IE 30301 – Time-series Analysis (FALL 2022)

Ulsan Air pollution project

20201352 Jisu Shin

20202040 Tsoy Roman

20201191 Taean Yoo

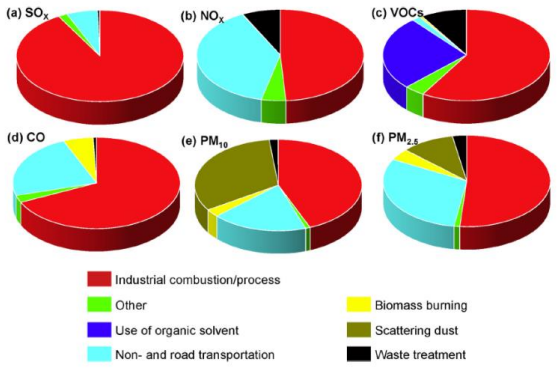
Contents

I. Introduction and Problem Setting

1.1 Introduction

Ulsan is a representative industrial city in South Korea with the largest petrochemical, automobile, and shipbuilding industries. Since many industry complexes are concentrated in Ulsan, the emission of CAPs(Criteria air pollutants) is also serious.

SOx, NOx, VOCs, CO, PM10 and PM2.5 are major pollutants emitted from Ulsan. At this time, Volatile Organic Compounds (VOCs) generate PMs, and O3 is generated by VOCs and NO2. These substances are not only harmful to the human body, but also have a serious adverse effect on the Korean environment.



Source contributions to CAPs based on the emission inventory data in 2017 in Ulsan, South Korea.[1]

'The Ulsan Institute of Health and Environment' announced that as a result of operating the ozone warning system this year, warnings were issued on the 23days (48 times). This is an increase of 10 days (26 times) from 13 days (22 times) during the same period last year. In 2019, 25 times were issued on the 10th and 17 times on the 7th in 2020. Considering that it was issued 25 times in 2019 and 17 times in 2020, it is on the rise.[2] In addition, a study found that every time the concentration of fine dust rises by 10㎍/㎥ in Ulsan, the risk of death increases by 4.9%. On the other hand, in Seoul, the risk of death

1 Vuong, Q. T., Park, M.-K., Do, T. V., Thang, P. Q., & Choi, S.-D. (2022). Driving factors to air pollutant reductions during the implementation of intensive controlling policies in 2020 in Ulsan, South Korea. Environmental Pollution, 292, 118380. https://doi.org/10.1016/j.envpol.2021.118380

2 곽 시열. (2022, October 24). 지구온난화 탓?…울산 오존주의보 해마다 증가. Retrieved from http://www.munhwa.com/news/view.html?no=2022102401039927108001.

only increased by 0.6%. It is because of the difference in the quality of fine dust.[3] Plus, the degree of air pollution in Ulsan is higher than that of Seoul at certain times.

Also, this air pollution has a very big impact on our health. The cancer incidence rate of residents around the Ulsan National Industrial Complex was found to be significantly higher than that of other industrial complexes. According to the report, from 1999 to 2013, men had 1.61 times, and women had 1.33 times more cancer than the country in exposed areas within 14 kilometers of Ulsan Industrial Complex.[4]

1.2 Problem Setting

Therefore, the problem of air pollution in Ulsan is the most closely related problem in real life for us living in Ulsan. However, there is a more serious problem with air pollution in Ulsan. The reason is that the degree of air pollution in Ulsan varies greatly from season to season. Also, the season or time when the air pollution level is high differs significantly from other regions.

In the case of Ulsan, the high concentration of fine dust in the summer of 2018 (July 10-19) was unusually high compared to other regions.[5] Therefore, air pollution in Ulsan is a serious local problem, but it is difficult for residents to predict or prepare.

Therefore, air pollution in Ulsan is a serious local problem, but it is difficult for residents to predict or prepare. So, Our team would like to use Ulsan's emission of CAPs(O3, PM2.5) and seasonal time series data to making predict model and investigate when to prepare.

1.2.1 Emission of CAPs

Among the various emission of CAPs(Criteria air pollutants), we chose two types of PM2.5 and O3 which pollutants warning are issued and people can clearly know the severity.

1.2.2 Seasonal Datas

We also selected temperature and wind direction data that can clearly indicate seasonal changes in Ulsan. The reason wind direction is related to the season is because of the monsoon. Monsoons are the name of wind that direction change with the seasons. The word monsoon comes from the Arabic word mausim, which means "season." A monsoon is caused by a seasonal shift in the winds. The winds shift because the temperature of the land and the temperature of the water are different as seasons change. For example, at the beginning of summer, the land warms up faster than bodies of water. Monsoon winds always blow from cold to warm. In the summer, warm air rising off the land creates conditions that reverse the direction of the wind.[6] Therefore, in summer, the wind blows from the sea to the land, and in winter, the wind blows from the land to the sea.

Ulsan is greatly influenced by monsoons due to its locational characteristics that it faces the East Sea.

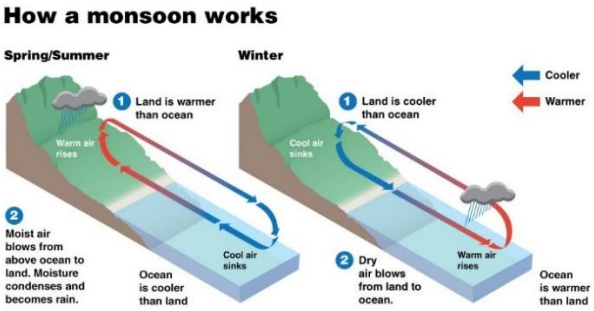
---

3 배 문규. (2019, February 14). 미세먼지 이제는 '질' 관리도…울산이 서울보다 사망위험이 높은 이유. Retrieved from https://m.khan.co.kr/national/national-general/article/201902141634011#c2b.

4 financial. (2020, June 14). [단독]산단 1 번지 '울산의 그늘'… 암 발생률 1 위. Retrieved from https://www.fnnews.com/news/202006141717282703.

5 정 세홍. (2019, January 1). 겨울보다 여름 더 심각한 울산…맞춤형 미세먼지 대책 시급. Retrieved from http://www.ksilbo.co.kr/news/articleView.html?idxno=675757#08fn.

6 What is a monsoon? NOAA SciJinks – All About Weather. (n.d.). Retrieved December 21, 2022, from https://scijinks.gov/what-is-a-monsoon/



II. Data Description and EDA

2.1. Data Description

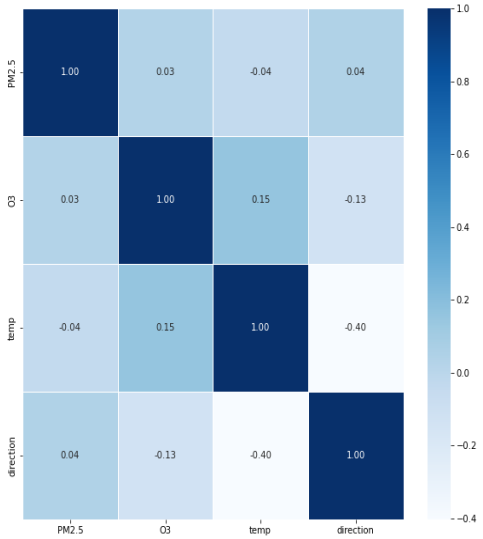
Fine dust and Ozone data are collected from "Air Korea", the air environment information website which is operated by the Korea Environment Corporation. This data is based on the National Air Pollution Information Management System (NAMIS) that collects and manages air information data such as sulfur dioxide, carbon monoxide, nitrogen dioxide, ozone, and fine dust. We concat each month's data and preprocessed it into three years data from January 1, 2019 to December 31, 2021. There were three null values in ozone. So, we replaced it as the average value. The temperature data and wind direction data (January 2019 to December 2021) are collected from the Korea Meteorological Administration and integrated with O_3 *and* $PM_{2.5}$data. In total, the final data have 5 column ($O_3, PM_{2.5}, temperature,$ $wind\ direction$) and 1096 rows. Then we model this data setting 'Datetime' column as an index.

	DateTime	PM2.5	O3	temp	direction
0	2019-01-01	19.0	0.024	0.1	340
1	2019-01-02	11.0	0.024	0.5	320
2	2019-01-03	13.0	0.019	1.4	340
3	2019-01-04	17.0	0.010	4.3	340
4	2019-01-05	52.0	0.029	4.3	340
...	...	...	...	...	...
1091	2021-12-27	7.0	0.034	-1.7	290
1092	2021-12-28	23.0	0.020	3.2	340
1093	2021-12-29	23.0	0.025	4.8	250
1094	2021-12-30	13.0	0.037	1.8	290
1095	2021-12-31	5.0	0.037	-1.5	320

1096 rows × 5 columns

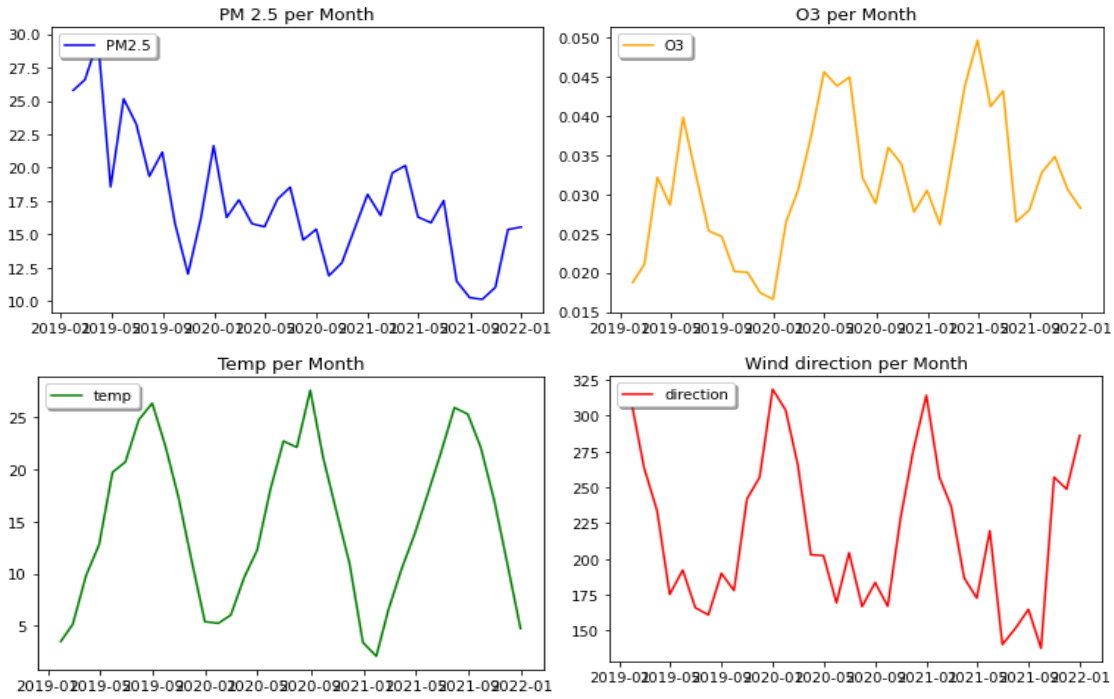
2.2 EDA

2.2.1 Correlation heatmap



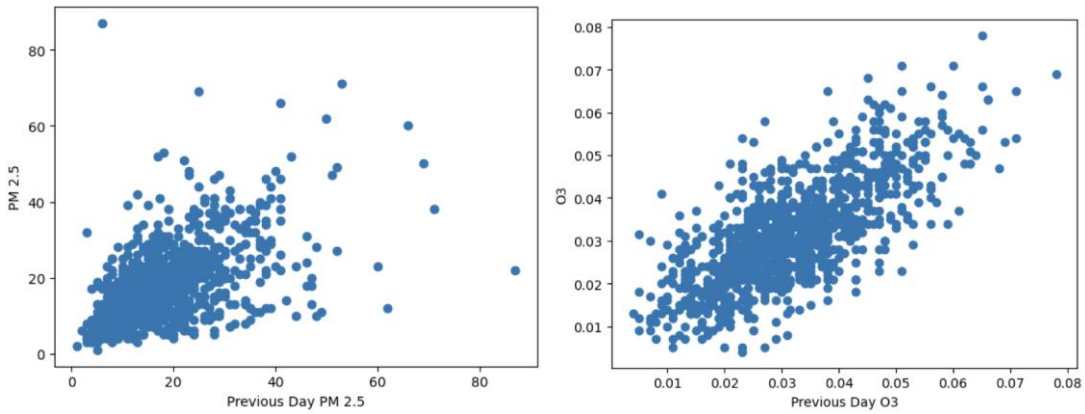
The table on the left shows the correlation between Data colum ($O_3, PM_{2.5}, temperature,$ $wind\ direction$) as heatmap. The map that has dark color means has relatively high correlation and bright color means relatively low correlation. Since the target variable we want to forecast to solve the problem we defined is O_3 *and* $PM_{2.5}$, we check the correlation focusing on the target variables.In the case of $PM_{2.5}$, there is no correlation with other variables exceeding 0.05. In other words, it shows a weak correlation with other variables like O_3 ,weather and wind direction . On the other hand, in the case of O_3, there is a high correlation with temperature and wind direction (over 10%). So, we will forecast how much warning will be occured by forecasting with the univariate model for $PM_{2.5}$ and the multivariate model for O_3.

2.2.2 Data Plot

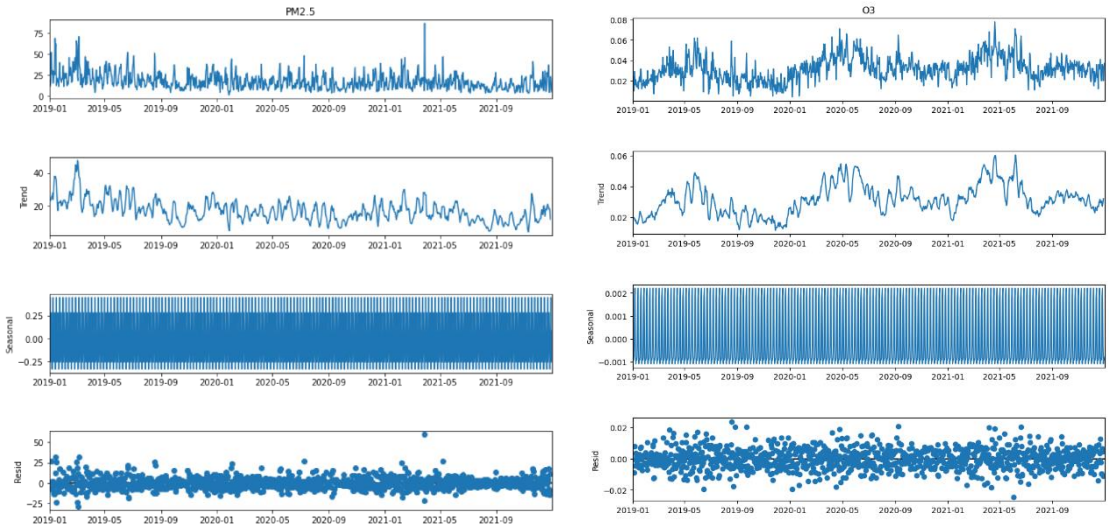


This is a graph that plotted with monthly data which is transformed from day data. Since it has been converted into monthly data, it is possible to understand the flow of data better. Since each data has a different scale, we plotted for each variable. As can be seen from the above heatmap, in the case of $PM_{2.5}$, seasonality does not appear, unlike other variables. However, in the case of O_3, Weather, and Temperature, it shows clear trends and seasonality, such as having a high value in a specific month and a low value in a specific month. This result represents the same result as the heatmap represented above meeting the domain knowledge that temperature has seasonality.

2.2.3 Scatter Plot & Time series decomposition



We draw 2 scatter plots for both PM2.5 and O3 variables in order to observe the dependence between current day and previous for each variable. As it is clearly seen on the plots, PM2.5 has a weak positive correlation with previous day data, and O3 has a strong positive correlation with previous day data.



Next, we performed time series decomposition for both variables - PM2.5 and O3. Both variables do not experience any trend, although for both PM2.5 and O3 there is some seasonality. In addition, it can be said that PM2.5 and O3 experience different seasonality trends, and it also can be explained due to low correlation between variables - 0.03.

III. Model and Evaluation

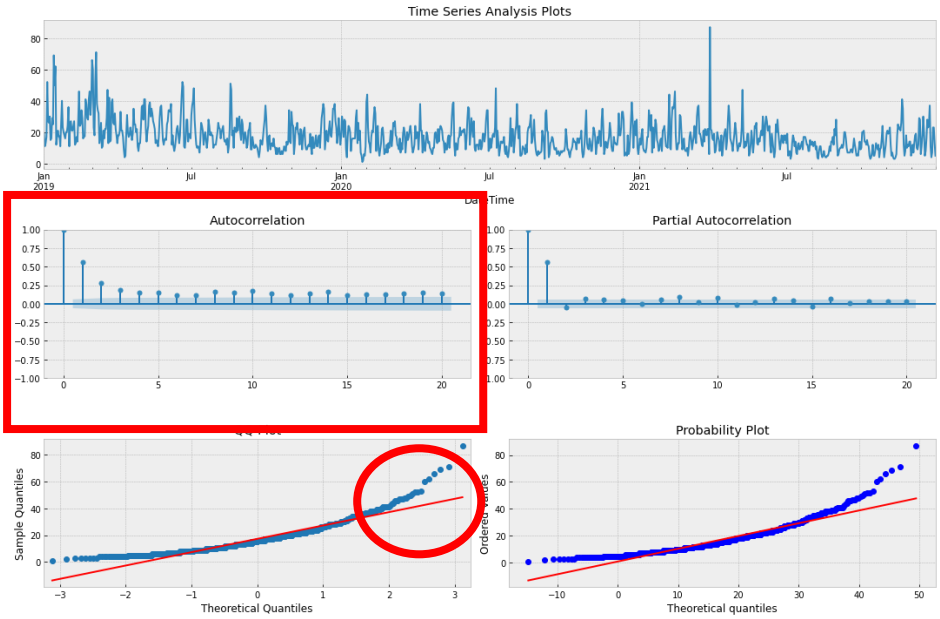
3.1 Traditional Model

3.1.1 Data transformation

We would like to model the $PM_{2.5}$ data as a traditional model at first. We can break traditional time series models into two categories: autoregressive (AR) and smoothing. The former contains models such as ARIMA and SARIMA, while the latter includes exponential smoothing and weighted averaging, to name a few. As I mentioned in the EDA, $PM_{2.5}$ does not have seasonality, so we will use the ARMA model. ARMA (Autoregressive Moving average) is a time series model which is partly autoregressive and partly moving average. There is one condition for data to statistical model such as ARMA.

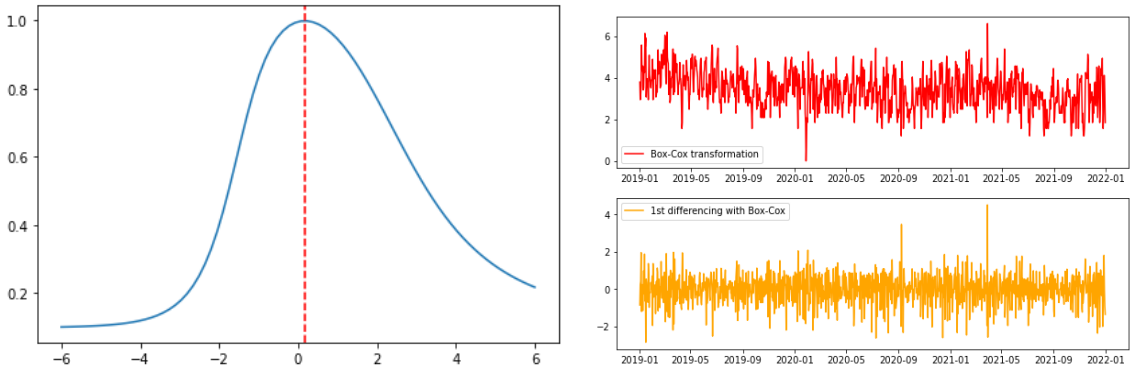
- The data that modeled are stationary.
  1. No significant autocorrelation for all lag k.
  2. Data is normally distributed.

To check our $PM_{2.5}$ meets these conditions, we plot the autocorrelation, Partial autocorrelation, Q-Q plot, and Probability of data.



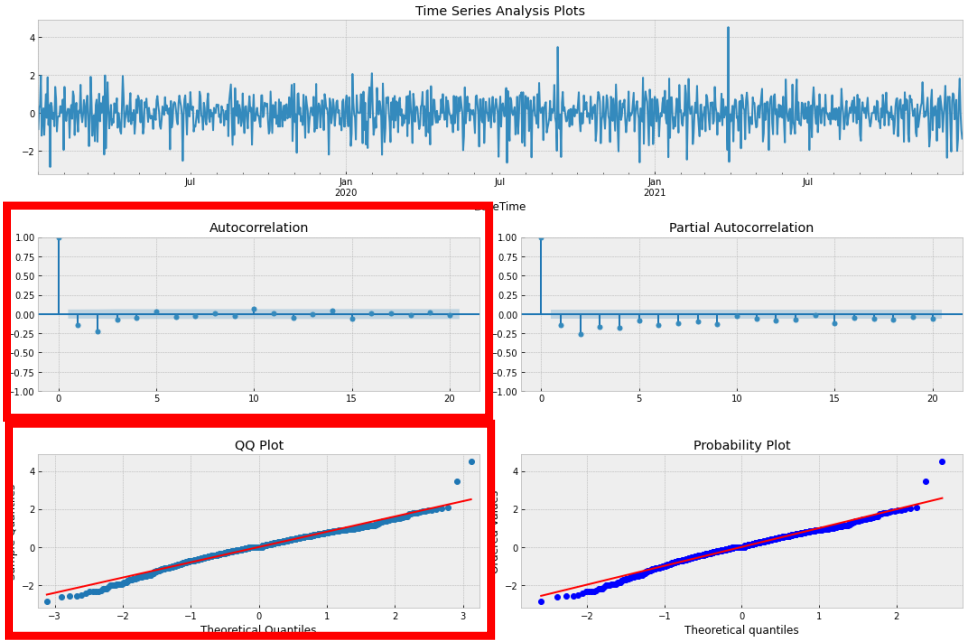
As can be seen in the autocorrelation plot, it shows significant autocorrelations for all lags k. In other words, all values are outside range of error. In addition, according to the Q-Q plot, most all the points are off the red line. Especially, some points at either end. In sum, our data is nonstationary, and it need preprocessing like Differentiating, logarithmic, and Box-Cox transformation etc. At first, we differentiated the data to eliminate the autocorrelation.

The autocorrelation for all lags become insignificant after one differentiation. However, many of the dots are still off the red line. In other words, the data is out of the normal distribution. So we appliled Box-Cox transformation to make the data into a normal distribution. Box-Cox transformation is a transformation method that makes a variable as a normal distribution, which varies depending on the Harper parameter λ..

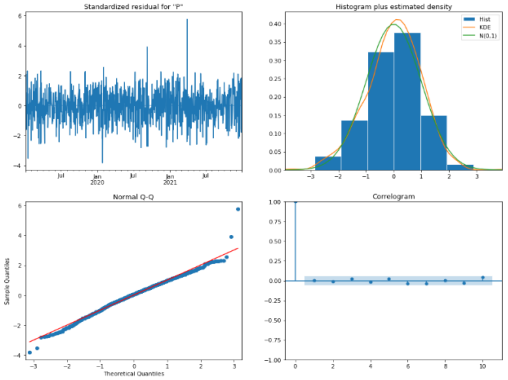


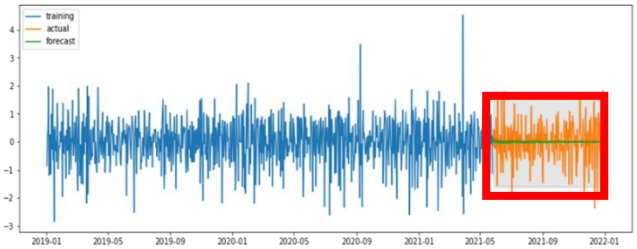
We calculate the optimal λ value (0.1650974401205356) and transform the data. We can see that the orange line which indicate the transformed data is stationary. To check the stationarity, we conduct ADF test. ADF test is the test that can be done via hypothesis testing that is useful to quantify the evidence of non-stationarity in the data-generating mechanism. If p-value is smaller than 0.05(critical value), we can say our data is stationary. Through this test we find that our data is stationary. (p-value is 0.00001).

3.1.2 Parameter estimation



This analysis is the analysis of data that applied first differencing with Box-Cox Transformation. We can find in the ACF plot and Q-Q plot that the data is stationary than the existing data. Before estimating parameters **p, q** based on the data, we can define our model is ARIMA (p,1, q) because we difference our data once. There are various methods of estimating the most suitable parameters. We will use the two methods, auto ARIMA and analysis of ACF and PACF. At first, we analysis the ACF and PACF. Then we find that both autocorrelation and partial autocorrelation die out. Here we can estimate the model is ARIMA (p>0, q>0). In addition, we can also estimate the parameter p will be 5, since autocorrelation die-out after lag 5. Using Auto Arima we find that the most suitable model (the one with the smallest AIC) is **ARIMA (5,1,5).** Therefore, we model the data with ARIMA (5,1,5) and analysis the residual of the model. When the residual becomes white noise, it means the trend of data is well removed. In other words, if the residual is stationary, we can determine that the model is suitable.

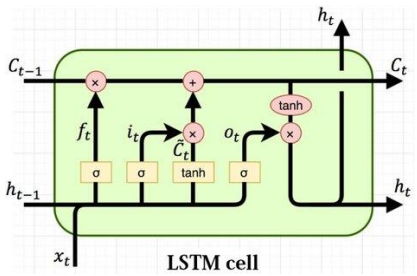


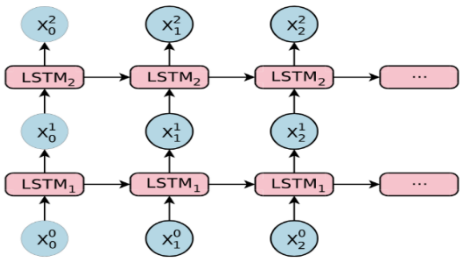


The left graph is the result of plotting the residual of the model [ARIMA (5,1,5)]. Although KDE is slightly skewed to the right, we can say the residual is normal distribution in general. Looking at Q-Q plot, almost all the residuals are on the red line, but some points are off the line at either end. According to the ACF plots there are no value outside range of error. To sum up the residual analysis diagnostic plots, the residual of models looks stationary. So, we can say that our model is fitted quite well. So, we modeled ARIMA (5,1,5) trained on 80% of the data (train set) and forecasted the remaining 20% (test set). The orange graph is the actual value, and the green is the forecast value. However, we can point out one problem. The forecast value is straight. In other words, the forecast value converges to the average value. It is the limitation of AR based model. As the time L (lead time) for the forecast be larger, the forecast value converges to the mean. So, we tried a deep learning-based model to solve the problem.

3.2 Deep learning model

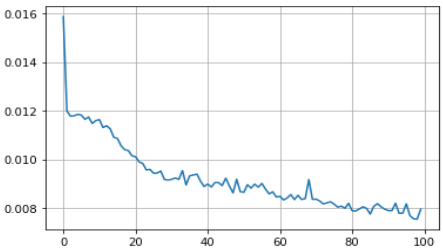
3.2.1 Stacked LSTM



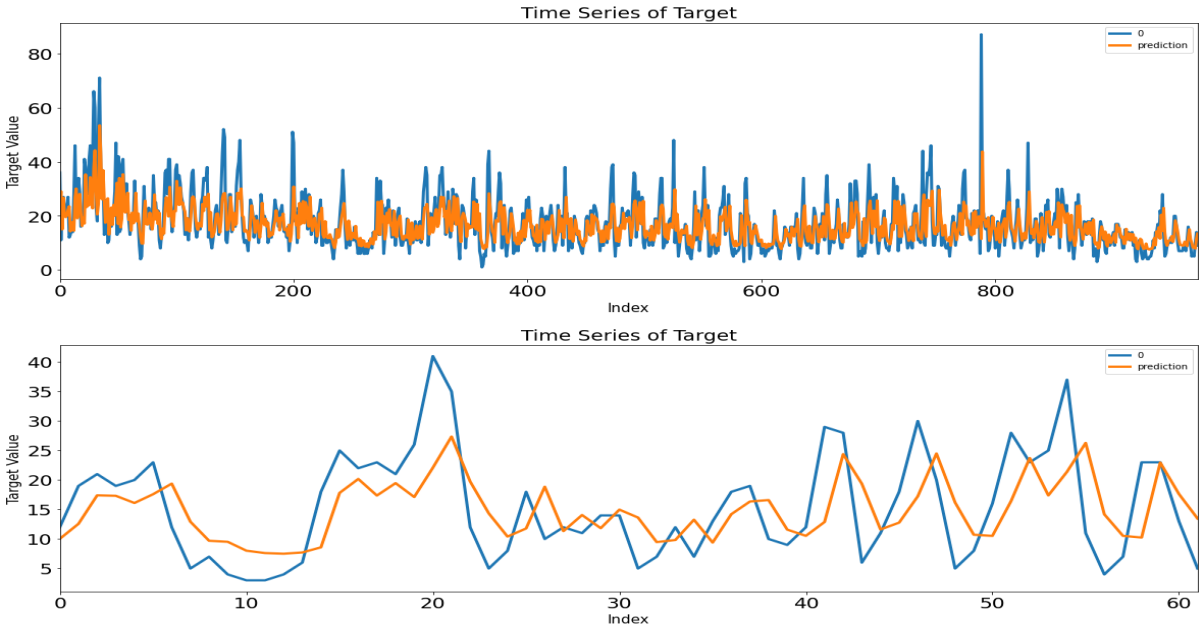


Long Short-Term Memory (LSTM) refers to the structure of a neural network designed to enable long-term/short-term memory, compensating for the shortcomings that conventional RNNs cannot remember long term memory from the output. It is mainly used for time series processing or natural language processing. Long-term memory is important because we will train the model with three years of data and predict 5 days, so we model the data with LSTM rather than RNN. In the presentation, the model learned stationary data(preprocessed) and has three layers. However, we modify the model by accepting the feedback that the model learning didn't occur properly.

```
In [251]: r2_score(Y_test,Y_test_pred)
Out[251]: 0.3243492062527754
```

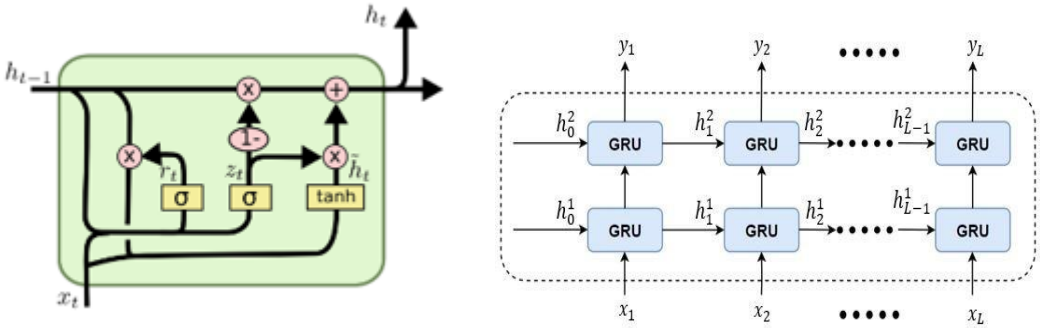


The model was trained with a train set, and the model was tested with a test set for checking the model well trained or not. Train set is the data set from January 2019 to September 2021, and test set is the data from October 2021 to December 2021. In the case of the existing model, the difference between the MSE (mean squared error) of the train set and the test set was large. In other words, the overfitting problem occurred. To solve this problem, we reduced the number of layers in the structure of the model to 3 layers (2 LSTM layers and 1 dense layer) and stacked dropout (ratio=0.2) for each layer. In addition, we increase epoch to 100. We set sequence to 30 as same before. As a result, the R2 score of improved models (0.325) was nearly doubled compared to the existing model. We can also confirm that model well learned both train set and test set with plots.



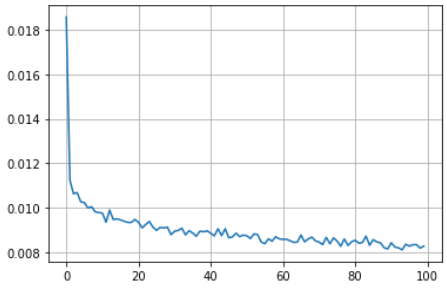
The model forecast 5 days based on 30 days data. We also used the predicted value as data to estimate the future value. In other words, we used 30 days data (December 2021) to predict from January 1 to 6, 2022. At this time, when predicting the value of second days (1/2), the data of first day (1/1) (estimated value) was used.

3.2.2 Stacked GRU

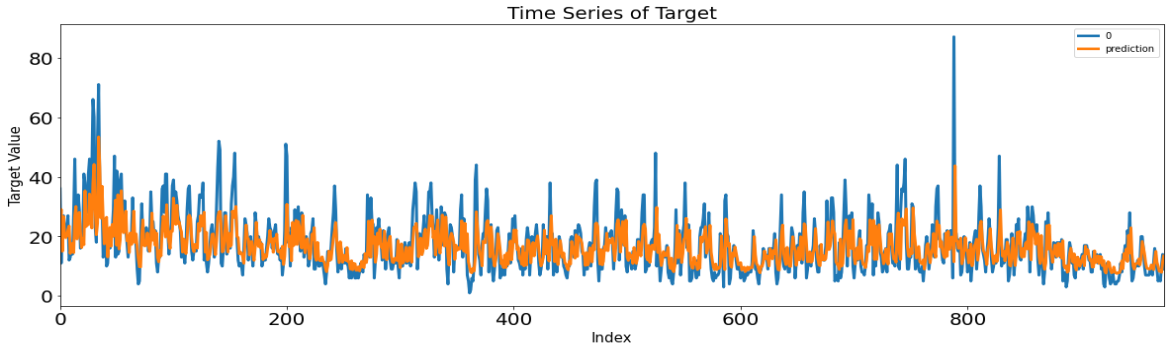


GRU is a simpler structure of the existing LSTM. For LSTM, there were three gates: forget gate, input gate, and output gate, but GRU uses only two gates: reset gate and update gate. In addition, the cell state and hidden state in LSTM are combined as one hidden state. The model has the advantage of having less computation because there are fewer parameteres than LSTM.

```
In [178]: r2_score(Y_test,Y_test_pred)
Out[178]: 0.33761943128303l4
```

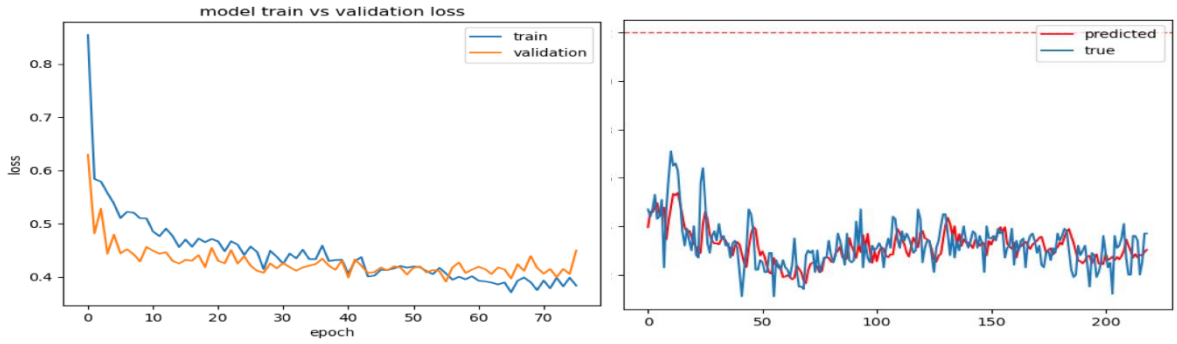


We modeled the GRU in the same condition with LSTM except for layers. The same train set, and test set were used. We also changed GRU structure. The existing model was a total of four layers (3 GRU layer 1 dense layer) but reduced to three layers (2 GRU layer 1 dense layer). Then compile with 'adam' optimizer measuring loss with MSE. In addition, the sequence was set to 30, and dropout was used to prevent overfitting for each layer. In addition, like LSTM, the model learned scale data. After training we forecast original value using inverse_transform in scaler. As a result, the R2 score was 0.337, which was relatively better than the LSTM. The orange graph is the forecast value for train set and the forecast value for test set. We can see that model well forecast the value. Since the model showed higher accuracy (R2 score), we will forecast future value using the GRU model.



3.2.3 Multivariate LSTM

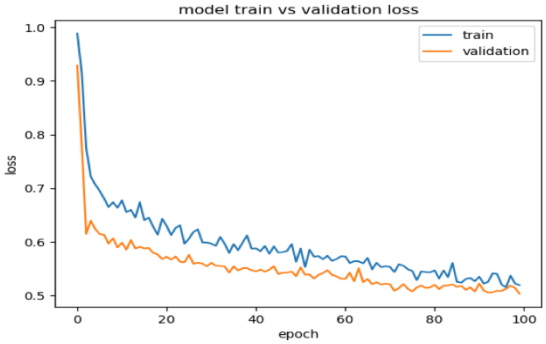
Due to relatively high correlation of temperature and wind with O3, it was decided to perform multivariate analysis for O3. Our team has chosen Multivariate LSTM with exogenous variables - temperature and wind, which will have an effect on forecasting of O3 values. We set the sequence value as 5, which means we will use 5 days to predict the next day. We set epochs number 100 and batch size 64. Next in order to prevent overfitting, we used Dropout Ratio (0.7) and Early stopping with patience 20 and min delta 0.001 which means it will wait 20 iterations for improve for at least 0.001 in validation less, in opposite case, it will stop. After setting all regularization methods, we used 2 layers of Stacked LSTM with input shape (5,3).

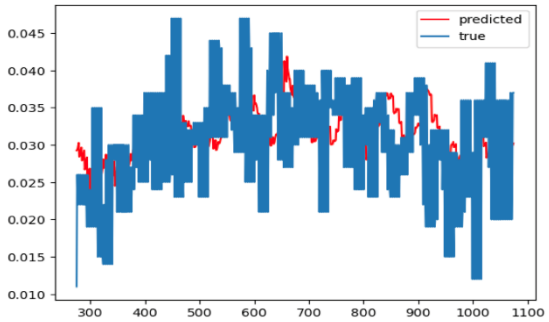


We plotted the train vs validation loss graph, as it is clearly seen, our LSTM model does not overfit, with validation loss almost equal to training loss. We made predictions on Test data, and compared with real value, and as it is clearly seen, our Multivariate LSTM makes a good prediction on Test dataset. Using MSE as an evaluation metric, MSE on the training set was 0.36247, while on the testing set it was 0.402303, which means our Multivariate LSTM with Dense layer (1) performed well. On the graph, there is also threshold value 0.12, which means that if values of O3 go beyond this threshold value, in real life a warning would be given. During this time, in real life no warnings were given, the same was predicted with our LSTM model.

3.2.4 Multivariate LSTM with Dense Layer (5)

Next, in order to make forecasting for future 5 days, we changed from Dense Layer (1) to Dense Layer (5) in the structure of our LSTM model. All other parameters remained the same.



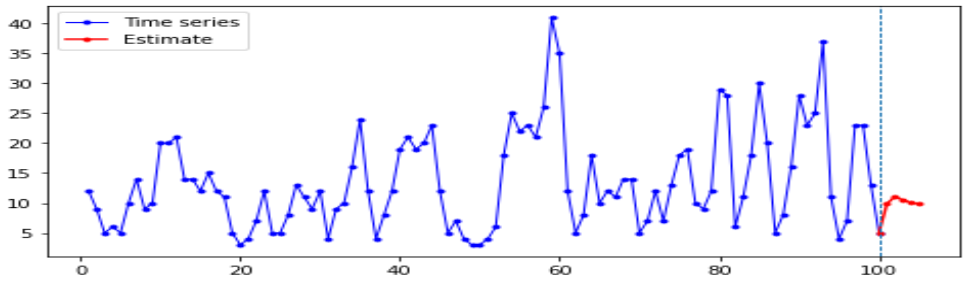


Observing train vs validation loss, our model does not overfit, with train loss almost equal to validation loss by the end of fitting the model. We performed forecasting on a test set, as it can be clearly seen, predicted values follow patterns and values of real test data. By using MSE as we used for Multivariate LSTM with Dense layer (1), it was 0.4687 on training set and 0.4896 on test set, although values are higher compared to Dense Layer (1), the difference between errors on training and testing sets is smaller, which means our LSTM model with Dense layer(5) performs well.

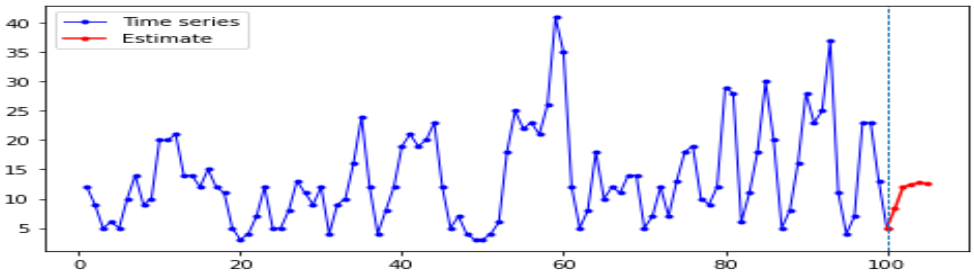
IV. Conclusions

4.1 Result interpretation

4.1.1 PM2.5 forecasting



PM2.5 forecasting using LSTM

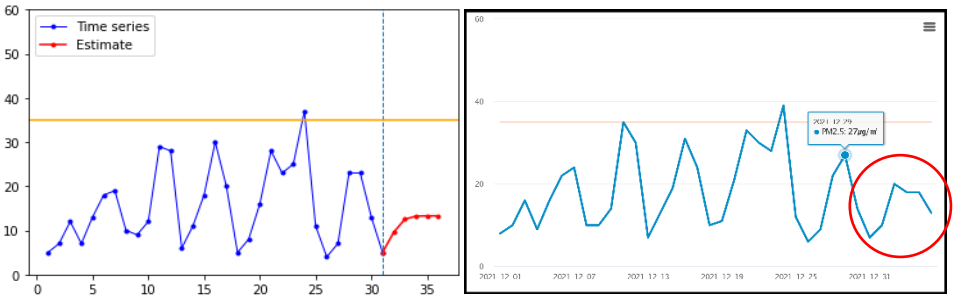


PM2.5 forecasting using GRU

These are Forecasting 5 days using 30 days of data using LSTM and GRU. It can be confirmed that both models are within the data range of the previous 30 days. It also peaked at a low concentration and predicted increasing data like the previous data. Considering these, we can say that both models are well-predicted.

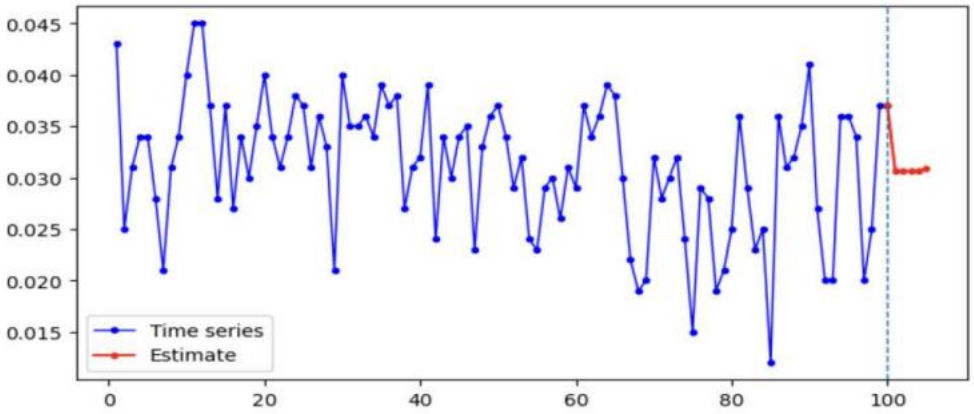
In Ulsan, PM2.5 warning is issued when the concentration of average PM2.5 lasts for 2 hours or more with 75㎍/㎥. Since there are no days that exceed this standard in the current forecasting data, warnings

were predicted not to be issued from January 1 to January 5, and they were not actually issued.



In addition, in order to check whether the model's forecasting came out well, we brought the 'Air Korea' graph and compared it. The left graph shows the measurements from 2021-12-01 to 2022-12-31 and the predictions from 2022-01 to 2022-01-05 predicted by the model, and the right graph shows the measurements from 2021-12-01 to 2022-01-05 drawn by 'Air Korea'. It has a similar trend and value, so it can be said that forecasting has worked well.

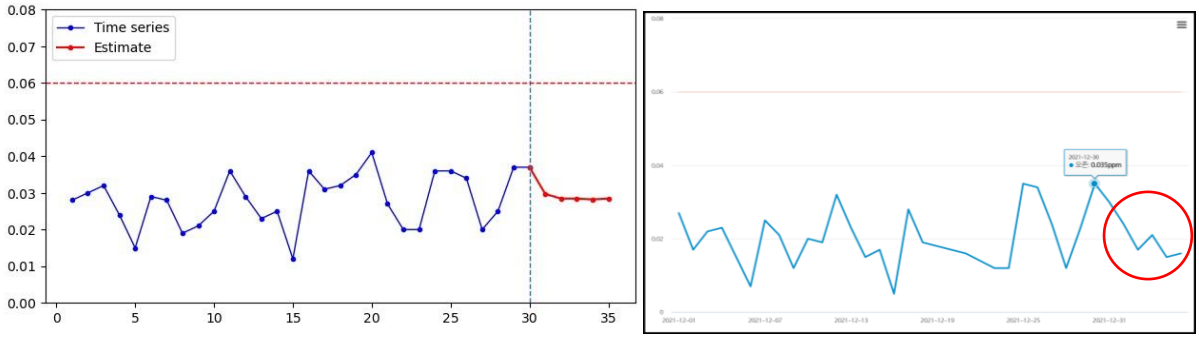
4.1.2 O3 forecasting



PM2.5 forecasting using LSTM

These are Forecasting 5 days using 30 days of data using LSTM. You can see values converge to mean value which is 0.032.

In Ulsan, O3 warning is issued when the concentration of average O3 lasts for one hour or more at 0.12ppm. Since there are no days that exceed this standard in the current forecasting data, warnings were predicted not to be issued from January 1 to January 5, and they were not actually issued.



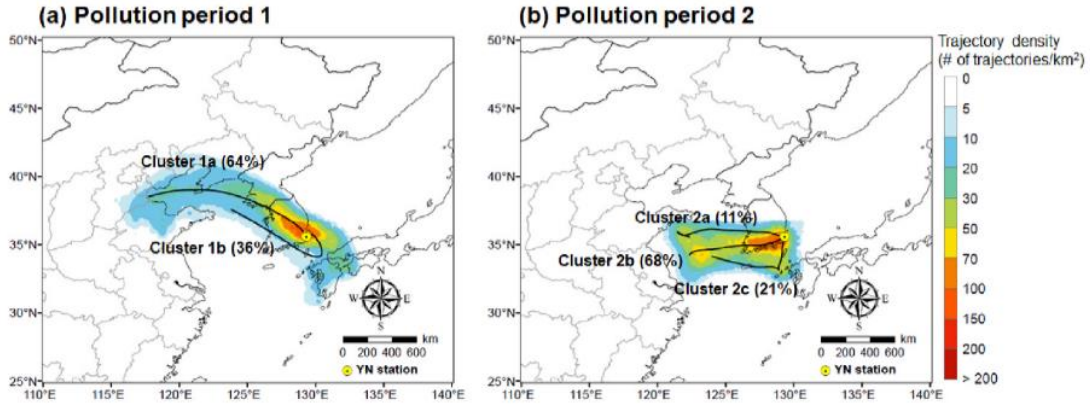
In the same way as before, it was compared with the 'Air Korea' graph. The left graph shows the measurements from 2021-12-01 to 2022-12-31 and the predictions from 2022-01 to 2022-01-05 predicted by the model, and the right graph shows the measurements from 2021-12-01 to 2022-01-05 drawn by 'Air Korea'. It has a similar trend and value, so it can be said that forecasting has worked well.

4.2 Discussion

4.2.1 PM2.5 of Ulsan

Currently, episodes of high levels of particulate matter (PM) with an aerodynamic diameter of less than 2.5 μm (PM2.5) frequently occur in South Korea as a result of both local emissions and the long-range at- mospheric transport (LRAT) of yellow dust and haze events from the Asian continent. The proportion of PM2.5 in South Korea that originates from other countries has been estimated to account for 30–50% of the annual average and reach 60–80% during high PM2.5 episodes.[7] Similar numbers have been reported by the Joint Research Project for Long-range Transboundary Air Pollutants in Northeast Asia, with a domestic contribution to the annual average concentration of PM2.5 in South Korea of 51.2% and the contributions from China and Japan of 32.1% and 1.5%, respectively.[8]

In addition, according to a study by Lee et al. (2023, UNIST), The major source of PM2.5 for the pollution period during winter was LRAT from eastern China and North Korea.[9] According to this study, PM flows into Ulsan through the trajectory shown in the figure below. Therefore, the concentration of PM2.5 is determined by LRAT, regardless of the season. It is also possible to explain for this reason that There is no correlation between PM2.5 and other variables (O3, temperature and wind direction data).



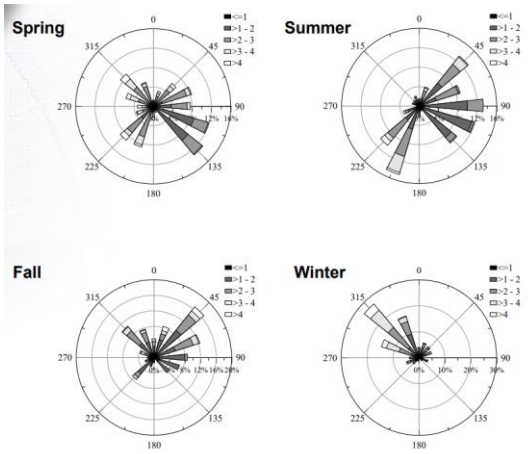
7 Bae, C., Kim, B. -., Kim, H. C., Yoo, C., & Kim, S. (2020). Long-range transport influence on key chemical components of PM2.5 in the seoul metropolitan area, south korea, during the years 2012-2016. Atmosphere, 11(1) doi:10.3390/ATMOS11010048

8 Diapouli, E., Manousakas, M., Vratolis, S., Vasilatou, V., Maggos, T., Saraga, D., Grigoratos, T., Argyropoulos, G., Voutsa, D., Samara, C., & Eleftheriadis, K. (2017). Evolution of air pollution source contributions over one decade, derived by PM10 and PM2.5 source apportionment in two metropolitan urban areas in Greece. Atmospheric Environment, 164, 416–430. https://doi.org/10.1016/j.atmosenv.2017.06.016

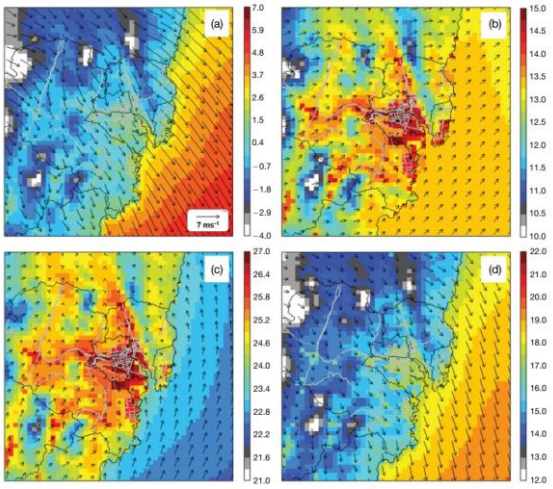
9 Lee, S.-J., Lee, H.-Y., Kim, S.-J., Kang, H.-J., Kim, H., Seo, Y.-K., Shin, H.-J., Ghim, Y. S., Song, C.-K., & Choi, S.-D. (2023). Pollution characteristics of PM2.5 during high concentration periods in summer and winter in Ulsan, the largest industrial city in South Korea. Atmospheric Environment, 292, 119418. https://doi.org/10.1016/j.atmosenv.2022.119418

4.2.2 O3 of Ulsan

In the case of Ulsan, it is in contact with the East Sea and is greatly influenced by seasonal winds (Monsoon). In particular, the northwesterly wind is excellent in winter, and the frequency of the easterly wind is relatively high in late spring and summer. It can be found in the wind rose and the wind vector of the below figure.



Seasonal winds in Ulsan (2005-2009)



Horizontal distributions of the simulated monthly mean temperature and wind vectors for (a) January, (b) April, (c) July, and (d) October. The wind vectors are displayed every three grid points and their lengths are proportional to the wind speeds.

Since the easterly wind prevails in summer and the CAPs emitted from the industrial complex are brought inland, the concentration of CAPs is high. In fall and winter, the northwesterly winds bring large amounts of CAPs emitted from the industrial complexes to the East Sea.

4.3 Further considerations

This project could not predict a long period of time. Since January data were predicted using December data, it did not show seasonality. Therefore, using the models made through this project, we want to predict the long-term with a longer dataset and consider the seasonality. It's worth noting that if predictable with these models, it will be possible to be more careful and prevent them because we know when PM2.5 and O3 concentrations are high, and warnings will be issued

V. Team member contribution

Regarding our responsibilities during doing team project, 유태안 was responsible for good problem setting, and domain knowledge. She analysed situation of air pollution in Ulsan, found variables that would be reasonable to use for our time series forecasting and interpreted results that we got by making time series models. She also made conclusions and explained reasons of obtaining certain results very well.

신지수 was responsible for visualizing monthly/ daily data, making correlation map to decide whether we use multi variate or univariate analysis for our target variables - PM2.5 and O3. She also made data preprocessing, which was used for building ARIMA model. Also, different univariate models like GRU, Stacked LSTM were developed and evaluated by her.

Roman was responsible for finding t and t+1 relation for 2 target variables - PM 2.5 and O3. He also checked each of the variables for seasonality and trend. Next, he developed Multivariate LSTM for O3 variable, and making forecasting for next 5 days. In addition, he also performed Univariate LSTM modeling for O3 and compared performance with Multivariate one to decide which one to use to solve our problem.

Appendix. What we improve compared to the presentation

*** The comments in the blue boxes are feedback from classmates**

(1) Increase the model accuracy

1. GRU for PM 2.5

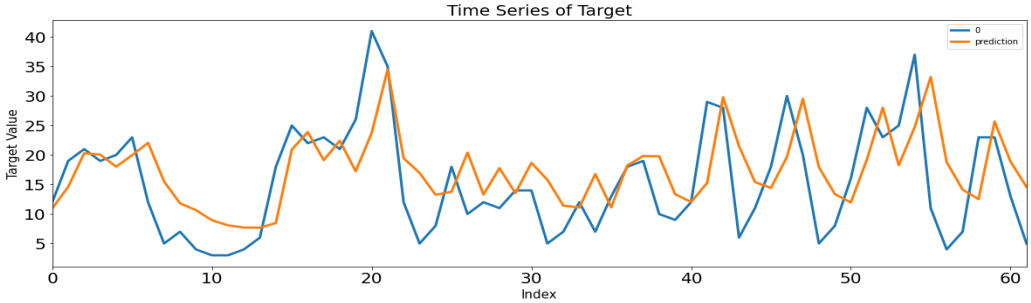
Previously, we learned data with three models, Bidirectional LSTM, Stacked LSTM, and GRU, with stationary data (processed) . However, there were some comments that the models' accuracy are so low (3 feedbacks).

> LSTM model doesn't perform great because of small learning rate, overall LSTM is nice
>
> good EDA, but low model accuracy
>
> good: Detail explanation of LSTM / bad: Model learning doesn't seem to have been done properly.

By accepting this part, we tried to improve the model accuracy. In the case of the existing model, the difference between the loss values of the train set and the test set was large. In other words, the overfitting problem occurred. To solve this problem, four layers were reduced to three layers (2 LSTM/GRU layer and 1 dense layer), and the dropout ratio for each layer was increased from 0.1 to 0.2. To increase accuracy, we increase epoch from 80 to 100. In addition, we changed the train data from stationary data to original data since it would be difficult for model to learn the processed data (stationary) that does not have any trend. These changes were reflected in both the LSTM and GRU. Then, the R2 score of the existing model increase from 0.16 to 0.34. In addition, we can check that the accuracy of the test set has improved using the plot.

(1) GRU forecasting (existing)

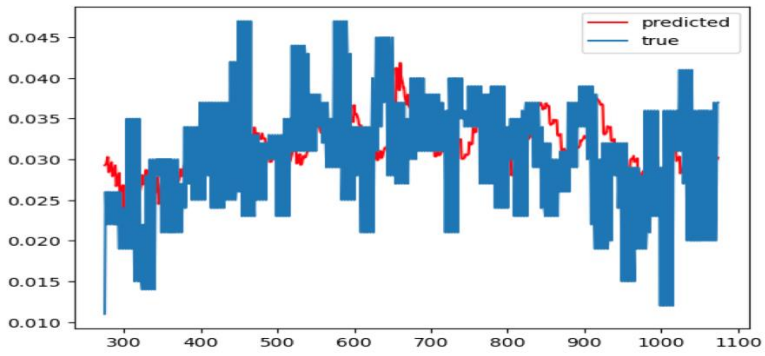
(2) GRU forecasting (Improved)



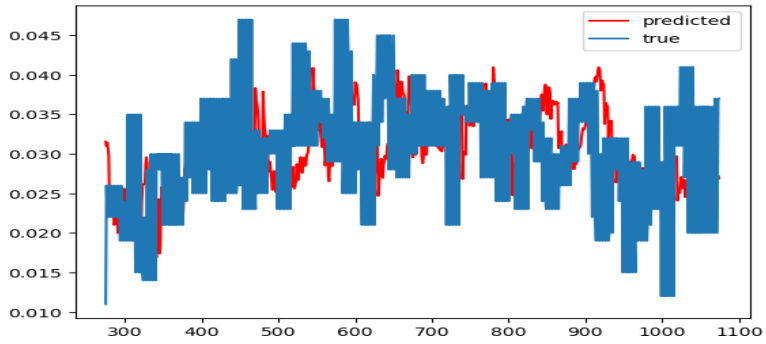
2. Multivariate LSTM for O3

Although our Multivariate LSTM model for O3 performs well, we also tried to improve it. We changed the activation function from relu to linear, and it made our forecasting much better. It improved the MSE on the testing set from 0.4896 to 0.486397, and the difference between training set and testing set became much smaller. We also changed drop out ratio from 0.7 to 0.5, as we used both early stopping and drop out ratio, we decreased drop out ratio value in order to make a model to capture data better.

(1) Multivariate LSTM (existing)



(2) Multivariate LSTM (Improved)



(2) Delete the Bidirectional model

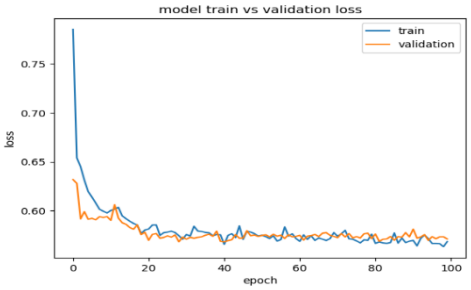
> They performed good EDA, but they should have not used Bidirectional LSTM for their problem.

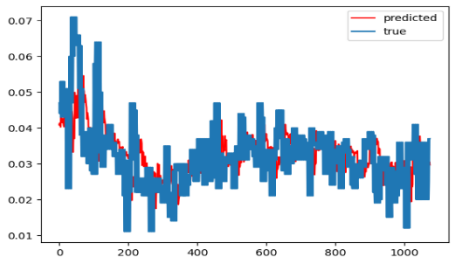
In addition, we received feedback that bidirectional LSTM was not suitable for this data. Before accepting the feedback, we study the model and data more specific and know that the $PM_{2.5}$ is not suitable for the bidirectional LSTM because current data does not affect past data like natural language. So, we removed that model and reduced the number of models from four to three accepting the feedback.

(3) Applying Univariate LSTM for O3 and comparing it with Multivariate one

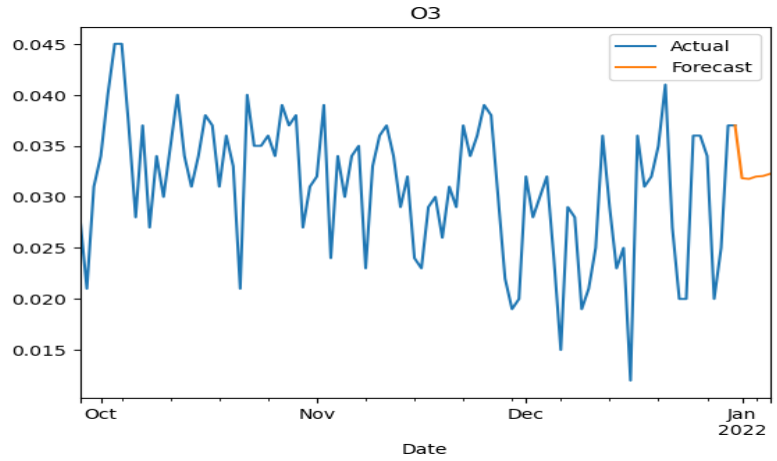
> good: detailed explanation of all parts, bad: maybe would be good to try univariate LSTM and compare with multivariate

We also received feedback that it would be good to compare our Multivariate LSTM with performance of the Univariate LSTM model. So, we did the Univariate LSTM model for O3. We stacked our model with 2 layers of LSTM and drop out ratio 0.5, and with input shape (5,1), instead of (5,3) that we used for the Multivariate one. However, we did not apply early stopping for the univariate model. We also used an activation function - linear, and Dense layer (5) meaning we will predict the next 5 days.





As it is clearly seen on the train vs validation loss graph, our model does not overfit, and performs pretty good on forecasting our test data. But overall, comparing with Multivariate LSTM, it can be concluded that Univariate performs a little bit better. It was also proved by evaluation on MSE. While for Multivariate LSTM it was 0.48639732540323716 on the test set, for the Univariate it was 0.45676965072396203 on the same test set.



We also performed forecasting for the next 5 days from 1st January 2022 to 5th January 2022 that we did for Multivariate LSTM, and as it is clearly seen, both models perform forecasting pretty much the same. The reason why Univariate LSTM performs a little bit better could be that correlation between temperature and O3, wind and O3 is relatively small, and these variables create more noise rather than bring value to the model.

(4) More explain of the relationship with the problem and wind direction data

good point : good choice of problem/bad point: It is hard to think about realtionship problem with data(wind direction)

In the feedback, there was an opinion that the relationship between the problem and the wind direction data was not understandable. By accepting this part, we further strengthened the investigation of the academic part. An explanation of the Monsoon, regional characteristics of Ulsan, and research results on it were added.

References

1. Vuong, Q. T., Park, M.-K., Do, T. V., Thang, P. Q., & Choi, S.-D. (2022). Driving factors to air pollutant reductions during the implementation of intensive controlling policies in 2020 in Ulsan, South Korea. Environmental Pollution, 292, 118380. https://doi.org/10.1016/j.envpol.2021.118380
2. 곽 시열. (2022, October 24). 지구온난화 탓?…울산 오존주의보 해마다 증가. Retrieved from http://www.munhwa.com/news/view.html?no=2022102401039927108001.
3. 배 문규. (2019, February 14). 미세먼지 이제는 '질' 관리도…울산이 서울보다 사망위험이 높은 이유. Retrieved from https://m.khan.co.kr/national/national-general/article/201902141634011#c2b.
4. financial. (2020, June 14). [단독]산단 1 번지 '울산의 그늘'… 암 발생률 1 위. Retrieved from https://www.fnnews.com/news/202006141717282703.
5. 정 세홍. (2019, January 1). 겨울보다 여름 더 심각한 울산…맞춤형 미세먼지 대책 시급. Retrieved from http://www.ksilbo.co.kr/news/articleView.html?idxno=675757#08fn.
6. What is a monsoon? NOAA SciJinks – All About Weather. (n.d.). Retrieved December 21, 2022, from https://scijinks.gov/what-is-a-monsoon/
7. Bae, C., Kim, B. -., Kim, H. C., Yoo, C., & Kim, S. (2020). Long-range transport influence on key chemical components of PM2.5 in the seoul metropolitan area, south korea, during the years 2012-2016. Atmosphere, 11(1) doi:10.3390/ATMOS11010048
8. Diapouli, E., Manousakas, M., Vratolis, S., Vasilatou, V., Maggos, T., Saraga, D., Grigoratos, T., Argyropoulos, G., Voutsa, D., Samara, C., & Eleftheriadis, K. (2017). Evolution of air pollution source contributions over one decade, derived by PM10 and PM2.5 source apportionment in two metropolitan urban areas in Greece. Atmospheric Environment, 164, 416–430. https://doi.org/10.1016/j.atmosenv.2017.06.016
9. Lee, S.-J., Lee, H.-Y., Kim, S.-J., Kang, H.-J., Kim, H., Seo, Y.-K., Shin, H.-J., Ghim, Y. S., Song, C.-K., & Choi, S.-D. (2023). Pollution characteristics of PM2.5 during high concentration periods in summer and winter in Ulsan, the largest industrial city in South Korea. Atmospheric Environment, 292, 119418. https://doi.org/10.1016/j.atmosenv.2022.119418
10. Oh, I., Bang, J.-H., & Kim, Y. (2015). Meteorological characteristics in the Ulsan Metropolitan Region: Focus on air temperature and winds. Journal of Korean Society for Atmospheric Environment, 31(2), 181–194. https://doi.org/10.5572/kosae.2015.31.2.181
11. Kim, S.-J., Kwon, H.-O., Lee, M.-I., Seo, Y., & Choi, S.-D. (2019). Spatial and temporal variations of volatile organic compounds using passive air samplers in the multi-industrial city of Ulsan, Korea. *Environmental Science and Pollution Research*, *26*(6), 5831–5841. https://doi.org/10.1007/s11356-018-4032-5