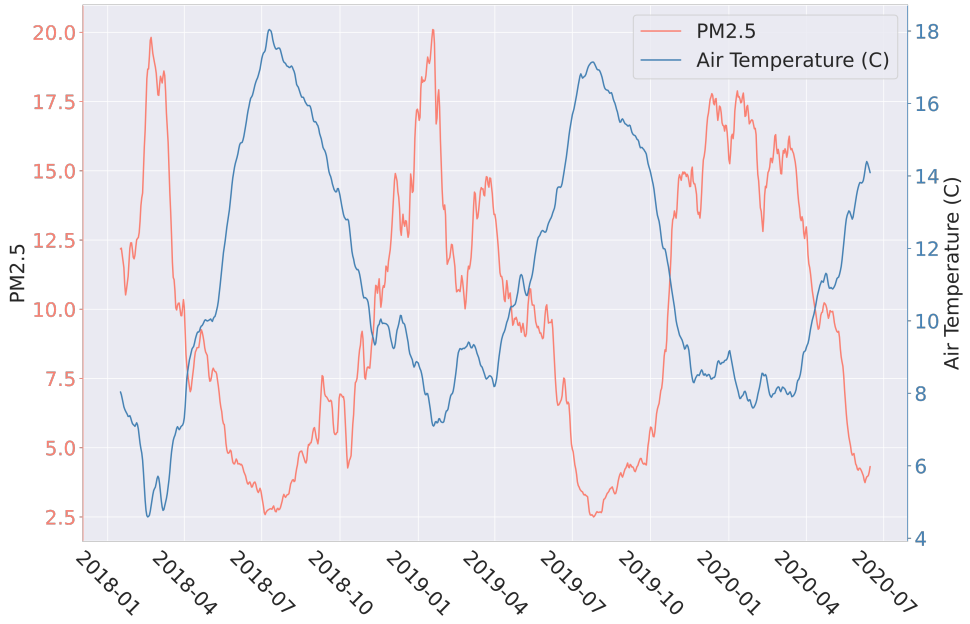
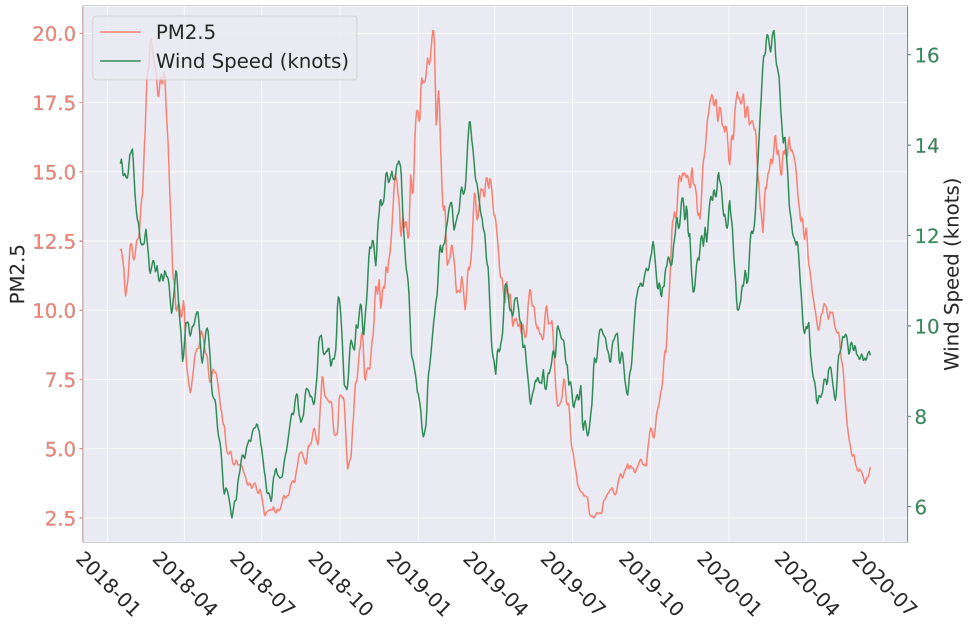


Figure 4. Partial autocorrelation plots for the original series, 1st order difference, and second order difference. These were used to estimate the order of the autoregressive term for the models

not perform significantly well and has on average a large error when predicting future air quality. This being said it shows the benefit of including both the AR and MA components of an ARIMA model and performs better when including weather data as a covariate. Given the high correlation value between both air temperature and wind speed between PM2.5 levels, and the



(a)



(b)

Figure 5. Extracted trend from the PM2.5 series plotted against the extracted trend of the (a) air temperature series and (b) wind speed series

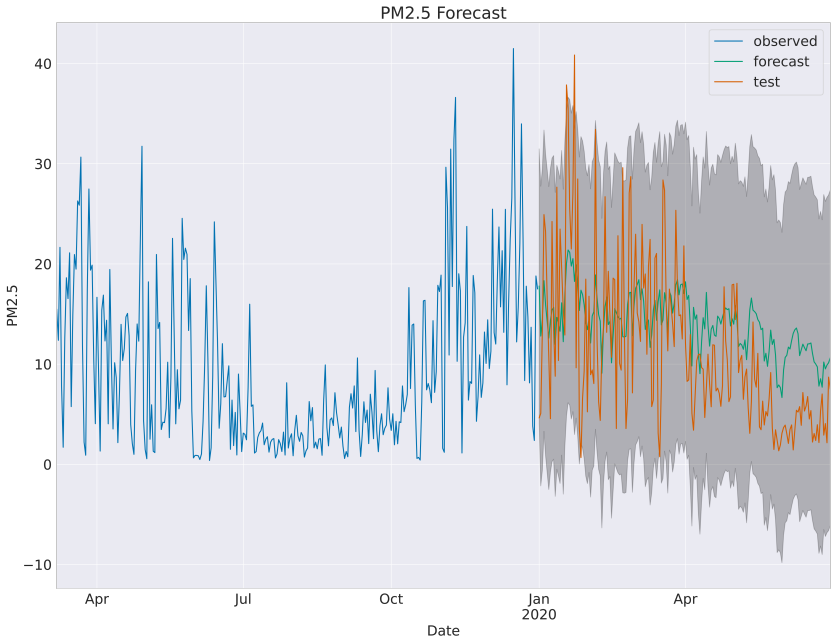


Figure 6. Thirty day prediction results using the ARIMAX_temp_wind model with a p, d, q value of 5, 1, 3 respectively and uses air temperature and wind speed as exogenous values

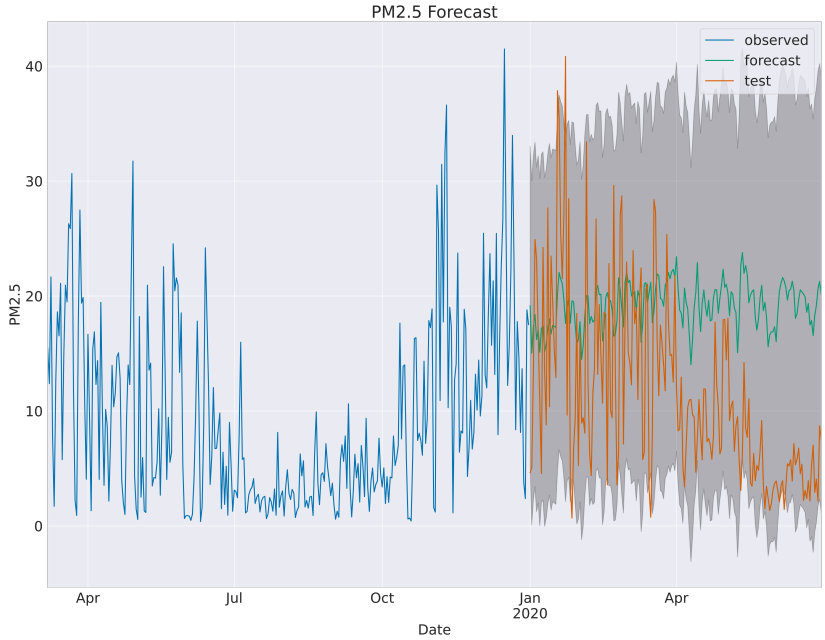


Figure 7. Thirty day prediction results using the SARIMAX model with a p, d, q, P, D, Q, m value of 5, 1, 3, 5, 1, 3, 12 respectively and uses air temperature as an exogenous value

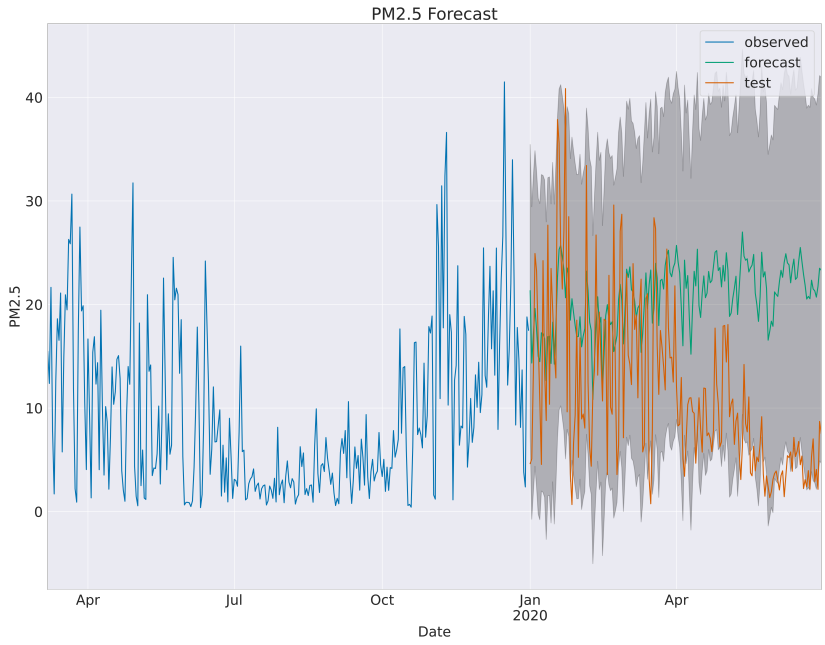


Figure 8. Thirty day prediction results using the SARIMAX_temp_wind model with a p, d, q, P, D, Q, m value of 3, 1, 1, 3, 1, 1, 12 respectively and uses air temperature and wind speed as exogenous values

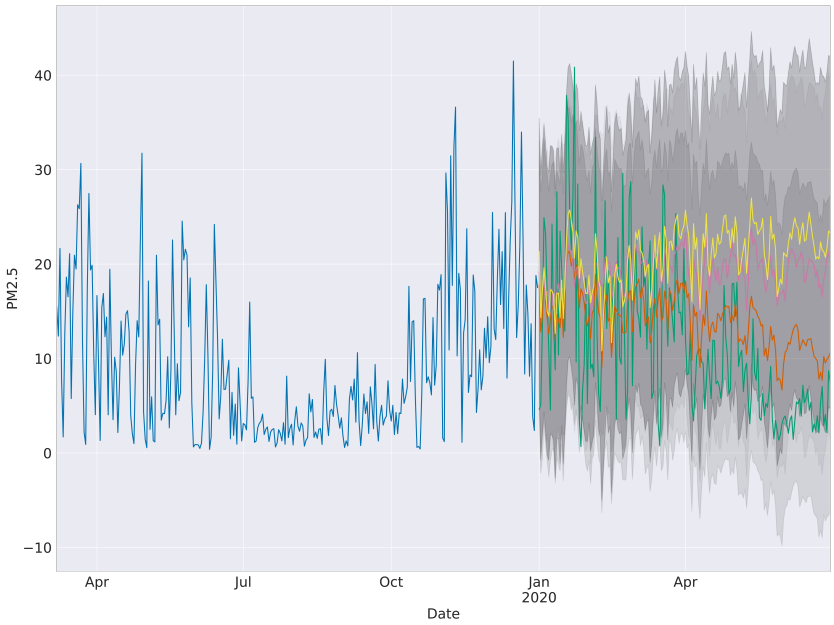


Figure 9. Comparison of the 30 predictions made using $\mathrm{ARIMAX}_{t}emp_{w}ind, SARIMAX, and SARIMAX_{t}emp_{w}ind compared to the test data$

increase in the performance when these features were included in the model, future work looking to predict air quality should look to include forecasted weather as features. Additionally, one area not explored in this research which could help to improve the performance would be the demographics and behaviour of people in the house. For example, the number of people living in the house could affect how much doors and windows get opened, or people who cook more meals at home may increase the levels of PM2.5 in their home. As can be seen in this analysis using a regressive approach to predicting air quality is not sufficient enough and other factors such as these should be included to get an accurate prediction.

Other machine learning approaches may have provide more accurate results and artificial neural networks, decision trees, and random forest models have been used previously in an attempt to predict air quality.

PROJECT CODE AND DATA

All code for the project can be found at GUTHUB

REFERENCES

[1] W. H. Organization, *Ambient air pollution: a global assessment of exposure and burden of disease*. World Health Organization, 2016, 121 p.

[2] J. Gonzalez-Martin, N. J. R. Kraakman, C. Perez, R. Lebrero, and R. Munoz, "A state–of–the-art review on indoor air pollution and strategies for indoor air pollution control," *Chemosphere*, vol. 262, p. 128 376, 2021.

[3] E.-.-E. Commission *et al.*, "Indoor air pollution: New eu research reveals higher risks than previously thought," *Joint Research Center*, 2003.

[4] S. Feng, D. Gao, F. Liao, F. Zhou, and X. Wang, "The health effects of ambient pm2. 5 and potential mechanisms," *Ecotoxicology and environmental safety*, vol. 128, pp. 67–74, 2016.

[5] N. R. Martins and G. C. Da Graca, "Impact of pm2. 5 in indoor urban environments: A review," *Sustainable Cities and Society*, vol. 42, pp. 259–275, 2018.

[6] GOV.UK, *Concentrations of particulate matter (PM10 and PM2.5)*, en. [Online]. Available: `https://www.gov.uk/government/statistics/air-quality-statistics/concentrations-of-particulate-matter-pm10-and-pm25` (visited on 05/15/2022).

[7] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8–16, 2018.

[8] A. Kumar and P. Goyal, "Forecasting of daily air quality index in delhi," *Science of the Total Environment*, vol. 409, no. 24, pp. 5517–5523, 2011.

[9] M. Abhilash, A. Thakur, D. Gupta, and B. Sreevidya, "Time series analysis of air pollution in bengaluru using arima model," in *Ambient Communications and Computer Systems*, Springer, 2018, pp. 413–426.

[10] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[11] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[12] G. Grell and A. Baklanov, "Integrated modeling for forecasting weather and air quality: A call for fully coupled approaches," *Atmospheric Environment*, vol. 45, no. 38, pp. 6845–6851, 2011.

[13] I. Jhun, B. A. Coull, J. Schwartz, B. Hubbell, and P. Koutrakis, "The impact of weather changes on air quality and health in the united states in 1994–2012," *Environmental research letters*, vol. 10, no. 8, p. 084 009, 2015.

[14] J. K. Sethi and M. Mittal, "Analysis of air quality using univariate and multivariate time series models," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, 2020, pp. 823–827.