THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Time Series Modeling & Analysis

DATS 6450 – Section 15

Fall 2020

Instructor: Dr. Reza Jafari

Term Project Report

NO2 Concentration Prediction

Tran Hieu Le

Columbian College of Arts and Sciences

TL December 9, 2020

Abstract

In recent years, the air quality has been getting worse due to the increasing emission of air pollutants from human activities and environmental changes. This project presents time series analysis and modeling to predict the ground-level concentration of nitrogen dioxide (NO2) – one of 6 criteria air pollutants identified by United States Environmental Protection Agency (EPA). The purpose is to provide useful information and forecast the future concentration of the air pollutant, which would help governments and authorities to make appropriate plans and solutions to protect the air quality.

Introduction

The global industrialization and the world population explosion since the mid of 20th century have led to the increasing levels of pollutants in the atmosphere. The air pollution not only has detrimental effects on human health but also contributes to a variety of negative environmental changes. In 2016, the World Health Organization (WHO) reported that an estimation of 4.2 million deaths annually is attributed to outdoor air pollution. Moreover, the increase of air pollutants is a major cause of severe environmental problems such as acid rain, eutrophication, haze, ozone depletion, wildlife destruction and forest fires.

Due to these highly concerned impacts, it is important to analyze and study the progress of air quality. Since there is a wide range of factors to the concentration of air pollutants and their relationships with the gases are not explicit, it is difficult for explanatory models to make accurate forecasts. A time series analysis is a good solution to this situation as the concentration of air pollutants is time dependent. For example, during daytime industrial activities and traffics would lead to a high level of gases in the atmosphere, whereas there is a lower concentration of pollutants at night. The variety in temperature, humidity and pressure due to the change in season also effects the chemical reaction that produces the gases.

NO2 is a popular air pollutant as it is a daily production of burning fuel from road traffics and power plants. The gas contributes to the development of breathing issues such as asthma and respiratory infections, and is a major cause to acid rain, hazy air and polluted coastal waters. Since its negative impacts on human health and environment, NO2 is selected as the target air pollutant for time series analysis and modeling.

This project consists of 6 main parts: data preprocessing; data exploration; baseline models using simple forecasting methods; Holt's Linear and Holt-Winter; multiple linear regression (MLR) model; and ARIMA models (ARMA, non-seasonal ARIMA and SARIMA).

A final evaluation of the proposed time series forecasting methods is presented at the end of the project.

Dataset

The data comes from the UC Irvine Machine Learning repository:

https://archive.ics.uci.edu/ml/datasets/Air+Quality

The dataset contains 9358 observations and 15 variables. There are 5 variables representing the hourly averaged concentrations of 5 different gases (Carbon monoxide, non-Methanic Hydrocarbons, Benzene, total Nitrogen oxides and Nitrogen dioxide), and 5 variables indicating the responses from an array of 5 metal oxide chemical sensors (tin oxide, tungsten oxide NOx targeted, tungsten oxide NO2 targeted and titania). The meteorological variables are temperature, relative humidity and absolute humidity. The datetime is split into two components: day-month-year and hour. The target is the concentration of NO2, and the remaining 14 variables can be considered as predictors. Missing data is recorded by -200.

Methodology and Procedures

Data preprocessing

This is the first step before working on the analysis and modeling. In this part, the two datetime components are combined into a variable. Since the majority of the concentrations of Benzene (90% of the observations) are missing, this variable is removed from the dataset. For the remaining variables, the missing values are imputed using the average of available data in the same day. If there is no record in a day, the missing data is imputed as the average of two closest available data (using the combination of forward fill and backward fill) with the assumption that the current data is related to the previous and the next observations.

Data exploration

The data is split into a training set (80%) for the time series analysis and modeling, and a testing set (20%) for evaluation.

A graph of the first 1000 observations in the training set is provided to visualize the time series. The autocorrelation function (ACF) plot of the time series is used to visualize autocorrelations at different lags, and to identify trend and seasonality.

Augmented Dickey-Fuller or ADF tests for the training data, its rolling mean and its rolling variance perform hypothesis tests for the presence of unit roots. If the p-value of the test is smaller than the significance level, the time series has a unit root. In contrast, if the p-value is bigger than the significance level, the null hypothesis that there is a unit root in the time series is rejected. The training data is stationary if we fail to reject the null in all 3 ADF tests.

Seasonal and Trend decomposition using Loess or STL decomposition is used to split the time series into three components, which helps to illustrate underlying patterns separately and measure the strength of trend and seasonality.

The observation in ACF plot, the statistics from ADF tests and the results from STL decomposition are used as evidences for the stationarity of the time series.

Baseline models

Simple forecasting methods such as average, naïve, drift and exponential smoothing methods such as simple exponential smoothing are used as baselines to evaluate advanced forecasting methods.

Advanced exponential smoothing methods

Using advanced forecasting methods such as Holt's Linear and seasonal Holt-Winter models to forecast.

Multiple Linear Regression

Assuming the relationship between NO2 and the other variables, MLR is used to forecast the concentration of NO2 considering the remaining variables as predictors. Backward stepwise regression and forward stepwise regression are performed for feature selection to derive the most optimal and simplified model.

In empirical application, MLR uses the time series forecasting of predictors as the inputs to calculate the target variable. In this dataset, the sensor responses and the concentrations of gases are greatly affected by a variety of unpredicted factors such as visitors, buildings, vehicles, the systematic errors of the sensors and the surrounding condition. This leads to the difficulty in forecasting these variables correctly. In fact, all of gases' concentrations and the sensors' responses are the measurements from the sensors, so any of them can be the target variable as NO2 concentration. Their time series forecasting is expected to be challenging, which reduces performance of MLR model.

Therefore, it is not advisable to use these variables as the predictors. Meteorological variables are much better as the features since they have strong and clear relationships with time, and their time series forecasting is more accurate. A more realistic MLR model using temperature, relative humidity and absolute humidity as predictors are built to make comparisons with other models.

ARMA model

Transformations such as logarithm transformation and differencing are used to make the training data stationary if necessary. ADF tests and ACF/PACF plots are used to check the stationarity of the transformed time series. GPAC table is used to find appropriate orders of autoregressive (AR) and moving average (MA) processes. The corresponding parameters of the ARMA model are estimated using Levenberg-Marquardt algorithm. Confidence interval and zero/pole cancellation are used to simplify the ARMA model.

ARIMA model

STL decomposition is used to remove the seasonal component from the time series. Additional transformations are applied to the make the seasonal adjusted data stationary. The number of 1^{st} differencing transformations is the degree of 1^{st} differencing involved in the ARIMA model. GPAC table is used to find the orders of AR and MA processes for the non-seasonal ARIMA model.

SARIMA model

Seasonal differencing followed by one or more 1st differencing is used to make the time series stationary. The lag of seasonal differencing is the seasonal period of the SARIMA model. Non-seasonal terms are determined using GPAC table, and the seasonal terms are determined using ACF and PACF plots.

Results and Explanations

Data exploration

Figure 1. Hourly averaged NO2 concentration of 300 samples

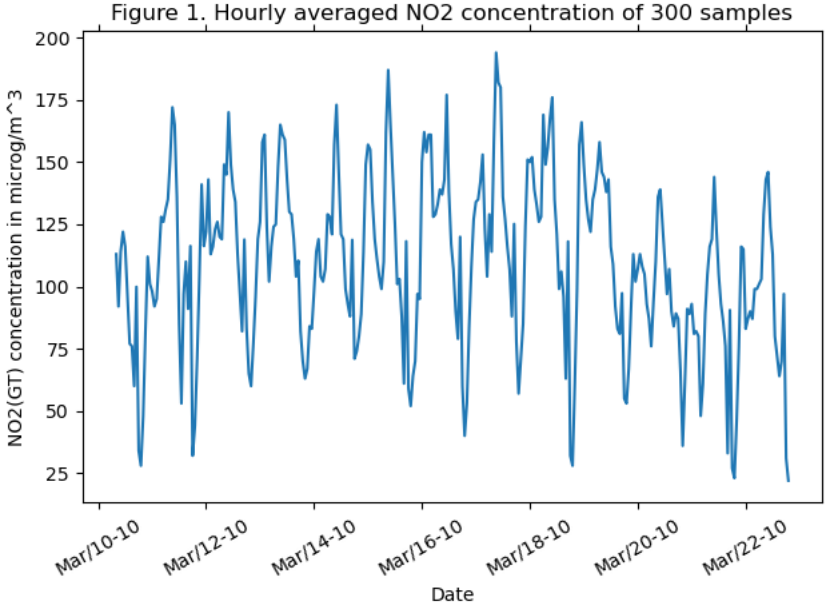
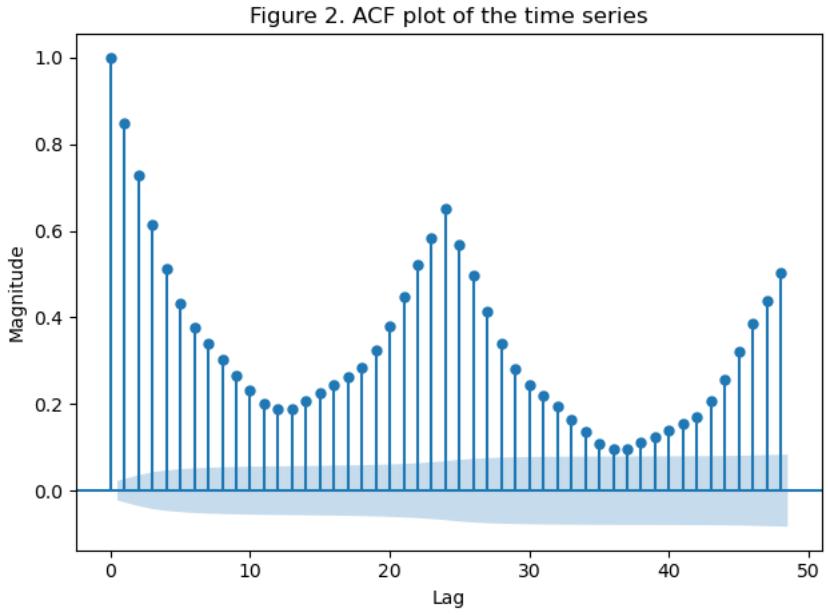


Figure 2. ACF plot of the time series



According to Figure 1, the time seems to have a seasonal behavior, but it is difficult to observe a trend from 300 samples. The ACF plot in Figure 2 shows both trend and seasonality in the time series. The autocorrelations at small lags are large and positive and slowly decays as the lags increase, which means that the data have a trend. For each 24 lags, there is a spike in the ACF plot, indicating a seasonality with frequency of 24. Therefore, the time series tend to be non-stationary.

Since the season period is 24, the number of lags for Box-Pierce test will be 48.

Table 1. ADF test with significance level 0.05.

Target of ADF test	ADF statistics	p-value
Training data	-7.4890	0.0000
Rolling average	-3.8139	0.0028
Rolling variance	0.8294	0.9921

According to Table 1, the rolling variance have lower p-values than the significance level 0.05, so we reject the null that there is a unit root in the rolling variance. In other words, the training data is non-stationary since its variance is not constant over time.

Figure 3. STL decomposition (showing 1000 samples)

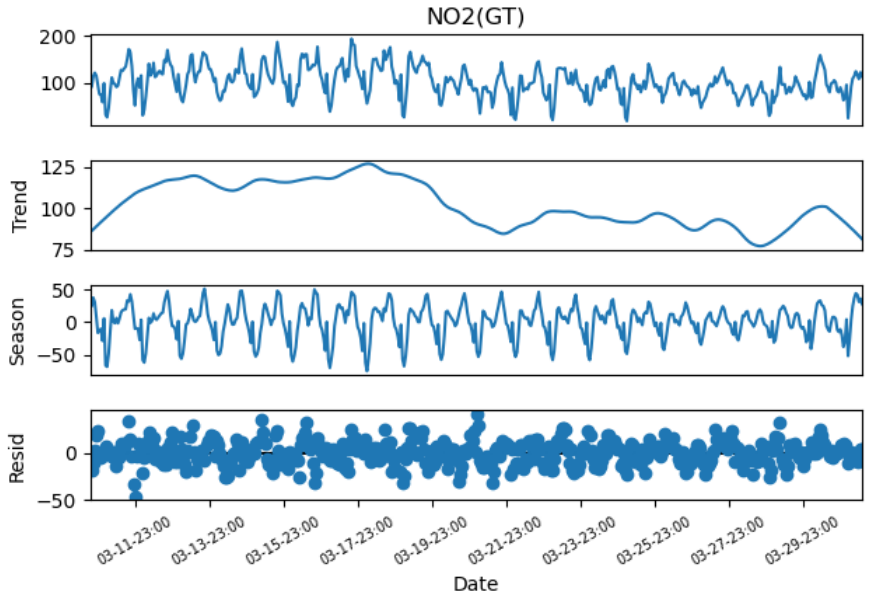
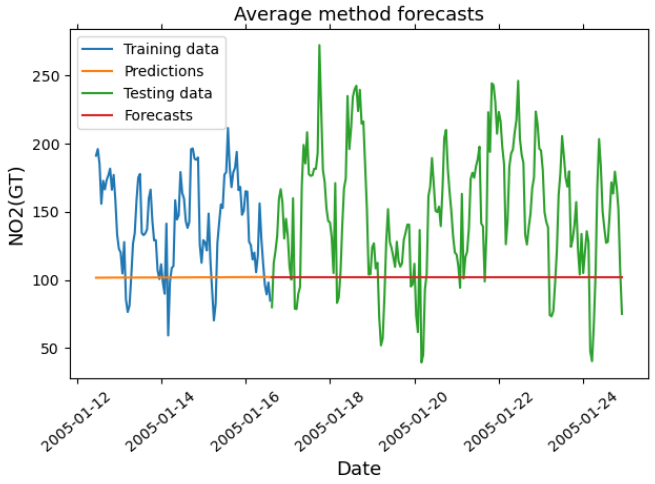


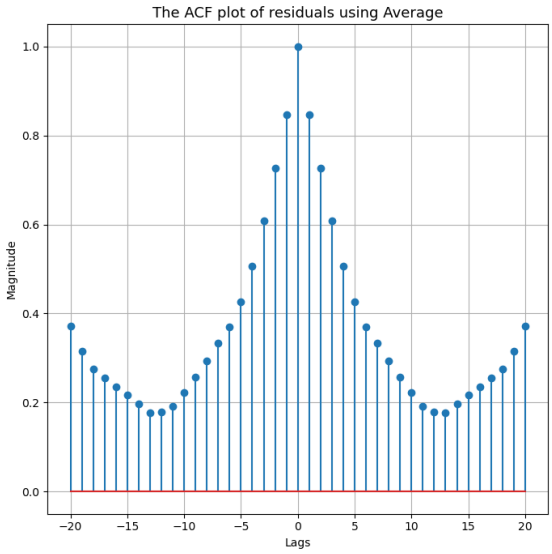
Figure 3 shows a direct STL decomposition for the training set without any transformation. As can be seen from the figure, the seasonal component tends to have a constant magnitude of fluctuations over time, indicating that the time series is additive. There is no need to apply logarithm transformation to the training data and reproduce the decomposition. The trend component does not show any clear pattern. This means that the forecasts from time series models are expected to have neither increasing nor decreasing trends.

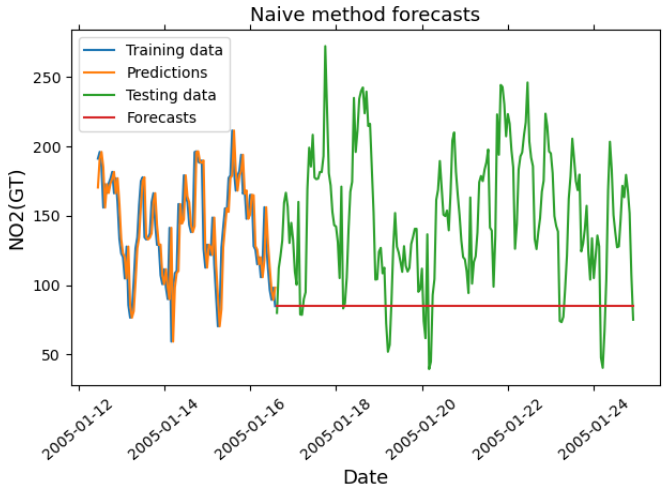
The strength of trend-cycle and seasonal components are 0.5690 and 0.8164, respectively. These measurements show that the trend-cycle is presented but not strong, whereas there is a strong seasonal behavior in the time series. The seasonal component must be removed to be able to fit the data into non-seasonal models.

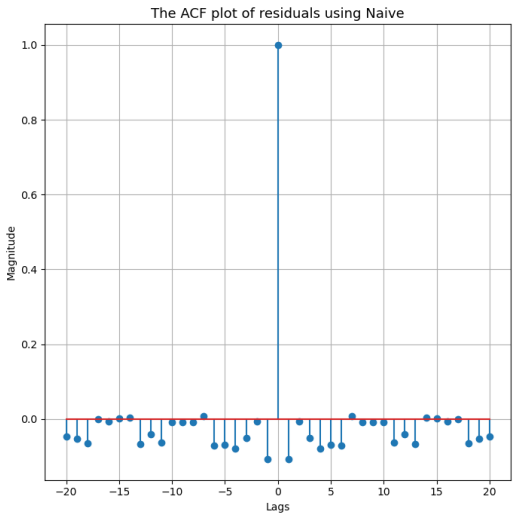
In summary, the observations from ACF plot, the results from ADF tests and the measurements from STL decomposition agree that the training data is non-stationary. Transformations such as seasonal differencing and 1st differencing need to be applied to make the time series stationary before finding appropriate seasonal and non-seasonal orders of AR and MA processes in ARMA, ARIMA and ARIMA models.

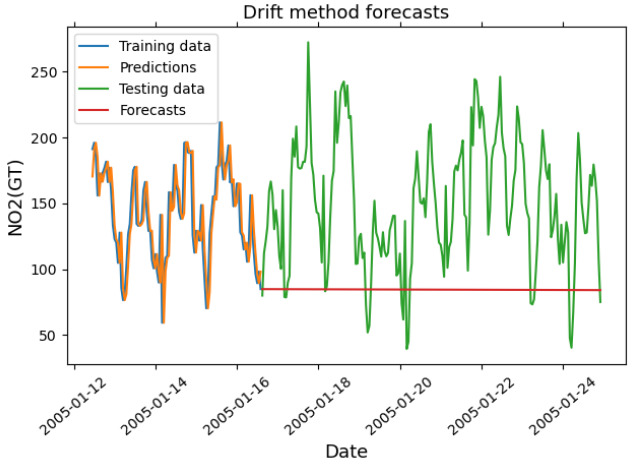
Baseline models

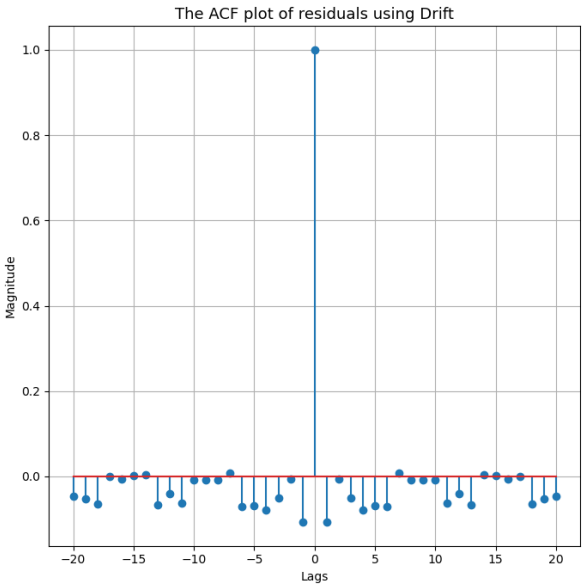


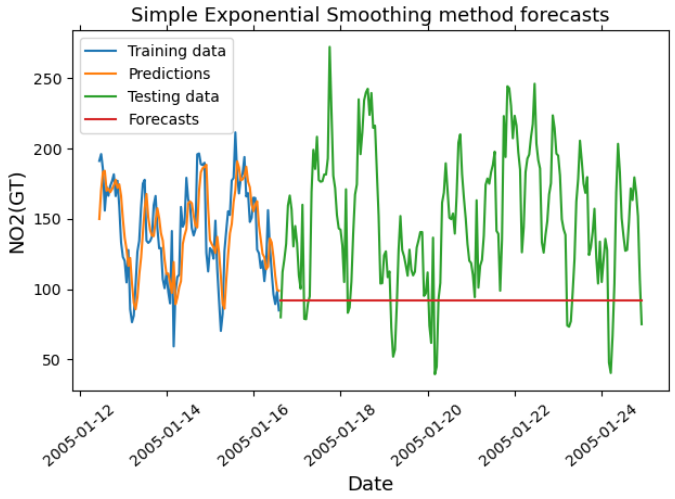












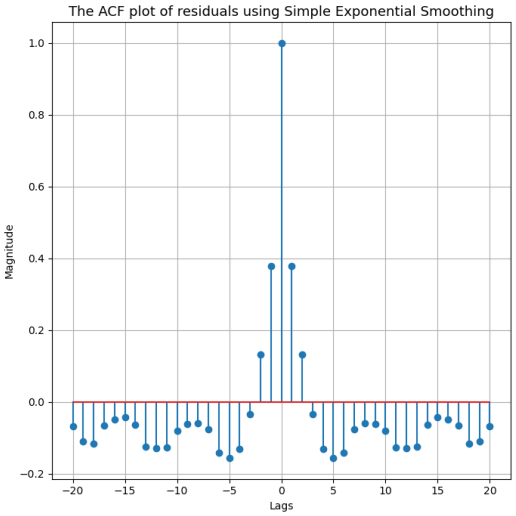


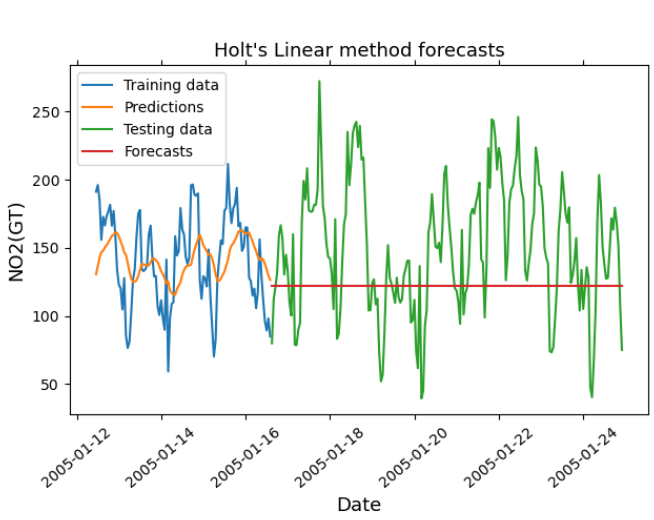
Table 2. Statistics of simple forecasting methods

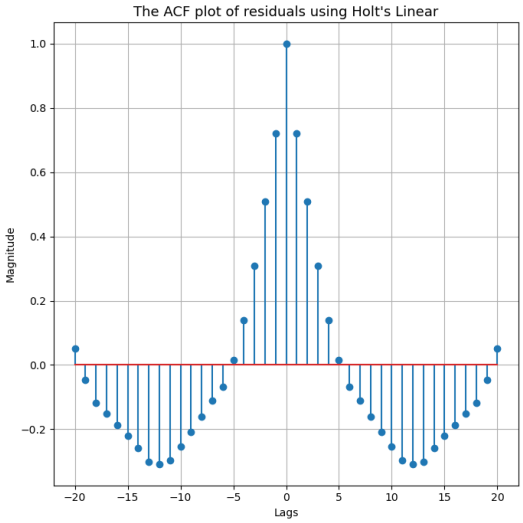
Model	Training			Testing		Q value of residuals
	Mean of residuals	MSE	Variance of errors	MSE	Variance of errors	
Average	4.1502	1538.6385	1521.4342	4571.4732	2820.4764	47251.5828
Naive	-0.0039	465.5377	465.5377	6315.1318	2820.4764	3940.4228
Drift	0.0083	466.2587	466.2585	6712.9523	2790.2018	3938.5814
SES ($\alpha = 0.5$)	-0.0044	533.9946	533.9946	5539.5756	2820.4764	804.1091

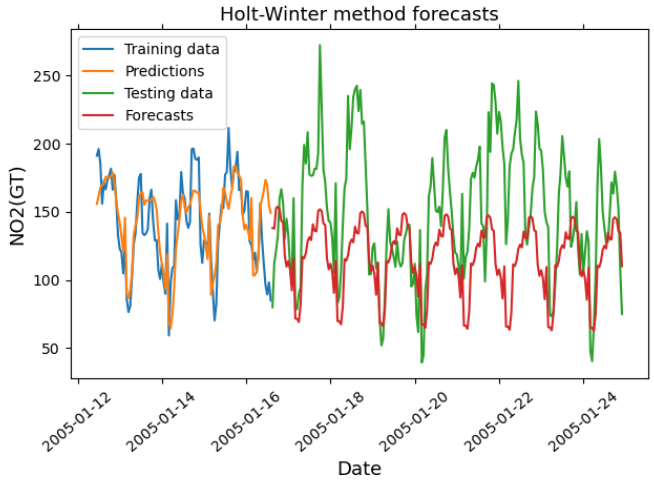
According to Table 2, all simple forecasting methods have high Q-value, meaning that the autocorrelation functions of residuals are significant from 0. The ACF plots for residuals also show the strong autocorrelations among residuals, meaning that the residuals from these baselines are not white noise. Average method seems to be the best baseline model since it has the lowest mean squared errors in the testing set. The difference between the predictions and forecasts of average method is also the lowest as the method gives the highest ratio of variances of residuals and forecasted errors. However, the mean of predicted errors is different from 0, meaning that the average method is a biased estimator.

In summary, the simple forecasting methods are not good models for real application but baselines to evaluate advanced time series forecasting methods such as Holt-Winter, MLR and ARIMA.

Holt's Linear and Holt-Winter







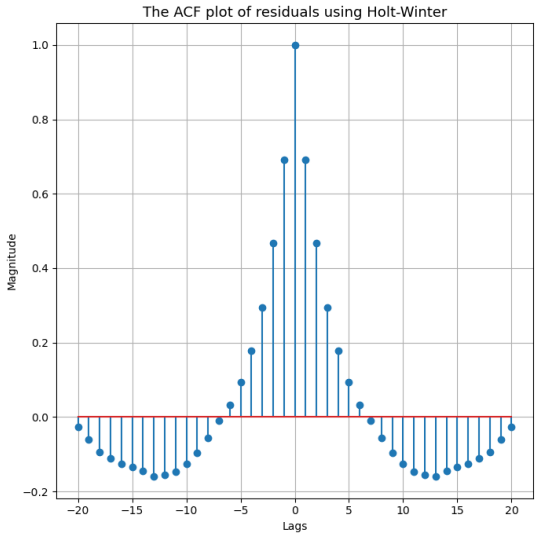


Table 3. Statistics of Holt's Linear and Holt-Winter methods						
Model		Training		Testing		Q value of residuals
	Mean of residuals	MSE	Variance of errors	MSE	Variance of errors	
Holt's Linear	0.03110	905.9211	905.9199	3289.4052	2820.4764	25329.7241
Holt-Winter	-0.0149	521.6586	521.6584	3129.2914	2038.5278	8513.8148

Both Holt's Linear and Holt-Winter give better results than simple forecasting methods. Holt-Winter model is better than Holt's Linear given its lower MSE and lower Q value from Box-Pierce test. The mean of residuals is approximately 0, meaning that Holt-Winter is an unbiased estimator. However, the autocorrelation functions of residuals are much higher than 0, and there are clear patterns of seasonality in ACF plots, making the residuals not white noise.

Multiple Linear Regression

1. Multiple Linear Regression using all available features

Table 4. Feature selection and corresponding AIC, BIC and Adjusted R squared value				
Method	Removed features	AIC	BIC	Adjusted R^2
Regression with all features	None	11244.7874	11327.8353	0.7374
Backward stepwise regression	PT08.S1 (tin oxide)	11244.5350	11320.6623	0.7374
Forward stepwise regression	PT08.S3 (tungsten oxide)	11248.0254	11324.1526	0.7373

According to Table 4, backward stepwise regression has the lowest AIC and BIC, whereas the adjusted R squared value is similar to this value in the regression using all features. In other words, backward stepwise regression gives the most optimal regression, and the feature PT08.S1 can be removed from the model.

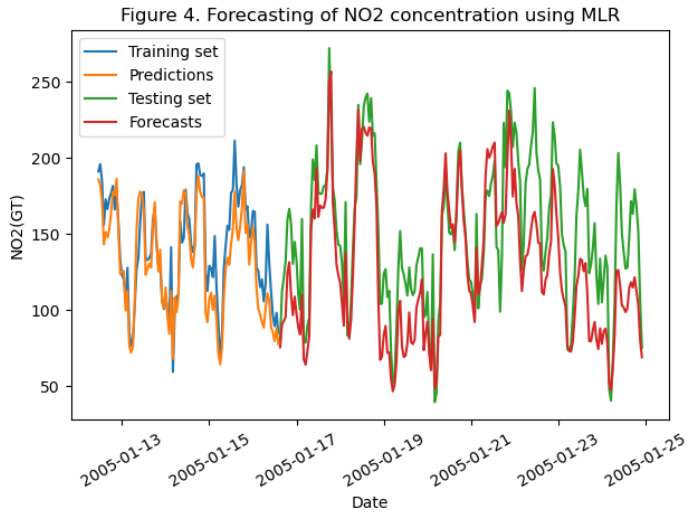
The summary of the model

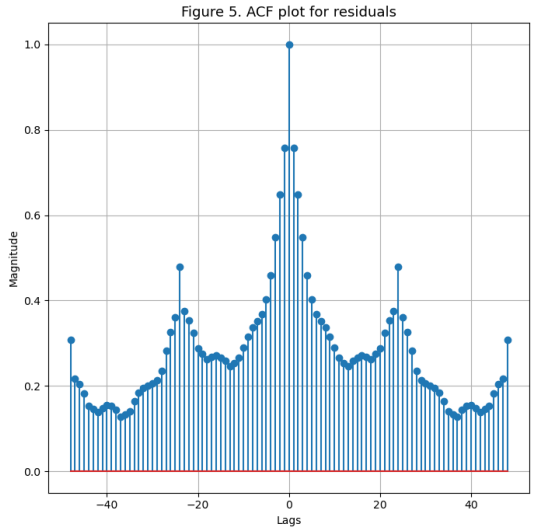
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 NO2(GT)   R-squared:                       0.738
Model:                            OLS   Adj. R-squared:                  0.737
Method:                 Least Squares   F-statistic:                     2103.
Date:                Sun, 13 Dec 2020   Prob (F-statistic):               0.00
Time:                        23:15:13   Log-Likelihood:                -5611.3
No. Observations:                7485   AIC:                         1.124e+04
Df Residuals:                    7474   BIC:                         1.132e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.267e-16      0.006   -7.2e-14      1.000      -0.012       0.012
AH               -0.3212      0.018    -17.901      0.000      -0.356      -0.286
C6H6(GT)         -0.7918      0.040    -19.786      0.000      -0.870      -0.713
CO(GT)            0.1695      0.013     13.133      0.000       0.144       0.195
NOx(GT)           0.5609      0.014     40.678      0.000       0.534       0.588
PT08.S2(NMHC)     0.5275      0.048     10.882      0.000       0.433       0.623
PT08.S3(NOx)     -0.0329      0.014     -2.301      0.021      -0.061      -0.005
PT08.S4(NO2)      0.1928      0.019      9.888      0.000       0.155       0.231
PT08.S5(O3)       0.3004      0.016     18.368      0.000       0.268       0.332
RH               -0.1516      0.019     -7.815      0.000      -0.190      -0.114
T                 0.1047      0.023      4.495      0.000       0.059       0.150
==============================================================================
Omnibus:                      317.436   Durbin-Watson:                   0.487
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              458.450
Skew:                           0.407   Prob(JB):                    2.81e-100
Kurtosis:                       3.899   Cond. No.                         24.0
==============================================================================
```

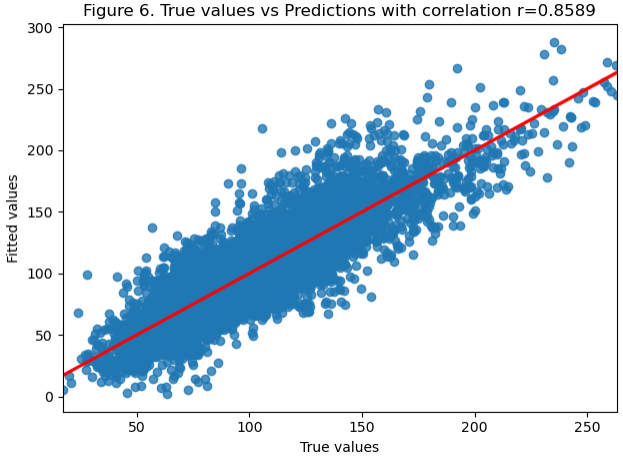
As can be seen from the summary of the model, every feature in the model is statistically significance from 0 since the p-value of t-test is smaller than the significance level 0.05. The intercept is in fact not significant from 0 and can be removed from the regression. The adjusted R squared value is around 0.7374, meaning that the model fits the historical data very well as it explains 73.74% of the variation in the NO2 concentrations.

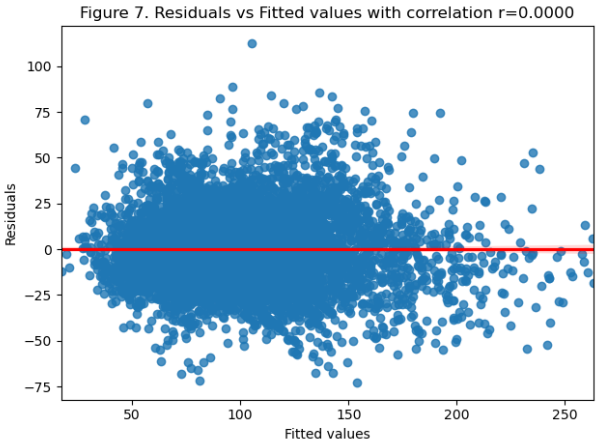
Table 5. Statistics of the most optimal multiple linear regression

Training set				Testing set	
Mean of residuals	MSE	Variance of errors	Q-value	MSE	Variance of errors
2.1871e-15	403.1102	403.1102	34084.5103	1632.5148	725.3511









According to Table 5, the mean of residuals is approximately 0, indicating that the regression is an unbiased estimator. The mean squared of residuals and forecast errors are both better than the baseline models, which means that the model is worth to be considered. The correlation coefficient between the predicted values and the true values in the training set is about 0.8589 (approximately the same as the square root of the Adjusted R squared value), which means that prediction is highly correlated to the real data.

Moreover, there is no linear relationship between the residuals and the fitted values, indicating that the variance of the residuals is constant. However, the autocorrelations among residuals are much different from zero as the Q-value of Box-Pierce test is very high at 34084.5103. The autocorrelation function plot in Figure 5 also shows that the residuals are highly correlated and not white noise.

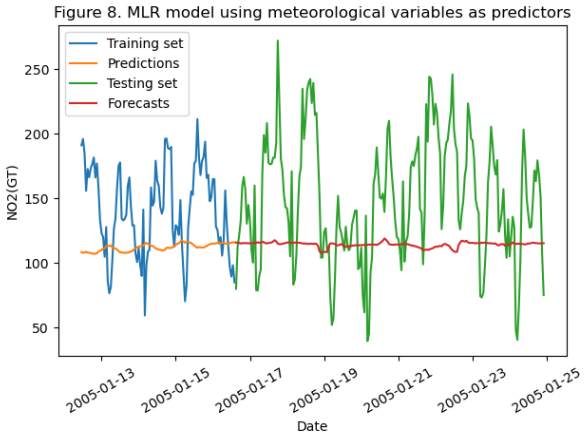
2. Empirical MLR model using meteorological features

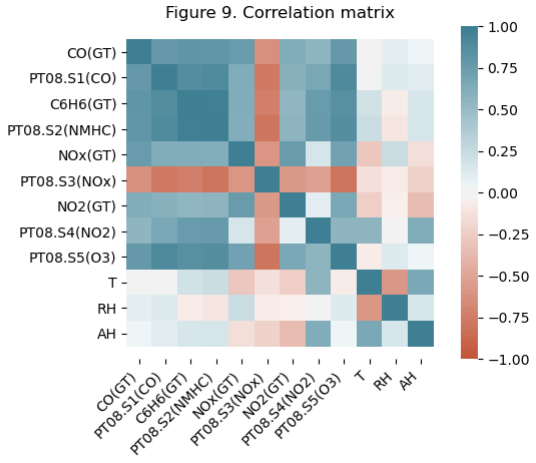
As was discussed in the "Methodology and Procedures" section, the time series forecasting of any variable (including NO2 concentration - the target variable of this project) recorded from the sensors is challenging and would be inaccurate. Therefore, the MLR model should only use meteorological features as predictors since they are highly time dependent and easy to be forecasted using time series methods.

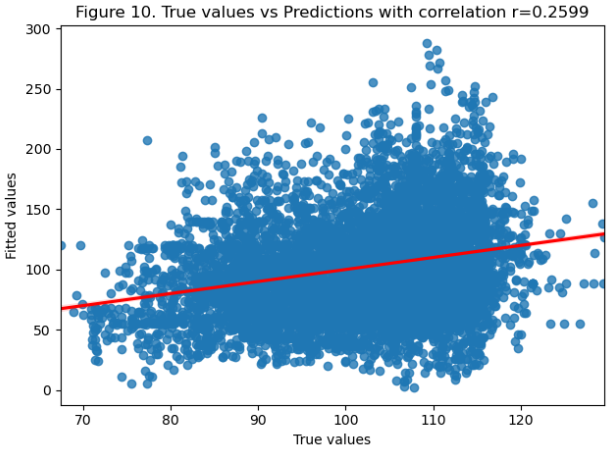
According to the feature selection from previous part, the meteorological variables: temperature, relative humidity and absolute humidity are important and help to improve the model performance, so the MLR model can include all of them.

The model summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 NO2(GT)   R-squared:                       0.068
Model:                             OLS   Adj. R-squared:                  0.067
Method:                  Least Squares   F-statistic:                     180.6
Date:                 Sun, 13 Dec 2020   Prob (F-statistic):          4.51e-113
Time:                         23:43:42   Log-Likelihood:                -10359.
No. Observations:                 7485   AIC:                         2.073e+04
Df Residuals:                     7481   BIC:                         2.075e+04
Df Model:                            3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.267e-16      0.011  -3.82e-14      1.000      -0.022       0.022
T              0.2220      0.042      5.284      0.000       0.140       0.304
RH             0.1628      0.035      4.597      0.000       0.093       0.232
AH            -0.4004      0.031    -13.108      0.000      -0.460      -0.341
==============================================================================
Omnibus:                      386.938   Durbin-Watson:                   0.330
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              486.161
Skew:                           0.523   Prob(JB):                    2.70e-106
Kurtosis:                       3.682   Cond. No.                         7.42
==============================================================================
```







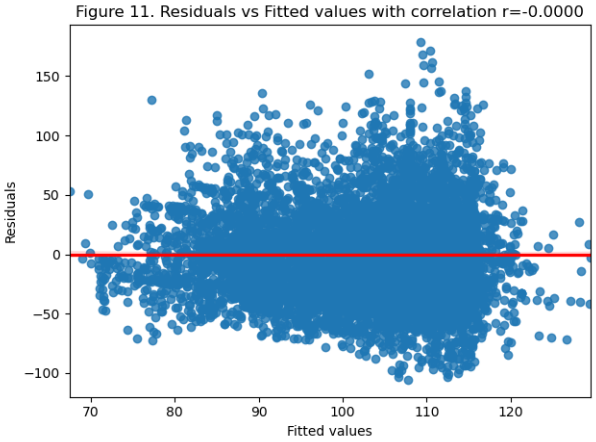


Table 6. Statistics of a realistic multiple linear regression using meteorological features						
Training set					Testing set	
Mean of residuals	MSE	Variance of errors	Q-value	Adjusted R^2	MSE	Variance of error
1.1664e-14	1433.4221	1433.42208	36938.2032	0.0672	3732.4824	2762.2650

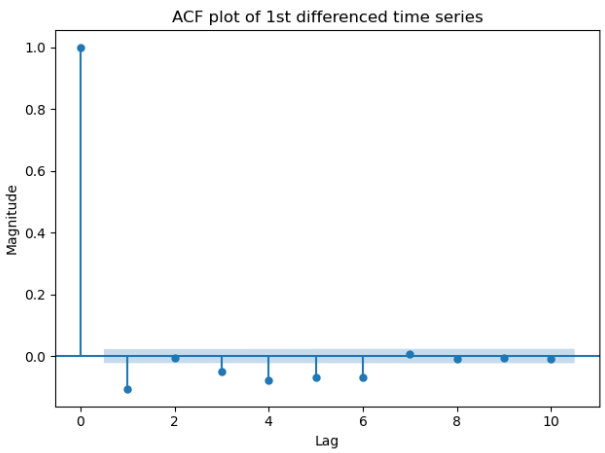
The MLR model is unbiased since the mean of residuals is approximately zero. However, the model does not fit the training data well since it can only explain 6.72% of the variance in the observations. As can be seen from Figure 8, the model is not able to capture the seasonal pattern of the dataset, which is a major cause of the bad performance. Moreover, the correlation matrix in Figure 9 shows that the relationships of temperature, relative humidity and absolute humidity with NO2 concentration are very weak, leading to the ineffectiveness of explanatory methods using these meteorological features. To minimize this issue and improve the predictive power of MLR model, more meteorological features such as wind speed, wind direction, pressure, precipitation, cloud cover … should be recorded in the dataset to provide more information for regression.

ARIMA models

1. ARMA (na, nb)

Since the original training set is not stationary, it is necessary to apply transformations to make the data stationary.

1st differencing transformation:



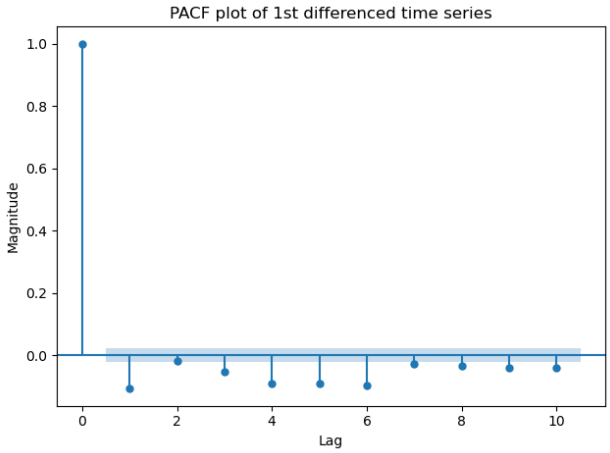
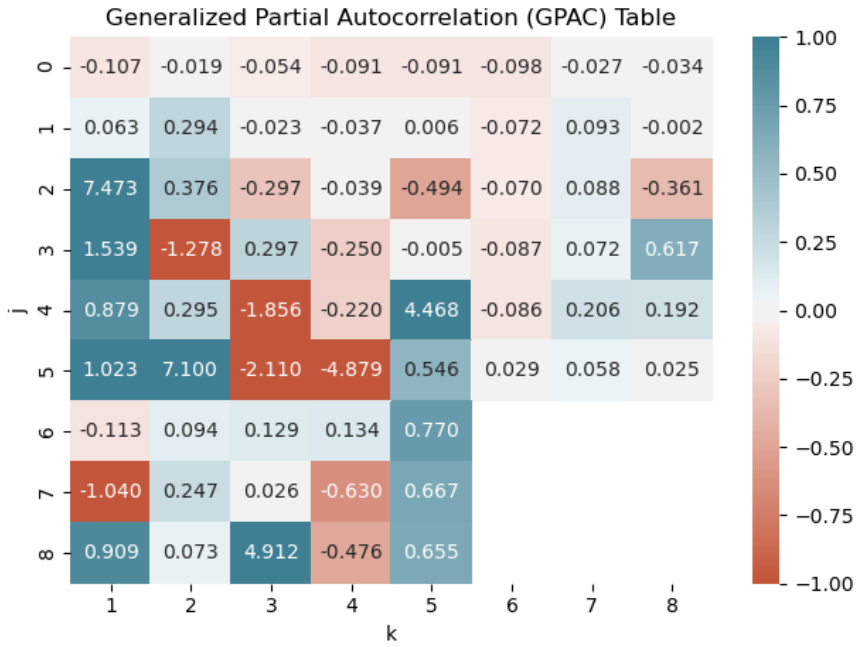


Table 7. ADF tests for the transformed time series with significance level 0.05.

Target of ADF test	ADF statistics	p-value
1st differenced data	-7.4890	0.0000
Rolling average	-9.0617	0.0000
Rolling variance	-7.3220	0.0000

According to ACF and PACF plots, the autocorrelation functions quickly decay to approximately 0 after few lags, indicating that there is no trend and seasonality after 1st differencing transformation. Moreover, the ADF tests for the differenced data results in p-values which are smaller than the significance level 0.05. This mean that we reject the null hypothesis of the presence of a unit root in the time series. In other words, the data after 1st differencing is stationary.

In the next steps, GPAC table is used to find appropriate orders of AR and MA processes. Levenberg-Marquardt algorithm is the method to estimate the corresponding parameters. After that, confidence intervals and zero/pole cancellation are applied to simplify the ARMA model if necessary.



According to GPAC table, there are two possible cases for ARMA model. The first one is ARMA (5,5) as we can see a row of 0s at j = 5, and a column with some similar values (0.546, 0.770, 0.667, 0.655) at k = 5. The second possible model is ARMA (2,1) with the similar observation.

ARMA (5,5):

Process	Parameter	LM estimated value	Confidence interval (95%)	Roots
AR	a1	-0.4982	[-0.5132, -0.4832]	-0.8636
	a2	-0.7089	[-0.7138, -0.7040]	-0.8635
	a3	0.6843	[0.6774, 0.6911]	0.7048
	a4	0.7404	[0.7355, 0.7453]	0.7048
	a5	-0.8156	[-0.8306, -0.8006]	0.9894
MA	b1	-0.6718	[-0.6775, -0.6661]	-0.8659
	b2	-0.7617	[-0.7667, -0.7567]	-0.8659
	b3	0.7549	[0.7492, 0.7606]	0.7070
	b4	0.6782	[0.6732, 0.6832]	0.7070
	b5	-0.9766	[-0.9822, -0.9711]	0.8160

According to the above table, all the parameters are significant from 0 since their confidence intervals do not include 0. However, since the roots of a1 and a2 are roughly similar to the roots of b1 and b2, and the roots of a3 and a4 are roughly similar to the roots of b3 and b4, the model can be simplified to ARMA (1,1) due to zero/pole cancellations.

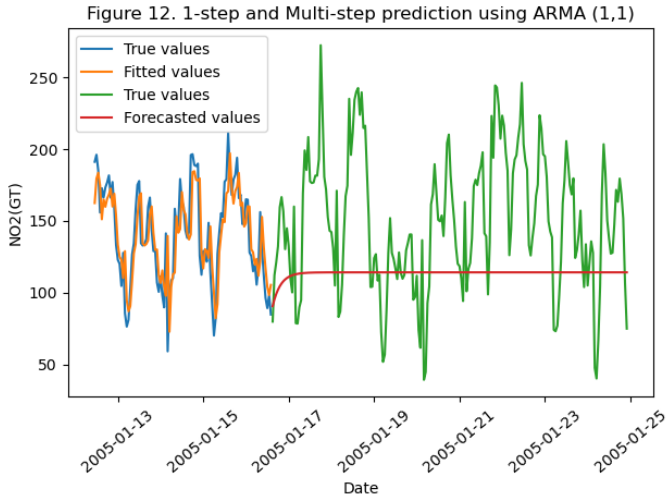
ARMA (2,1):

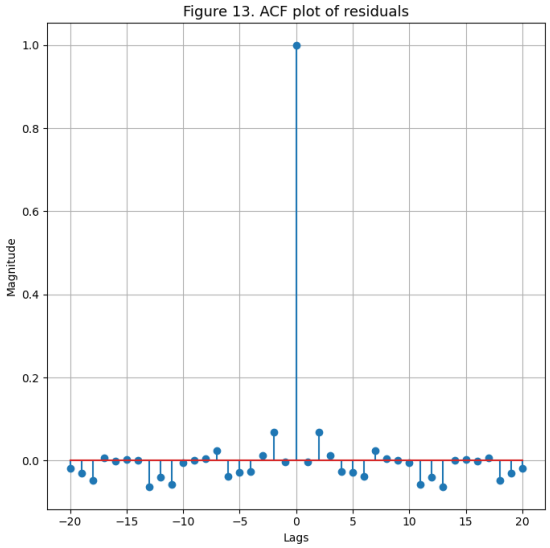
Process	Parameter	LM estimated value	Confidence interval (95%)	Roots
AR	a1	-0.7941	[-0.8176, -0.7707]	0.9891
	a2	-0.0095	[-0.0329, 0.0139]	0.8056
MA	b1	-0.9891	[-0.9930, -0.9852]	-0.0118

According to the above table, the confidence interval of the second parameter of AR process contains 0, which means that the parameter is not significant from 0 with 95% confidence level. Therefore, the parameter can be removed to simplify the model to ARMA (1,1).

In summary, both ARMA (5,5) and ARMA (2,1) reduced to ARMA (1,1). Levenberg-Marquardt algorithm is used to estimate the parameters for the simplified ARMA model.

The final ARMA (1,1) is described as: $y(t) - 0.8010(t-1) = e(t) - 0.9887e(t-1)$





Training set					Testing set		
Mean of errors	MSE	Estimated variance of errors	Variance of errors	Q-value	MSE	Estimated variance of errors	Variance of errors
-0.1469	425.6576	424.0375	425.6360	3781.1640	3706.7210	3710.6854	2818.3680

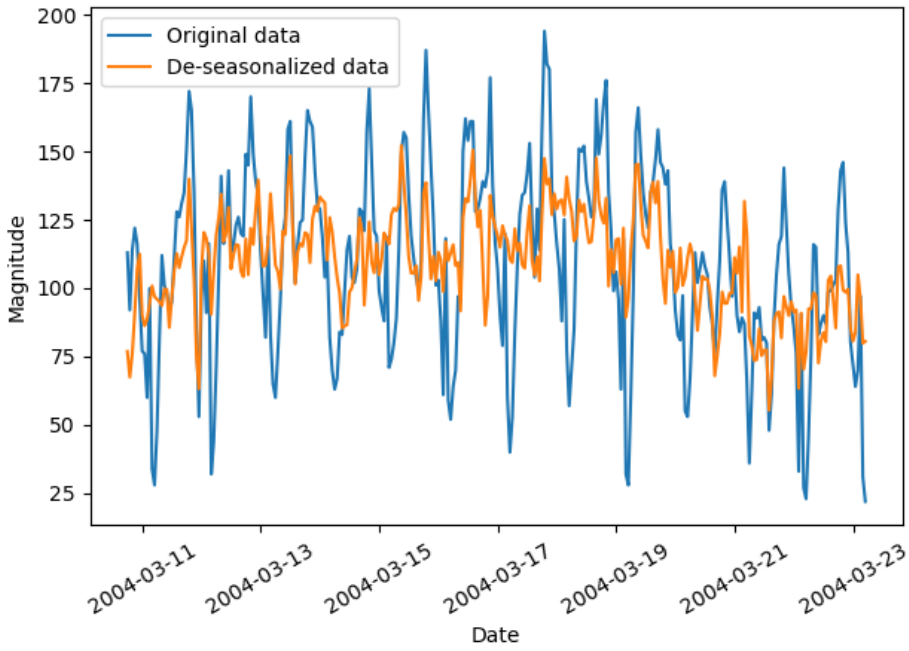
According to the table, the ARMA model is an unbiased estimator since the mean of predicted errors is close to 0. The mean squared value of forecasted errors using ARMA is smaller than using the baseline models, meaning that ARMA is better than the simple forecasting methods and worth to be considered. The ACF plot shows that the residuals from ARMA model are less autocorrelated comparing to other discussed methods.

2. Non-seasonal ARIMA (p, d, q)

To apply ARIMA model to the forecasting, it is necessary to de-seasonalize the original time series and apply transformation to make the seasonal adjusted data stationary.

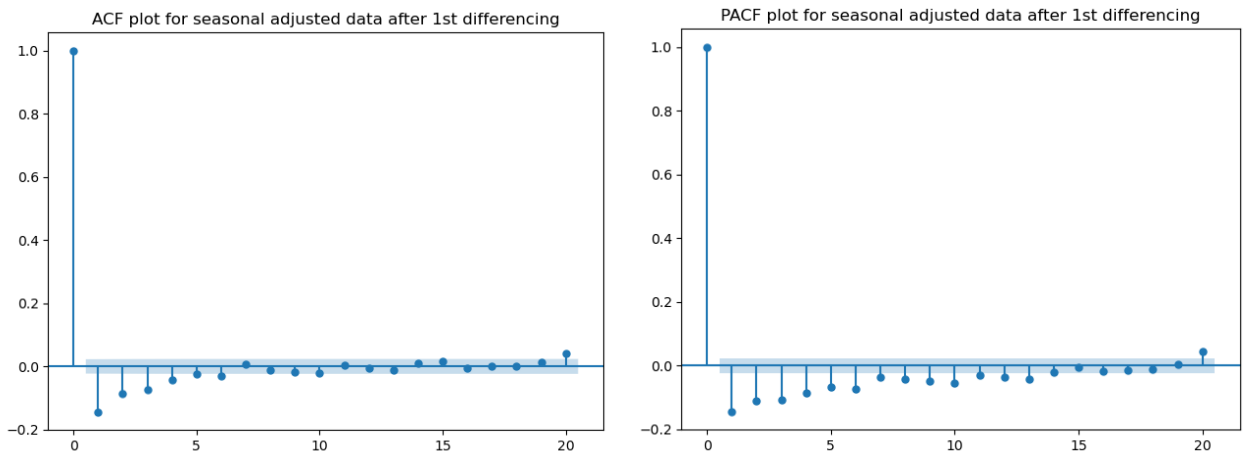
STL decomposition is used to remove the seasonal component from the original data.

Figure 14. Seasonal adjusted data (300 samples)

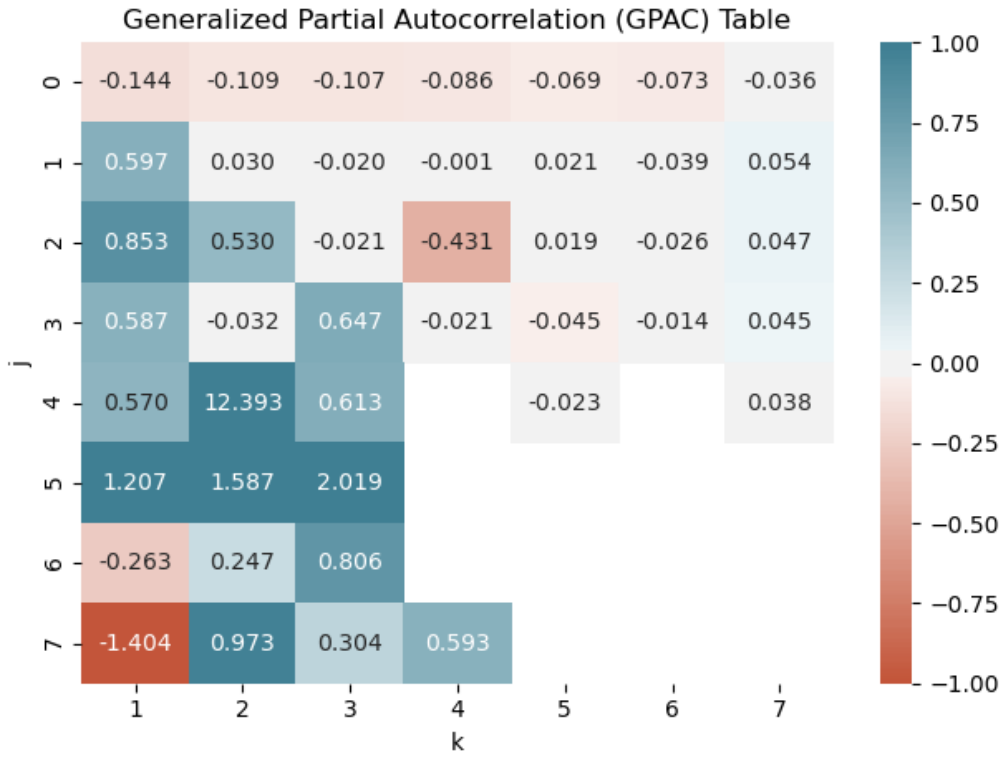


The ADF tests for the seasonal adjusted data indicate a unit root in the rolling variance, which means that the time series is not stationary. An additional 1st differencing helps to make the data stationary since the ADF tests for the differenced data, its rolling mean and its rolling variance reject the null hypothesis of the presence of a unit root in the time series. Moreover, the ACF and PACF plots show a decay to 0 within 10 lags.

The degree of 1st differencing involved in the ARIMA model is 1 as an extra 1st differencing is applied.



After the data is stationary, GPAC table is used to find the appropriate orders of AR and MA processes.



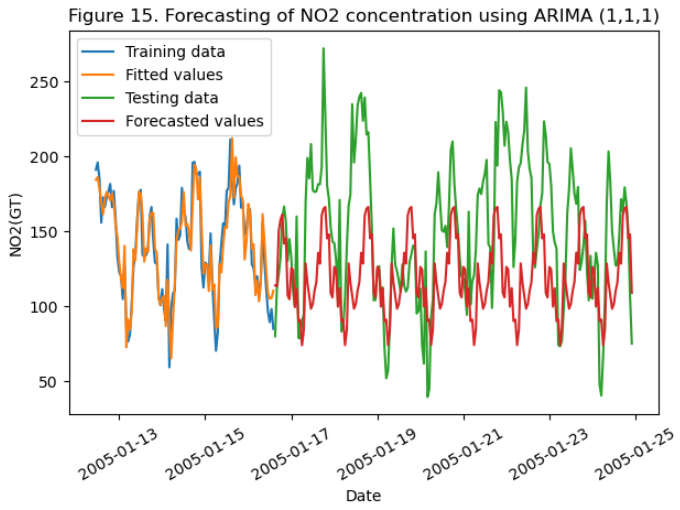
There are two possible cases for AR and MA orders: (p, q) = (1, 1) and (p, q) = (3, 3). Therefore, there are two possible models: ARIMA (1,1,1) and ARIMA (3,1,3). After calculating the roots of the parameters in ARIMA (3,1,3), the model can be reduced to ARIMA (1,1,1) due to zero/pole cancellation.

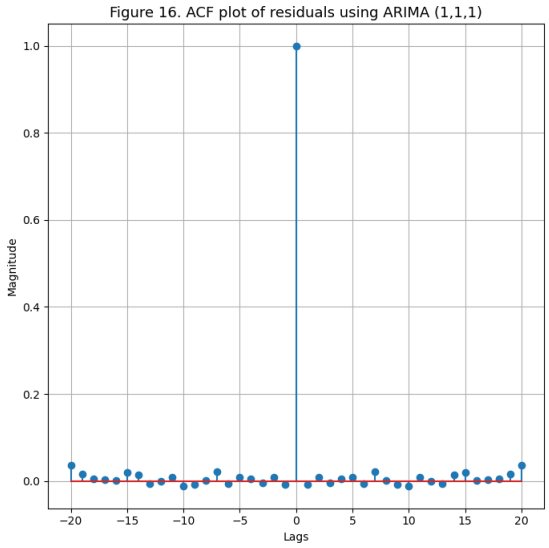
The table below shows the calculation of the roots of AR and MA parameters in ARIMA (3,1,3):

Process	Parameter	LM estimated value	Roots
AR	a1	-1.0836	0.9974
	a2	-0.2474	0.8434
	a3	0.3450	-0.5383
MA	b1	-1.3024	-0.5429
	b2	-0.1497	0.9737
	b3	0.4528	0.6528

As can be seen from the table, a1 can be cancelled by b2, and a3 can be cancelled by b1. The zero/pole cancellation removes two parameters in each process, so the orders of AR and MA are both simplified to 1. The optimized ARIMA model is ARIMA (1,1,1).

ARIMA (1,1,1):





Training set					Testing set		
Mean of errors	MSE	Estimated variance of errors	Variance of errors	Q value of residuals	MSE	Estimated variance of errors	Variance of errors
0.0260	141.7548	141.7927	141.7541	387.0506	2940.7349	2943.8800	2292.1860

According to the table, the mean of residuals is approximately 0, meaning that the ARIMA (1,1,1) is an unbiased estimator. The ACF plot shows that the autocorrelation functions for residuals are close to 0 within 20 lags. However, the critical Q-value is 62.8296 which is smaller than the Q-value of residuals, indicating that the residuals are still not white noise. The autocorrelation functions are in fact much different from 0 for lags beyond 20. The mean squared values of residuals and forecasted errors are both much better than those from simple forecasting methods. In other words, ARIMA model has better performance than the baselines.

3. SARIMA (p, d, q) (P, D, Q) m

According to the ACF plot in Figure 2, the seasonal period of NO2 concentrations is 24, so the seasonal component m is 24. Since the training data is non-stationary with clear seasonality, we first apply a seasonal differencing transformation with lag 24. ADF tests are used to check if the seasonal differenced data is stationary.

Table 8. ADF tests for the seasonal differenced time series with significance level 0.05.

Target of ADF test	ADF statistics	p-value
Seasonal differenced data	-14.8814	0.0000
Rolling average	-15.9123	0.0000
Rolling variance	-0.1775	0.9411

According to Table 8, the rolling mean of seasonal differenced data has a unit root since the p-value of the ADF test is bigger than the significance level 0.05. Therefore, the seasonal data is not stationary, and one or more 1st differencing transformations need to be applied.

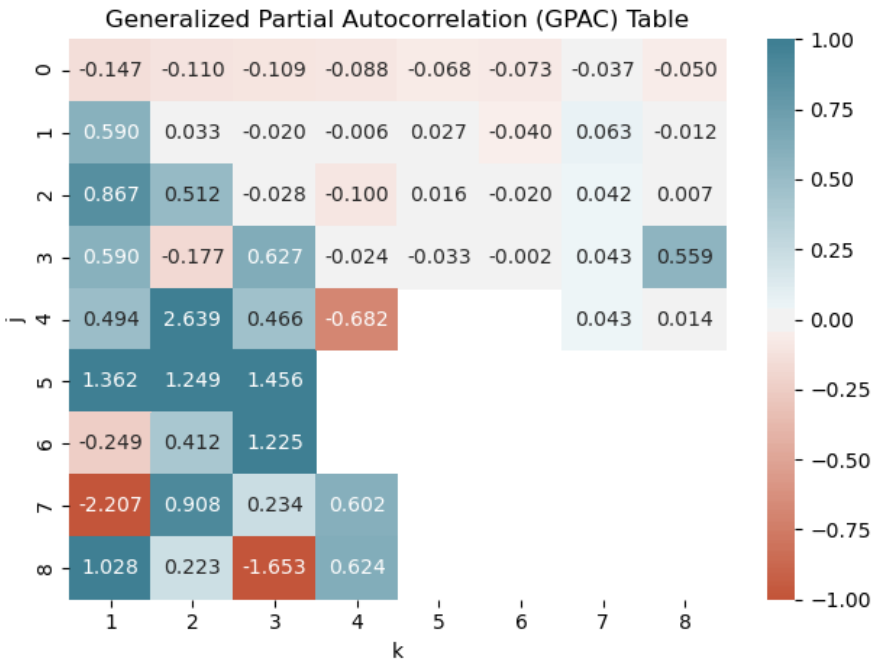
Table 9. ADF tests for the time series after seasonal differencing followed by a 1st differencing

Target of ADF test	ADF statistics	p-value
Seasonal differenced data	-22.0990	0.0000
Rolling average	-12.2562	0.0000
Rolling variance	-8.3777	0.0000

According to Table 9, all three ADF tests result in smaller p-value than the significance level 0.05, meaning that the differenced data, its rolling average and its rolling variance have no unit roots. In other words, the differenced data is stationary.

Since a seasonal differencing and a 1st differencing transform the training data to be stationary, the degrees of seasonal differencing and 1st differencing involved are both 1. In other words, d = 1, and D = 1.

a. Non-seasonal orders of SARIMA model:



As can be seen from the GPAC table above, there is a row of 0s at j = 1, and the column k = 1 seems to have roughly similar values.

Therefore, a possible case for orders of non-seasonal AR and MA processes is: p = 1, and q = 1.

b. Seasonal orders of SARIMA model:

ACF and PACF plots of the differenced data:

Figure 17. ACF plot of differenced data

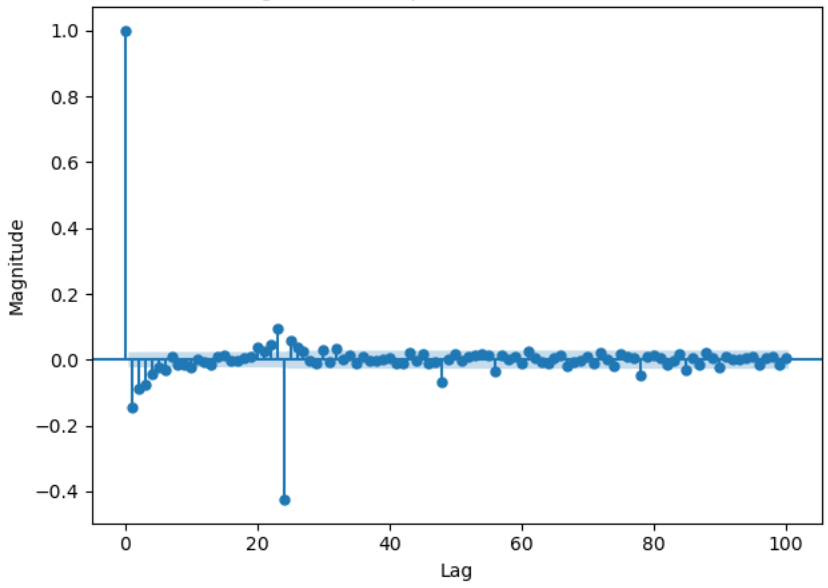
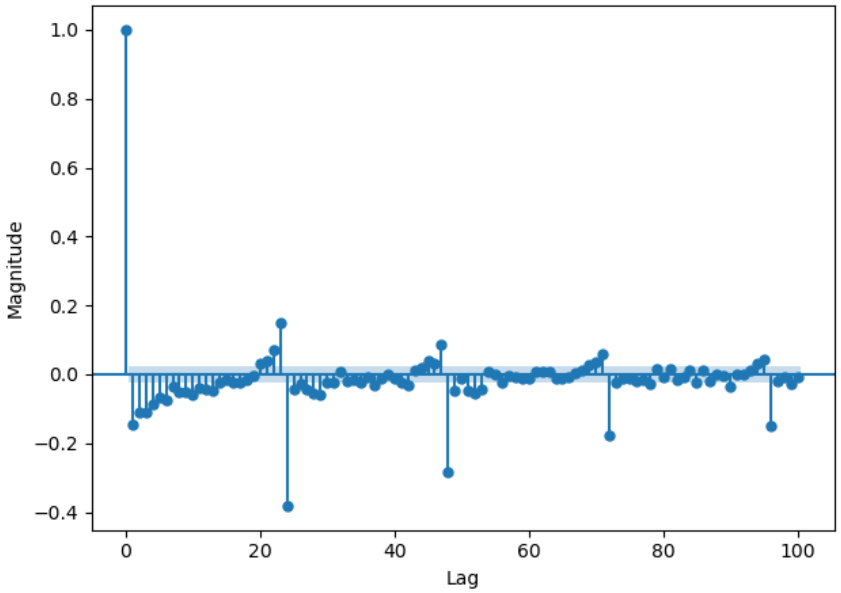


Figure 18. PACF plot of differenced data



According to Figure 17 and Figure 19, there is a significant spike at lag 24 in the ACF plot but no other significant spikes, whereas the PACF shows an exponential decay in the seasonal lags (at lags 24, 48, 72 …). These patterns suggest a seasonal MA (1) component, but there is no seasonal AR component. Therefore, the seasonal terms for SARIMA model are: P = 0, and Q = 1.

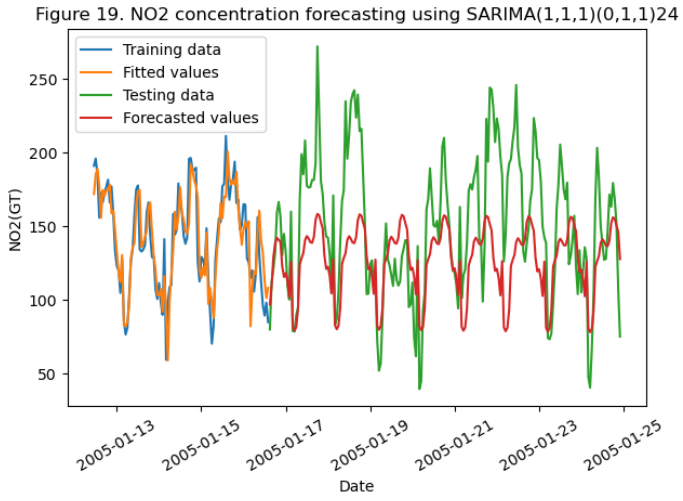
Consequently, an appropriate SARIMA model would be SARIMA (1, 1, 1) (0, 1, 1) 24.

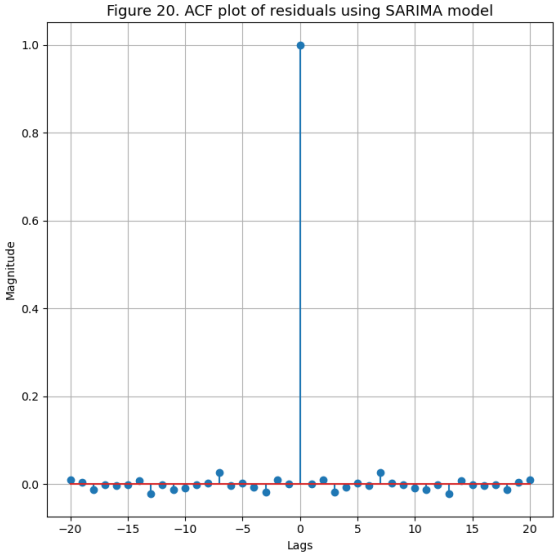
SARIMA (1,1,1) (0,1,1) 24

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                           NO2(GT)   No. Observations:                 7485
Model:             SARIMAX(1, 1, 1)x(0, 1, 1, 24)   Log Likelihood              -31382.611
Date:                            Tue, 15 Dec 2020   AIC                          62775.222
Time:                                    02:05:45   BIC                          62809.808
Sample:                                03-10-2004   HQIC                         62787.102
                                     - 01-16-2005
Covariance Type:                              opg
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
ar.L1                     0.7807      0.011     71.920      0.000       0.759       0.802
ma.L1                    -0.9634      0.004   -250.811      0.000      -0.971      -0.956
ma.S.L24                 -0.9079      0.004   -224.071      0.000      -0.916      -0.900
var.measurement_error     6.4726      3.482      1.859      0.063      -0.352      13.297
sigma2                  251.1862      7.223     34.774      0.000     237.029     265.344
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):              1307.30
Prob(Q):                              0.89   Prob(JB):                         0.00
Heteroskedasticity (H):               1.24   Skew:                             0.11
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.04
===================================================================================
```

The SARIMA model can be described as:

$$(1 - 0.7807q^{-1})(1 - q^{-1})(1 - q^{-24})y_t = (1 - 0.9634q^{-1})(1 - 0.9079q^{-24})e_t$$





Training set					Testing set		
Mean of errors	MSE	Estimated variance of errors	Variance of errors	Q value of residuals	MSE	Estimated variance of errors	Variance of errors
-0.0428	266.7485	266.8555	266.7467	48.5273	2762.6185	2767.0529	1917.7966

Since the mean of residuals are close to 0, the SARIMA model is an unbiased estimator. The model has lower mean squared value of errors than the baseline models, meaning that the method is worth to be considered.

It is worthwhile to note that the Q-value of residuals from SARIMA model is the lowest among discussed models. The critical Q-value of Box-Pierce test is 61.6562 which is higher than the Q-values of residuals, which suggests that the residuals are white noise. Figure 20 also shows that the autocorrelation functions of residuals are approximately zero since lag 1.

In summary, the SARIMA model is the only method that meets the expectation that the residuals are white noise.

Model Evaluation

Since all advanced time series forecasting methods have better performance than the baseline models, it is worth to consider them for the prediction of NO2 concentration.

The table below summarizes the evaluation metrics such as mean of squared errors, variance of errors and Box-Pierce test's Q-value from Holt's Linear, Holt-Winter, Multiple Linear Regression with meteorological variables, ARMA (1,1), ARIMA (1,1,1) and SARIMA (1,1,1) (0,1,1) 24.

Table 10. Mean of squared errors, variance of errors and Box-Pierce test results of time series methods.

Model	Training set				Testing set	
	MSE	Variance of errors	Q-value of residuals	Critical Q	MSE	Variance of errors
Holt's Linear	905.9211	905.9199	25329.7241	65.1707	3289.4052	2820.4764
Holt-Winter	521.6586	521.6584	8513.8148	65.1707	3129.2914	2038.5278
MLR	1433.4221	1433.42208	36938.2032	61.6562	3732.4824	2762.2650
ARMA	425.6576	425.6360	3781.1640	62.8296	3710.6854	2818.3680
ARIMA	141.7548	141.7541	387.0506	62.8296	2943.8800	2292.1860
SARIMA	266.7485	266.7467	48.5273	61.6562	2767.0529	1917.7966

According to Table 10, the MLR has the worst performance due to the highest mean of squared forecasted errors and the highest Q-value. The regression consists of only 3 meteorological features, which are not enough to provide explanatory information to capture the seasonal behaviors of the target variable. To improve the performance of the MLR, more time series weather variables such as pressure, cloud coverage and wind properties need to be recorded.

SARIMA model have the best performance as the model has the lowest mean of squared forecasted errors and the lowest Q-value. It is worthwhile to note that SARIMA is the only model that have smaller Q-value that the critical Q-value, which means that residuals of the method are white noise. Although the ARIMA model has the lowest mean of squared residuals, the difference between the variance of residuals and forecasted errors is relatively high comparing to the difference in the SARIMA model. This suggests potential overfitting problem in the ARIMA.

Future Improvements

There are two main limitations in this project: the missing data and the shortage of meteorological variables to explain the relationship between NO2 concentration and the weather.

There are many ways to address these two issues and improve the time series forecasting of NO2 concentration in this project in future research:

1. A more effective and complicated imputing method would provide better representation for missing values. For example, an imputation using a seasonal time series method can preserve the seasonal behavior of the original dataset.
2. A detailed fine tuning for parameters such as smoothing level, smoothing seasonality and smoothing trend in Holt-Winter is necessary for finding the most optimal setting for the model.
3. Recording data for other meteorological variables such as pressure, cloud coverage, wind speed and wind direction provides better explanation of the relationship between NO2 concentration and the weather, which can improve the performance of multiple linear regression.
4. Since the orders of AR and MA processes in ARMA, ARIMA and SARIMA are found using theoretical methods such as GPAC table, ACF and PACF plots, there is no guarantee that they are the most optimal values. Additional experiment and fine tuning for different orders of AR and MA processes would help to produce the best option for ARIMA models. Furthermore, we can improve the predictive power of ARIMA methods by including meteorological regressors into the models.

Conclusion

This project proposes a time series analysis and modeling to predict the concentration of NO2 at ground-level. The best time series forecasting method is SARIMA (1,1,1) (0,1,1) 24 since the model has the lowest mean of squared forecasted errors and it is the only model with white residuals. However, there is still a high difference between the prediction and the forecast due to the appearance of missing data. A better imputing method would be analyzed to maintain the seasonality in the imputed values. Moreover, SARIMA model with regressors is also a promising method to improve the predictive power of the time series forecasting.

References

https://otexts.com/fpp2/

https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

https://online.stat.psu.edu/stat510/lesson/4/4.1