Time Series Analytics Case Study Report

(AQI changes of Wan Hua district in 2021)

Background

In recent years, as the case of ischemic heart disease, stroke, chronic obstructive pulmonary disease (COPD) increase, severe air pollution issue had drawn a lot of attention. Air pollution is mostly derived from sources of domestic combustion, primarily the burning of fossil fuels, and the any other chemical particles within the urban air, such as PM 2.5, CO, NO. Taipei, surrounded by mountains, is more likely to fall victim to polluted air.

Combining all particle concentrations and the other pollution source together under consideration, the AQI (Air Quality Index) can shows the purity of the air. When AQI is too high, means the air quality is poor; when AQI is high, means the air is clear instead.

We would like to know how the quality of the air changes, whether the air of tomorrow is clear enough or not. Therefore, we try to formulate a time series model to fit the AQI changes, and make prediction of it. Moreover, evaluate the model's performance.

Dataset

This dataset is from the Environmental Protection Administration Executive Yuan, R.O.C (Taiwan), splitting into twelve csv files, in which show multiple particle concentrations within the urban area of the Wan Hua district in 2021 hourly. Link(https://data.epa.gov.tw/dataset/aqx_p_488).

Containing multiple points of interests with specific time stamp, such as concentration of CO, SO, NO, PM2.5 particle, the dataset also including the 8 hours average data for each particle concentrations.

Objective

The goal of the research is to find out the AQI fluctuation with time, whether these changes following a certain trend or pattern. In other words, the AQI is significantly differ form season, or having different mean value between day and night.

Using multiple different parameters setting ARIMA model to fit the training time series data, and accept the one with best performance. Then forecast the testing dataset, analyzing the forecast residual.

Detail steps in Analyzing

1. Generate Datasets

First gathering all AQI data from January to December (2021) at the website, then merged then into one data frame and output it as csv file.

The output csv is (8560x25), having multiple point of interest columns (ex. CO, SO, NO, PM2.5) with each specific time stamp. However, for now we only considering the AQI column, and drops the others.

2. Data Preprocessing

Reading the merged dataset, and checking the missing time stamp. For the accuracy of SARIMA model, we use

yearly mean to generate the missing time stamp data. There are 174 hourly data are not existing.
Also, we rearrange the sequence of the series by the Timestamp.

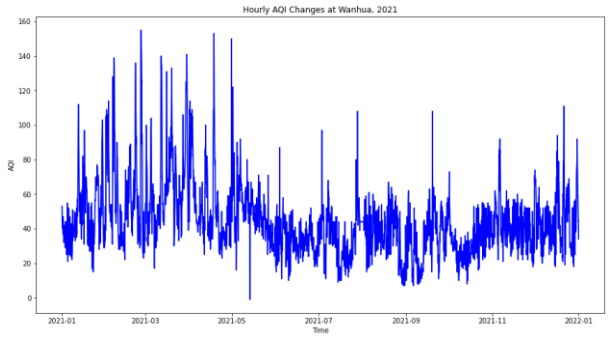
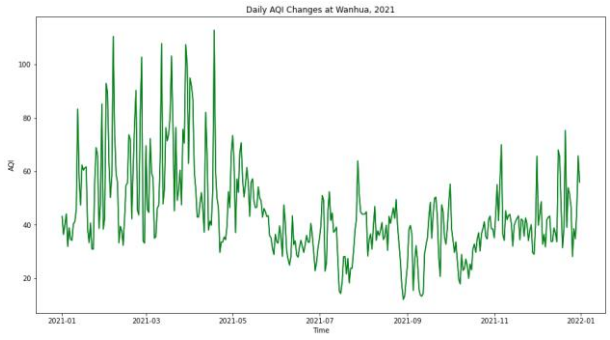
3. Time Series Analyzing

Comparing hourly and daily series

Daily series (365, 2) are generated by the Hourly series (8560, 2), grouping hourly data in the same date.

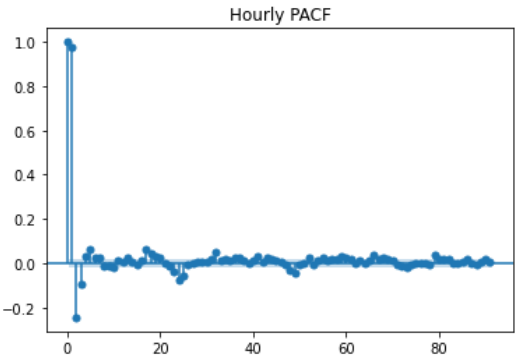
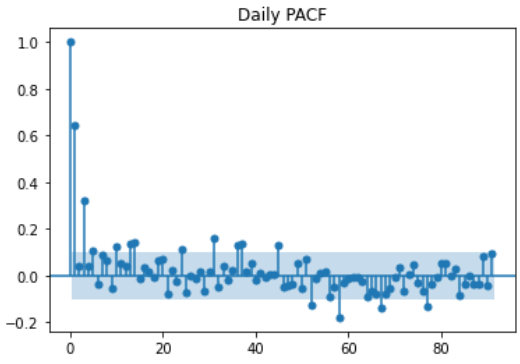
4. Series visualization & checking stationarity

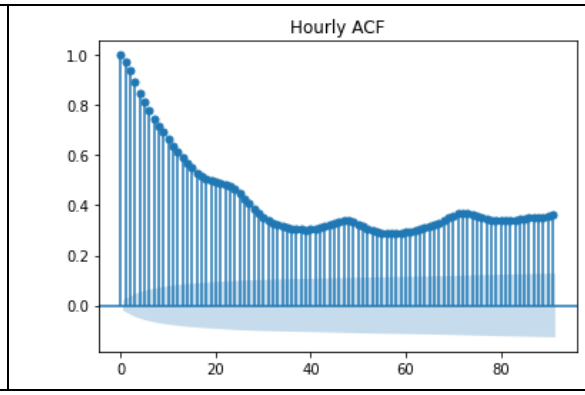
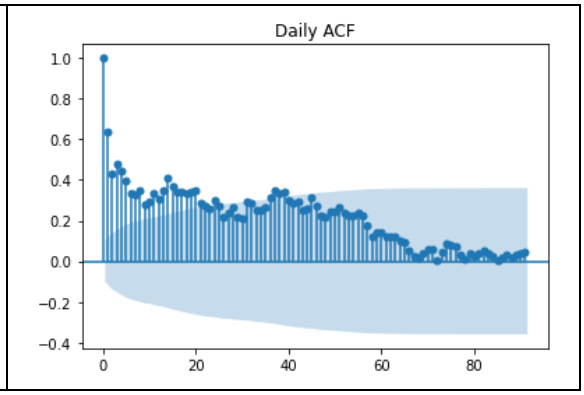
Then we plot the AQI with time changes to observe the trend, and doing ADF test to check the seasonality.

Hourly AQI changes	Daily AQI changes
	

5. Check the stationarity, ACF, PACF of the series

First we can check the stationarity of the series by the ADF test to find d, and then see the ACF, PACF plots to narrow down the searching area of p and q.

	Hourly ACF, PACF	Daily ACF, PACF
PACF		

ACF		

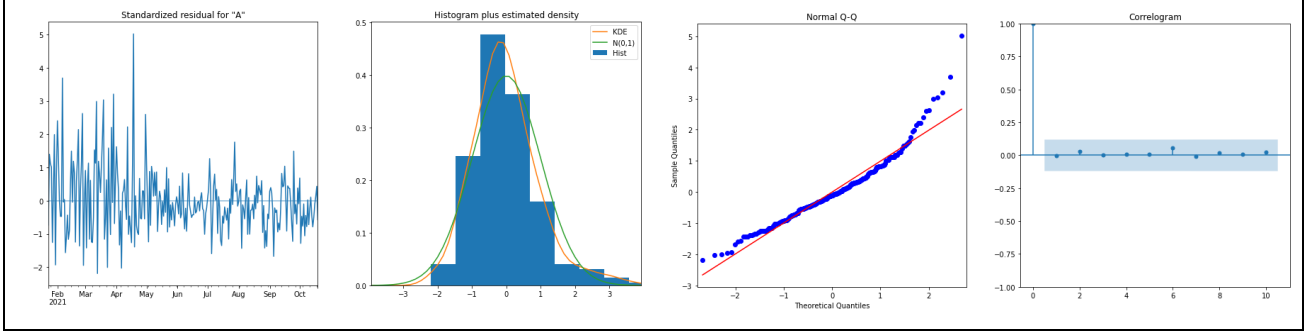
We can clearly observe that the ACF of the Hourly data keeps having significance along all the lags, in other words, q of hourly data is above all these lags. And from now on we can concentrate on discuss the daily data only, for it is the 24 sample mean of the hourly data.

Overserving the "cut of" of the daily PACF and ACF, we set the searching range of p [0, 2] and the searching range of q [0, 22].

6. Finding proper ARIMA model by AIC

Enumerate all possible ARIMA parameter combination within our searching range, to find out the smallest AIC, and define the fittest ARIMA model for the series. And our optimal parameter for ARIMA model is (p, d, q) = (0, 0, 21) with AIC = 2150.62.

Model diagnosis



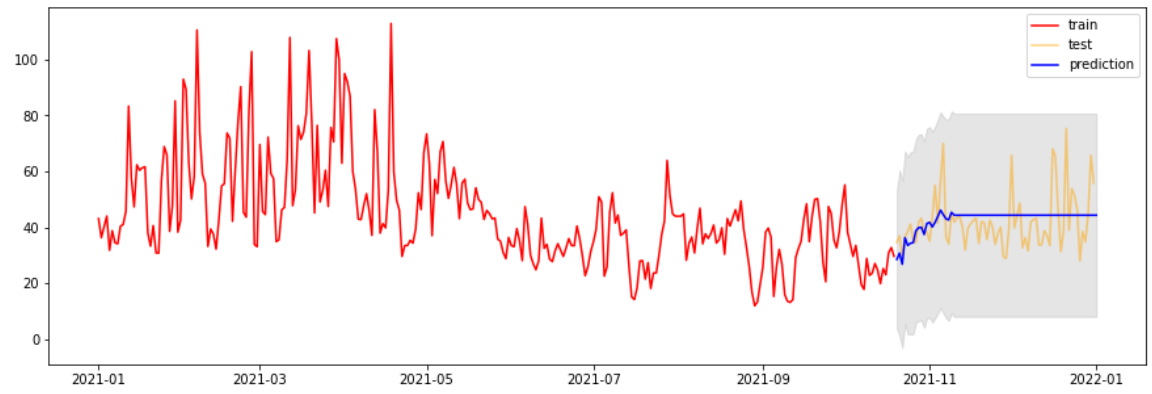
We can observe that the residual varies much from JAN to MAY, and then it start shrink till the end of the 2021, in other words, model performs better between MAY to DEC.

What's more, according to Q-Q plot of the residual, we can see that the center quantiles part follows the normal distribution better than the last quantiles part.

7. Forecast by the fitted SARIMA model

After fitting the model with the first 80% of data, we now try to forecast the last 20% of data, to check the model's ability in predicting future AQI data.

Prediction results



Discovery and conclusion

The real world is not as simple as the text book taught. Although we may see the CO2 changing following a certain trend in Canada yearly under a large scale, it's difficult to find a simple model to fit the AQI changes in both small-time scale and small location. Might be disturbance by the much human being activities, or not enough time span to observe.

Although we manage to fit an ARIMA (0, 0, 21) model to the data, the AIC is not as good as those models we generate for the homework. Perhaps we should try more parameters settings, or using SARIMA to fit the data after determining a seasonal factor, or consider other time-series model to deal with this dataset.