AUTOREGRESSIVE MODELS FOR AIR QUALITY INVESTIGATION

O. Zalieska[1] **and H. Yailymova**[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
[2] National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine
e-mail: zalieskaolena@knu.ua, yailymova.hanna@lll.kpi.ua

Abstract

The aim of the work is to build a forecast of air quality in Kyiv for some period of time. For this purpose we preprocessed and analized data, selected and fitted a model.

Keywords: machine learning, autoregression, air quality, time series

1. Introduction

The development of technology, increased production of certain products contribute to higher emissions of harmful substances into the air. Air pollution causes climate change, increases the number of people suffering from heart and respiratory diseases etc. Therefore, air quality in Kyiv is monitored in order to minimize possible negative consequences.

2. Prediction using the SARIMA model

2.1. Problem Statement

PM (particulate matter) - small particles that are air pollutants (dust, dirt, smoke, etc.) distinguished by diameter (PM1, PM10, PM2.5). The Air Quality Index (AQI) provides information about air pollution [1]. Usually the AQI is calculated for indicator PM2.5, because, as stated in [3], it is the most dangerous pollutant. Therefore, PM2.5 can be considered a target.

Table 1: 3 rows of the data

logged at	pm25
2020-12-01 00:08:10	0.518428
2020-12-01 00:13:04	0.729244
2020-12-01 00:16:25	0.770710

The data is taken from [2]. The dataset contains values: phenomenon - the measured indicator; value (of the indicator); logged_at - the exact time when the measurements were taken. Let’s try to identify any dependencies between PM2.5 feature and time. Table 1 shows the data after pre-processing.

Let’s look on the average value of the PM2.5 feature by hour (Figure 1).The average at about 2-3 pm is the lowest, the highest value is at 7 am, another peak occurs at 10 pm. This is probably due to the increase in the number of cars during the hours when people go to or from work.

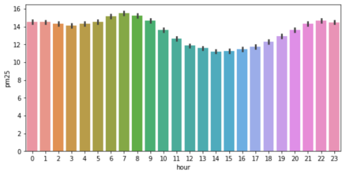

Figure 1: Average PM2.5 by hour

2.2. Time Series Research, Choosing and Fitting Model

We will consider the averaged values of the PM2.5 column for every 2 hours as a time series. We can find outliers by decomposing the series into trend, seasonality and error as those values that deviate significantly from the combined seasonal component and trend, i.e., there is a large error (with sigma rule).

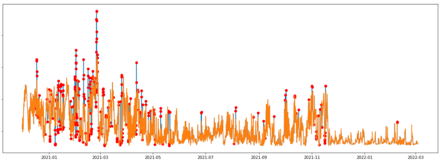

Figure 2: time series and outliers

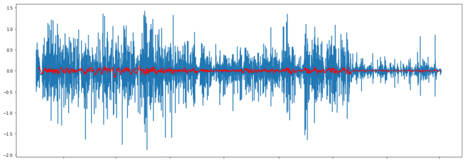

Figure 3: time series after the first differentiation

The figure 2 shows the points that were identified as outliers, the series before removing these points is colored blue, after removing - orange. Since there is a nonlinear trend, the series is not stationary. After trying to make it stationary by taking differences we get the series shown on Figure 3.

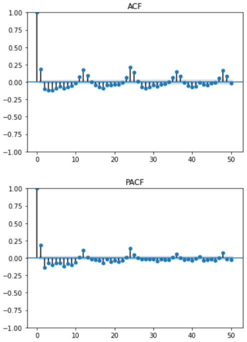

Figure 4: Values of ACF and PACF for the series after the first differentiation

Using ACF and PACF (Figure 4) we can see that there's a seasonality (a large correlation at the value of 12). This means that the data is correlated with what happened 24 hours ago. At 48, 72, 96 hours, the correlation decreases, but still remains quite high. So we need to get rid of seasonality.

For this series, the hypothesis in the Dickey-Fuller test is not rejected, so the series is indeed stationary.

One approach to time series forecasting is to use autoregressive and moving average models, as well as their modifications. The SARIMA (Seasonal autoregressive integrated moving average) model is used if there are both seasonality and trend in a time series.

So, we will use SARIMA ((p, 1, q)(P, 1, Q), 12) to predict the values of the series.

We will build the model on PM2.5 values until 2021-07-01 4pm and try to build a forecast for 30 hours ahead. We will search through the possible model parameters and determine the best ones using the Python and R built-in functions. The model defined in this way is SARIMA ((5,1,0), (2,1,0), 12). Let's build a forecast and display the predicted and real data (Figure 5).

The mean square error (MSE) is about 0.0967 on the test sample, and 0.1038 on the training sample. That is, the model made a fairly good prediction.

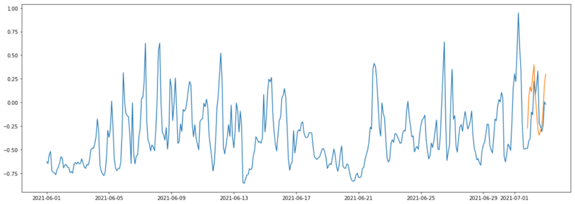

Figure 5: Prediction of PM2.5 for 30 hours

.

Conclusions

The data from one of the air quality monitoring stations in Kyiv was chosen. For this dataset the time series with the values of the concentration of PM2.5 pollutants in the atmosphere was analyzed. Based on this analysis, an autoregressive model was chosen to predict the values of this indicator in the future.

The SARIMA model was used to make a 30-hour forecast of PM2.5 in the atmosphere. The model made a good prediction, but there are other approaches and ways to improve the model (adaptive methods for building autoregressive models, neural networks, etc.)

References

1. *Beijing Air Pollution: Real-time Air Quality Index (AQI)*- https://aqicn.org,

3. *SaveEcoBot* - https://www.saveecobot.com,

3. *Undark - The Weight of Numbers: Air Pollution and PM2.5* -https://undark.org/breathtaking,