Scientific Research Publishing

Atmospheric and Climate Sciences, 2022, 12, 728-749
https://www.scirp.org/journal/acs
ISSN Online: 2160-0422
ISSN Print: 2160-0414

# Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air Pollutants in Zahleh, Lebanon

**Alya Atoui[1,2], Kamal Slim[3], Samir Abbad Andaloussi[1], Régis Moilleron[1], Zaher Khraibani[2,4]**

[1]Leesu, Univ Paris Est Creteil, Creteil, France
[2]Rammal Rammal Laboratory, PhyToxE Group, Lebanese University, Nabatieh, Lebanon
[3]Lebanese Atomic Energy Commission, National Council for Scientific Research (CNRS), Beirut, Lebanon
[4]Department of Applied Mathematics, Faculty of Sciences, Lebanese University, Hadat, Lebanon
Email: alyaatoui@gmail.com

**How to cite this paper:** Atoui, A., Slim, K., Andaloussi, S.A., Moilleron, R. and Khraibani, Z. (2022) Time Series Analysis and Forecasting of the Air Quality Index of Atmospheric Air Pollutants in Zahleh, Lebanon. *Atmospheric and Climate Sciences*, **12**, 728-749.
https://doi.org/10.4236/acs.2022.124040

**Received:** September 2, 2022
**Accepted:** October 28, 2022
**Published:** October 31, 2022



## Abstract

During the last decades, air pollution has become a serious environmental hazard. Its impact on public health and safety, as well as on the ecosystem, has been dramatic. Forecasting the levels of air pollution to maintain the climatic conditions and environmental protection becomes crucial for government authorities to develop strategies for the prevention of pollution. This study aims to evaluate the atmospheric air pollution of the city of Zahleh located in the geographic zone of Bekaa. The study aims to determine a relationship between variations in ambient particulate concentrations during a short time. The data was collected from June 2017 to June 2018. In order to predict the Air Quality Index (AQI), Naïve, Exponential Smoothing, TBATS (a forecasting method to model time series data), and Seasonal Autoregressive Integrated Moving Average (SARIMA) models were implemented. The performance of these models for predicting air quality is measured using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), and the Relative Error (RE). SARIMA model is the most accurate in prediction of AQI (RMSE = 38.04, MAE = 22.52 and RE = 0.16). The results reveal that SARIMA can be applied to cities like Zahleh to assess the level of air pollution and to prevent harmful impacts on health. Furthermore, the authorities responsible for controlling the air quality may use this model to measure the level of air pollution in the nearest future and establish a mechanism to identify the high peaks of air pollution.

## Keywords

Air Pollution, Air Quality Index, Times Series, Prediction

## 1. Introduction

Air pollution is one of the most important public health problems, specifically in urban areas where the majority of industries and traffic circulation. It has impacts on human health as well as on the environment. Nowadays, air pollution does not threaten only humanity, but also Earth's ecosystem [1]. Air pollution is a growing public health problem, and mortality due to air pollution is expected to double by 2050 [2]. According to recent estimates reported by the World Health Organization (WHO), more than 90% of the world's population is exposed to unhealthy air quality, which exceeds the guideline limits, which is the main cause of high incidence of premature mortality and morbidity. In addition, each year, 4.6 million people die from causes related directly to air pollution [3]. Health problems increase depending on the time of exposure and the type of pollutants. Hence, we can define several types of pollutants which can be classified into two groups based on their sources: primary and secondary. The primary category covers the pollutants which are directly transmitted from an atmospheric source of air pollutants. The secondary group consists of the pollutants which are formed from the chemical reaction of the primary pollutants within the atmosphere [4]. The Middle East region is characterized by specific weather, hot and dry in summer and moderate and humid in winter. This region is composed of an almost surrounded sea where the air constitutes a crossroads of pollution and natural emissions [5]. The Mediterranean basin is especially sensitive to atmospheric air quality due to cloudless weather and intense summer sunshine. Since Lebanon belongs to the MENA region, the assessment of air quality in this country has been of special importance in recent years. Lebanon is a small country in the Middle East, located at approximately 34˚N, 35˚E surface of 10,452 $km^2$, and its widest point is 88 kilometers, its narrowest is 32 kilometers; the average width is about 56 kilometers [6]. Air pollution in this country is very high, in proportion to its area and number of inhabitants. The increase in signs of pollution along with the number of people suffering from respiratory illness made it very important to monitor and control air pollution. According to the World Health Organization, the air quality in Lebanon is considered very dangerous to the ecosystem. Recently the annual average concentration of $PM_{2.5}$ in the country is 31 g|$m^3$, which exceeds the recommended maximum of 10 g|$m^3$. Several factors that contribute to poor air quality in Lebanon include cement industries, food processing, minerals and chemicals, petroleum refining, and vehicle emissions. Note that there are seasonal variations in Lebanon with the highest levels of air pollution in winter (December to March) due to heating. The most polluted cities in Lebanon are Baalbak, Beirut, Saida, and Zahleh have high levels of air pollution. Zahleh is a heavily populated city that is affected by several sources of air pollution. Several pollutants were identified that have constantly increasing levels in the atmosphere such as the road transport sector [7], and an unregulated private diesel generator sector [8]. In addition, the significant crisis in

solid waste treatment has increased the concentration of pollutants in the air, as well as the risks of short-term cancer and respiratory diseases [9]. Several works were carried out to study the air quality in the capital Beirut, these studies showed serious air pollution which has an impact on the health of the citizens. However, there are no recent studies to assess the air quality in the Bekaa region, especially since this region differs from the capital in geographical characteristics, its rivers, and its valleys. In this article, we are interested in modeling pollution in Zahleh, Lebanon, where we will be focusing on three main objectives; identifying the relationship between the parameters of air pollution and the meteorological elements, calculating the Air Quality Index (AQI), and making an air quality forecasting to predict future levels of pollution. Previously, time series analysis has been used to study the daily average concentration of air pollutants, compare the fluctuation of time series, and investigate the variation of air pollutants over time on a weekly, monthly, and yearly basis in Lebanon [10]. This method was also used worldwide in big cities in very crowded countries such as India [11] [12] [13]. The time series method was helpful in showing the variation of the daily and seasonal averages in order to identify the time of the year, and of the day, where some parameters have peak levels [14]. Also, the correlation between air pollution parameters and the meteorological elements was investigated many times in the literature, using different models such as the Random Forest regression [15], simple linear regression [16], and multi-linear and nonlinear regressions [17]. Findings were different in each study area and for different parameters, where some parameters showed a highly negative correlation with meteorological elements, and others were affected positively by these elements. Many statistical approaches are used to study air pollution and its health effects in order to analyze air quality and predict future values [12]. Note that descriptive statistics are limited in terms of understanding behavior and air quality variability [18]. In addition to probability models [19] and studying the temporal distribution of air pollutants, other studies have relied on time series analysis, which facilitates a better understanding of the cause-and-effect relationship in environmental pollution [20]. The ARIMA and SARIMA models were found very useful in different studies conducted in Tehran and Peninsular Malaysia respectively [21]. Then, combined with the ARIMA model, the principal component regression was chosen as the best model to calculate and forecast the future Air Quality Index (AQI) in Delhi-India [22].

This article is organized into several sections, firstly, the description of a monitoring station and the parameters of pollutants illustrate the evolution of pollution in this region. Secondly, a description is given regarding the statistical method, the linear model, and the performance of the model to model and forecast air quality for the near future. Then, in the next section, we present results, starting with descriptive statistics and graphic presentation per hour/week/month in order to assess the variation of atmospheric pollutants. Afterward, to investigate the correlation between different measurements, we compute the coefficient of correla-

tion for two different seasons. Furthermore, the calculation of the Air Quality Index (AQI) through a general formula that presents the pollutant parameter to correctly forecast atmospheric pollution and compare several forecast models together with selecting the SARIMA model as the most suitable one according to the selection guideline provided. In the end, an overall summary is presented whose aim is to evaluate the levels of air pollution to prevent harmful impacts on health.

## 2. Study Area

Monitoring of air quality in Lebanon dates back to 2013 from the installation of five monitoring stations as part of the Environmental Resources Monitoring Project in Lebanon (ERML). These stations use online analyzers connected to a control and Data Acquisition System (DAS) located at the MoE. All the analyzers installed within the air quality monitoring stations (AQMS) are based on reference methods meeting the requirements of the European directive on air quality 2008/50/EC. The monitoring stations were located in Memchiyeh garden with coordinates of 33˚59'52.44"N and 36˚12'16.43"E. Zahleh is the largest city in Bekaa and the fourth largest city in Lebanon with an area equal to 8 $km^2$. It is located 54 km to the east of the capital Beirut. It is located on the Eastern foothills of Sannine Mountain and surrounded by the Lebanese mountains and the Bekaa valley [10].

## 3. Data Monitoring

In this paper, we focus on the more prevalent atmospheric pollutants: Nitrogen oxides ($NO_x$), which contain nitrogen, and oxygen in varying levels, such as nitric oxide (NO) and nitrogen dioxide ($NO_2$), carbon monoxide (CO), sulfur dioxide ($SO_2$), ozone ($O_3$), and particulate matter (PM) of 10 and 2.5 microns in diameter. The pollutant concentrations are compared to the WHO Guidelines that are summarized in Table 1.

**Table 1.** WHO guidelines for air pollutants.

| Pollutant | Duration of exposure | WHO Guidelines |
|---|---|---|
| $PM_{2.5}$ | 1 year | 35 μg\|$m^3$ |
| $PM_{10}$ | 1 year | 20 μg\|$m^3$ |
| $NO_2$ | 1 year | 40 μg\|$m^3$ |
| $O_3$ | 8 hours | 100 μg\|$m^3$ |
| $SO_2$ | 24 hours | 20 μg\|$m^3$ |

The pollutant data is registered by the Department of Air Quality, Ministry of Environment, Lebanon, which is monitored at Memchiyeh Garden Station located in Zahleh-Bekaa region. The data were hourly observed from June 2017 to June 2018 (13,895 recorded observations). The variables available were NO, $NO_2$, $NO_x$, CO, $SO_2$, $PM_{2.5}$, $PM_{10}$, *Temperature*, *Humidity*, *Wind Direction*,

*Wind Speed* at Zahleh. All pollutants were measured in μg|m$^3$ except for CO which was measured in mg|m$^3$. The missing values were observed to be filled by various statistical approaches, such as replacing them with the neighbor's most similar value, daily mean/median/maximum, however, the data that was collected in Memchiyeh station, encounters missing values consecutively. Therefore, the Multivariate Imputation by Chained Equation (MICE) algorithm was suggested as the most suitable method to impute these missing values [23]. The MICE is a package **R**, which assumes that data are randomly missing. It prepares multiple values to be replaced by the missing values by designing an appropriate model, such as regression. Each variable is considered a dependent variable to be treated in this approach. The MICE process involves the prediction of the missing values for the selected variables by using the available data of the other variables. Then the missing values will be replaced by the predicted values to make a new data set called imputed data set, and through iterations, multiple imputed data sets are generated [24]. There are 22,485 missing values from a total of 152,895 observations (about 15% of the data are missed) in Memchiyeh station [25]. The detection of outliers is not important in time series analysis, but error detection is significant for analysis. Pollutants can have a high level due to numerous factors that serve the purpose of this study. However, for meteorological variables, there are some outliers that were considered to be seasonal errors. The errors detected have been replaced with the median of the time period they are included.

## 4. Methodology

The initial step prior to applying any models is to treat the raw data. For this purpose, we used the Multivariate Imputation by Chained Equation (MICE) algorithm to fill in the missing data, which was an **R** package that assumes data are missing randomly. There were 22,485 missing values out of a total of 152,895 observations (about 15 percent of the data are missing). MICE prepares several values to replace with the missing values through a suitable model design such as regression and then replaces the missing data with predicted data to make a new dataset called imputed dataset, and through iteration, multiply imputed datasets will be generated [11]. Finally, each dataset will be analyzed. Subsequently, the outliers were detected and replaced by the median of the period in which they are included. This is especially important because, in meteorological parameters, outliers are considered errors depending on the season. It is advisable to use numerical values for data aggregation (e.g. calculation of daily mean values or monthly mean values). We selected the daily average values for the calculation. Once the raw data was processed, R software was used for data analysis.

### Statistical Models

The goal of the statistical approach is to develop the most accurate model of the phenomenon, not based on the information that we know about it but on information that can be derived from a dataset describing this phenomenon. Therefore,

there is no necessity to formalize the operation of the phenomenon being studied and we do not have biases caused by imprecision's in the models we develop. However, we need to have a data set allowing the model to adequately represent the studied phenomenon, and we need to be able to exploit the information present in this data set. Furthermore, the inaccuracy of the data used will affect the forecast. An important aspect of modeling air pollution is statistical analysis, which involves the prediction of the future development of a data set from the observed data set up to the current observation time. To forecast univariate time series, we use Naïve, TBATS, exponential smoothing, and SARIMA models.

### Performance Measures

The objective of the assessment is to evaluate the ability of the model to perform well. Several measures are used to validate the regression model. Firstly, we focus on the computation of the Mean Absolute Error (MAE) that uses the absolute value of the difference between the observed and the predicted, thus measuring systematic error and random error. The MAE can be used to evaluate the efficiency of the model.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

Secondly, the Root Mean Square Error (RMSE) is computed by using the difference between the observed and the predicted rather than absolute values. As a result, this index gives more bias towards greater differences between observed and predicted, which is desirable for air quality prediction purposes as peak levels of pollution are most important for forecasting.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$$

It is preferred to provide RMSE results as its square root, the mean square error (MSE), which has the same information but has the same magnitude as the predicted variable. RMSE is most widely used due to its legibility and accurate precision estimate.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}$$

Finally, the Relative Error (RE) is a precision measure, which describes the accuracy of the measurement. RE identifies the relative precision between two or more measurements.

$$\mathrm{RE} = \frac{\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|}{\sum_{i=1}^{n} y_i} \times 100$$

## 5. Results and Discussion

Following data filtration, the daily averages were computed as well as descriptive statistics of the measured values in the monitoring station. The following step was to make hourly, weekly, and monthly variation graphs using temporal cor-

relation. Numerical values are recommended for data aggregation (e.g. calculation of daily average values or monthly average values). Here, daily average values have been selected to be computed. Once the pollutants data were processed, R software was used to perform the data analysis. The observations were taken from June 1, 2017 to December 31, 2018 in Zahleh. In Figure 1 and Figure 2, we represent some graphical descriptions.

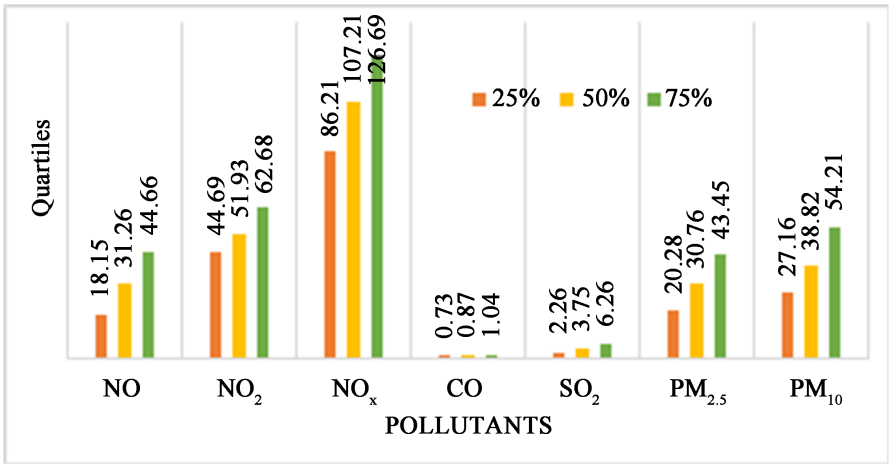


Figure 1. $Q_1$ (25%), $Q_2$ (50%), $Q_3$ (75%) for the pollutants in Memchiyeh station.

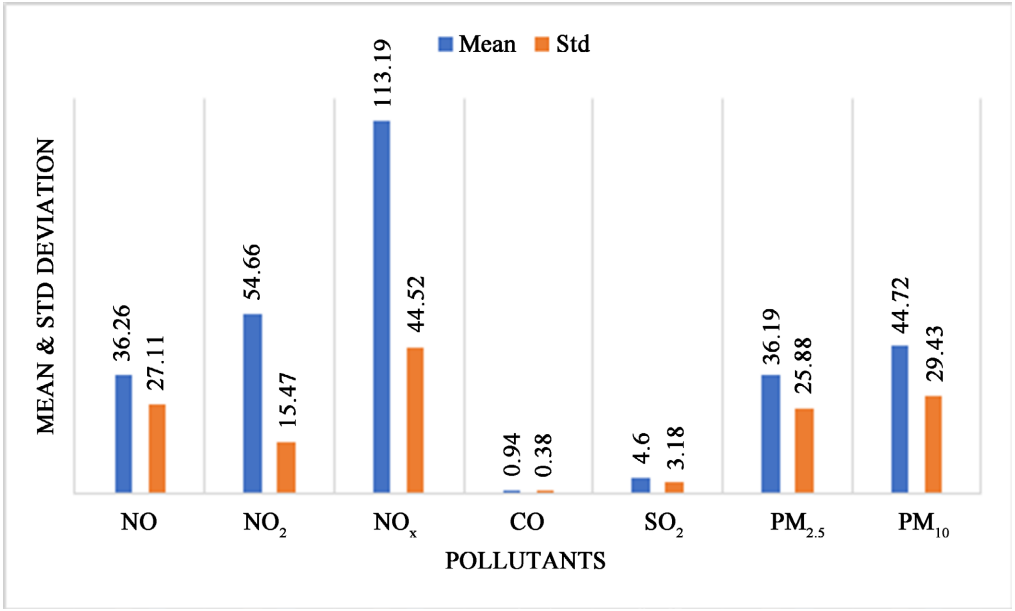


Figure 2. Mean and standard deviation for the air pollutants in the Memchiyeh monitoring station.

## 5.1. Hourly Variation

The hourly variation results indicated that there were some times of the day when the concentration of certain air pollution reached its peak, whereas others have no significant variation throughout the day. As an example, NO, $NO_2$, $NO_x$, and CO have two peak time periods, the first recorded between 6 and 10 AM, while the second recorded between 4 PM and 1 AM. The other parameters are

almost identical during the day. Figures 3-6 illustrate the hourly variation of air pollutants in Zahleh.

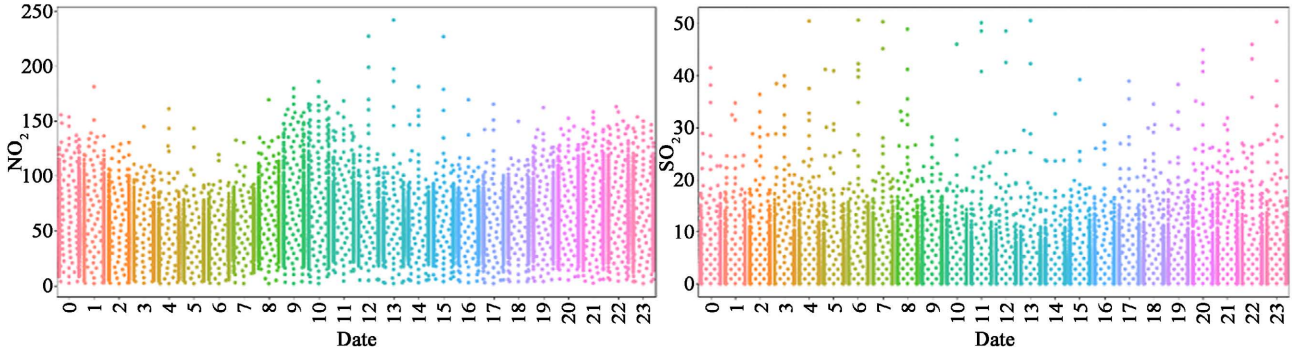


Figure 3. Hourly variation of $SO_2$ (left) and $NO_2$ (right).

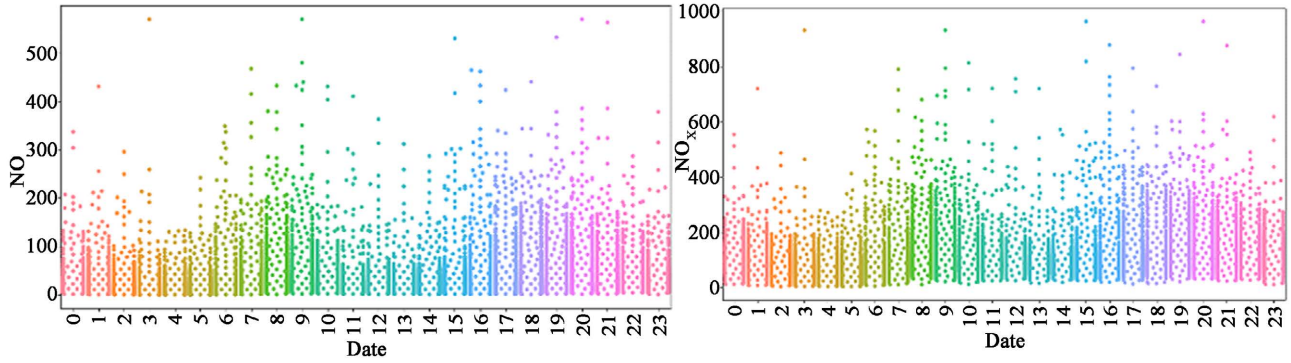


Figure 4. Hourly variation of NO (left) and $NO_x$ (right).

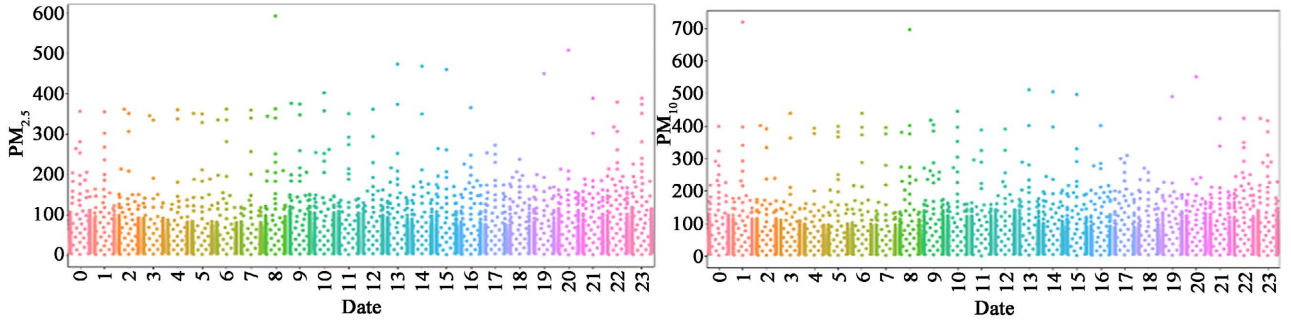


Figure 5. Hourly variation of $PM_{10}$ (left) and $PM_{2.5}$ (right).

Figures 3-6 indicate that the concentrations of NO, $NO_2$, $NO_x$, and CO have their first peaks between 6 and 10 AM. The second highest peaks are detected between 16:00 and 1:00 AM. $SO_2$, $PM_{2.5}$, and $PM_{10}$ pollutants do not exhibit any remarkable hourly variation. High daytime levels are correlated with increasing traffic and urban activities, and night-time concentrations are coincident with an increase in the number of recreation vehicles and outdoor activities. A previous study indicated that peak traffic hours occur between 7 and 11 AM and between 5 and 7 PM from Monday through Friday, although traffic remains relatively high throughout each day [26].

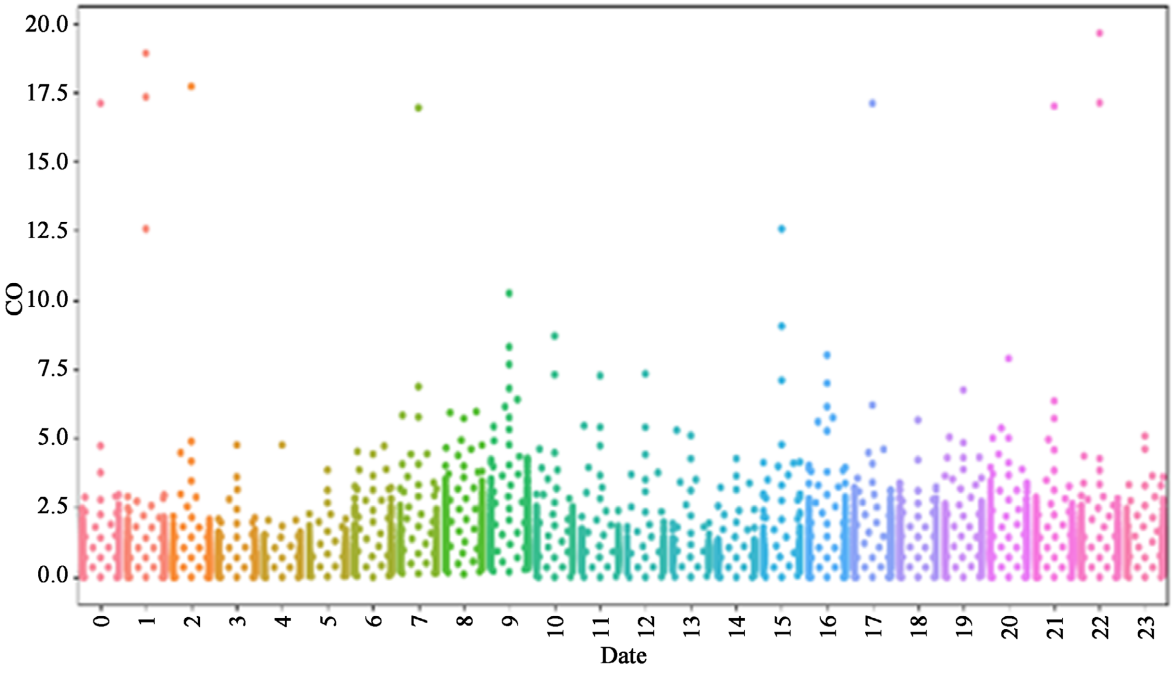


**Figure 6.** Hourly variation of CO.

## 5.2. Weekly Variation

This subsection investigates the weekly variation of the pollutants to identify the dependency of their concentration on their source activities which may affect the weekly atmospheric conditions in the Bekaa region, particularly in Zahleh (**Figures 7-10**).

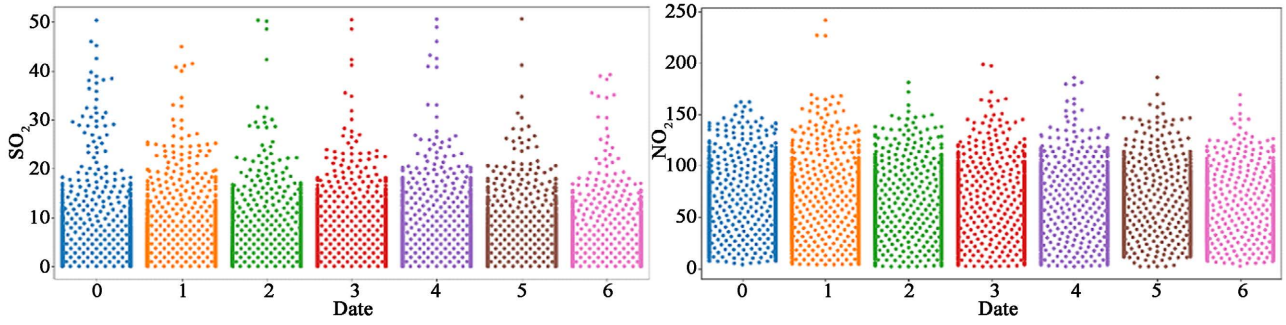


**Figure 7.** Weekly variation of $SO_2$ and $NO_2$.

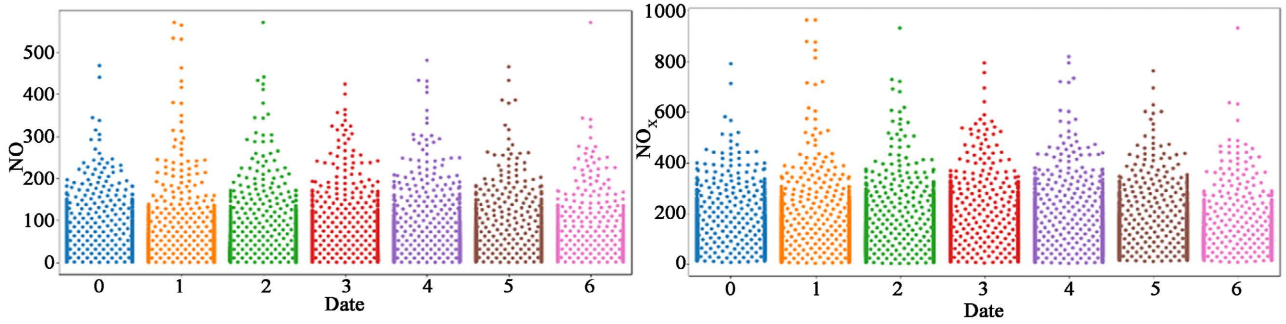


**Figure 8.** Weekly variation of NO and $NO_x$.

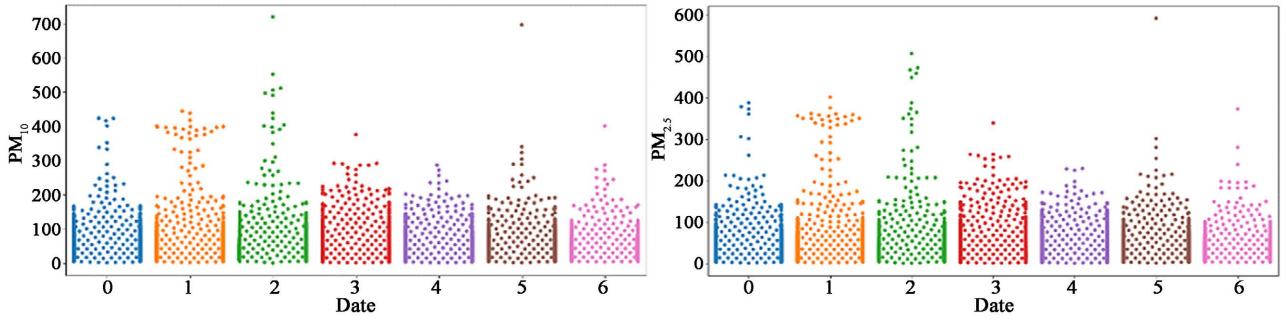


**Figure 9.** Weekly variation of $PM_{10}$ and $PM_{2.5}$.

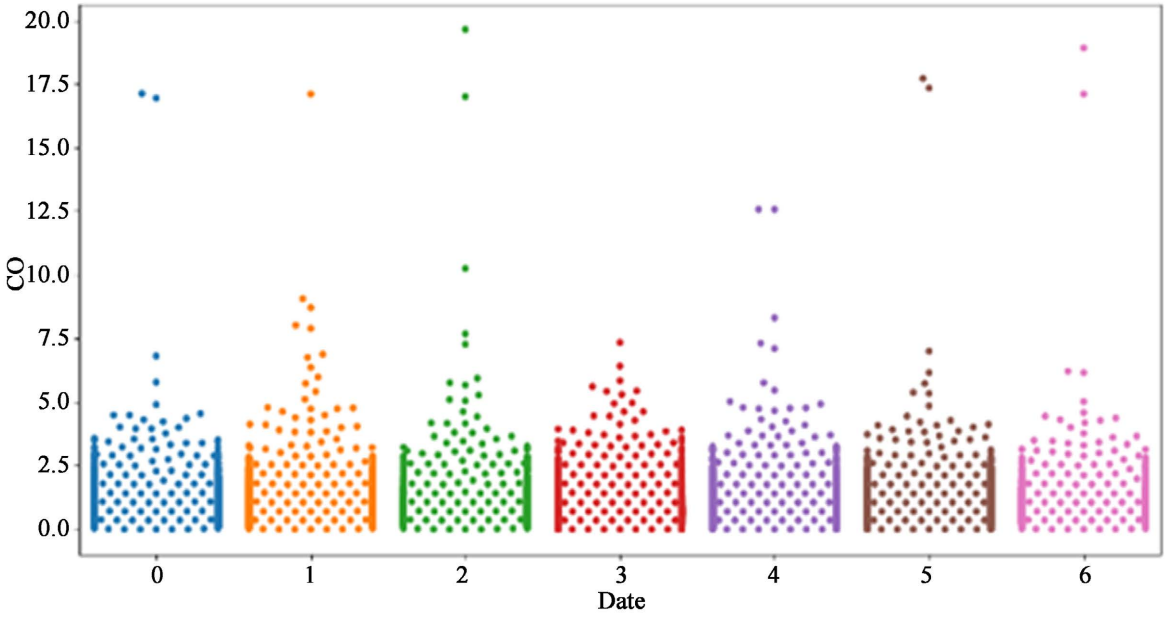


**Figure 10.** Weekly variation of CO.

Regarding weekly variation, we noticed no significant changes in pollutant levels which remain high throughout the entire week as the city of Zahleh is highly populated and has very frequented roads and driveways during the weekends as well as the weekdays. The principal reason for the absence of weekly variation is that concentrations of pollutants change quickly in urban areas and persist for several days before they disappear. Nevertheless, Zahleh is heavily inhabited. That led to a variety of issues, particularly in the transportation sector, where most transportation methods used in the area were private vehicles, and almost complete absence of any public transport mode, which caused a noticeable increase in pollution in this area.

### 5.3. Monthly Variation

Monthly graphs indicate that the concentrations of NO, $NO_x$, $SO_2$, and CO increase in the winter and decrease in the summer. In contrast, $NO_2$ concentration is highest in the summer. However, $PM_{2.5}$ and $PM_{10}$ have no remarkable variation over the year (Figures 11-14).

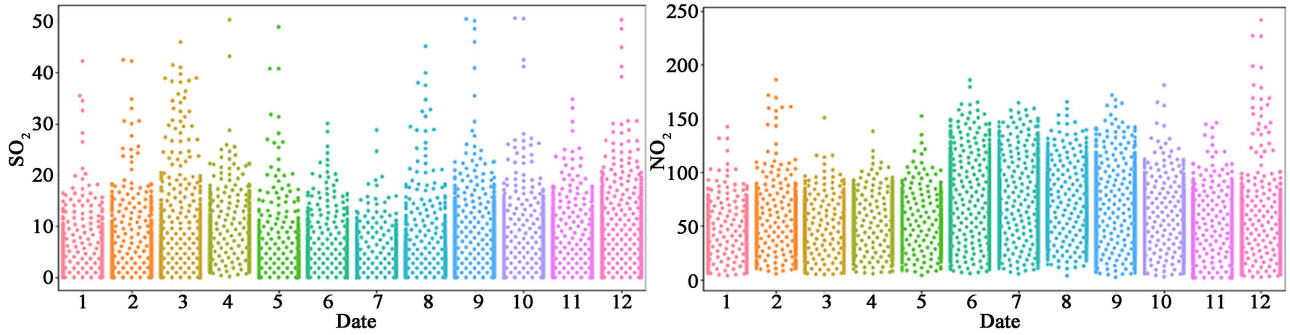


**Figure 11.** Monthly variation of $SO_2$ and $NO_2$.

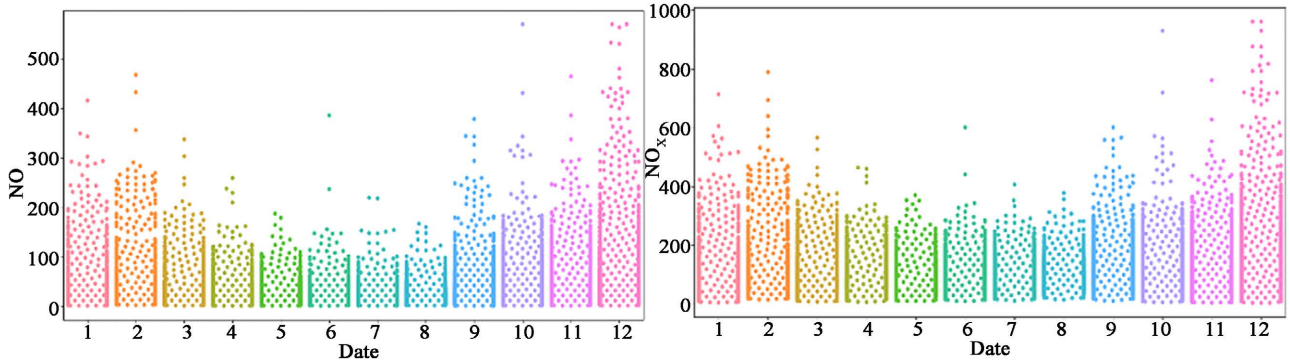


**Figure 12.** Monthly variation of NO and $NO_x$.

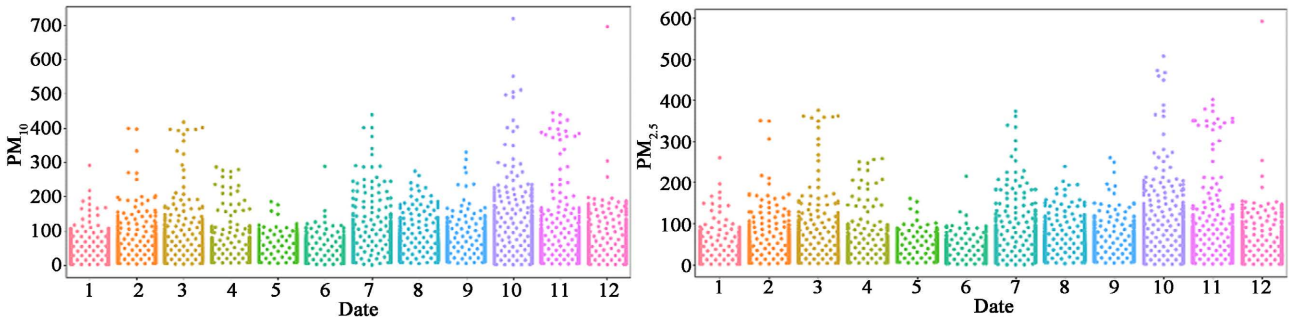


**Figure 13.** Monthly variation of $PM_{10}$ and $PM_{2.5}$.

High emissions of $SO_2$ during the winter may be caused by some factors including the development of a reversal atmosphere and delayed wintertime oxidation of $SO_2$. These pollutants are emitted from diesel vehicles and from central combustion engines. These operations are increased during the winter season, which causes an increase in the levels of pollutants during that season. The traffic levels increase during the winter season and reduce during the summer season as a result of the schools, the universities, and certain work activities that occur seasonally. Whereas during summer, activities of the schools and the universities come to a halt; furthermore, numerous people move out of the country for vacation. Also, heating equipment is not used during summer. Those factors reduce pollutant emissions. Pollutants that do not vary monthly are caused by their sources which do not vary throughout the year.

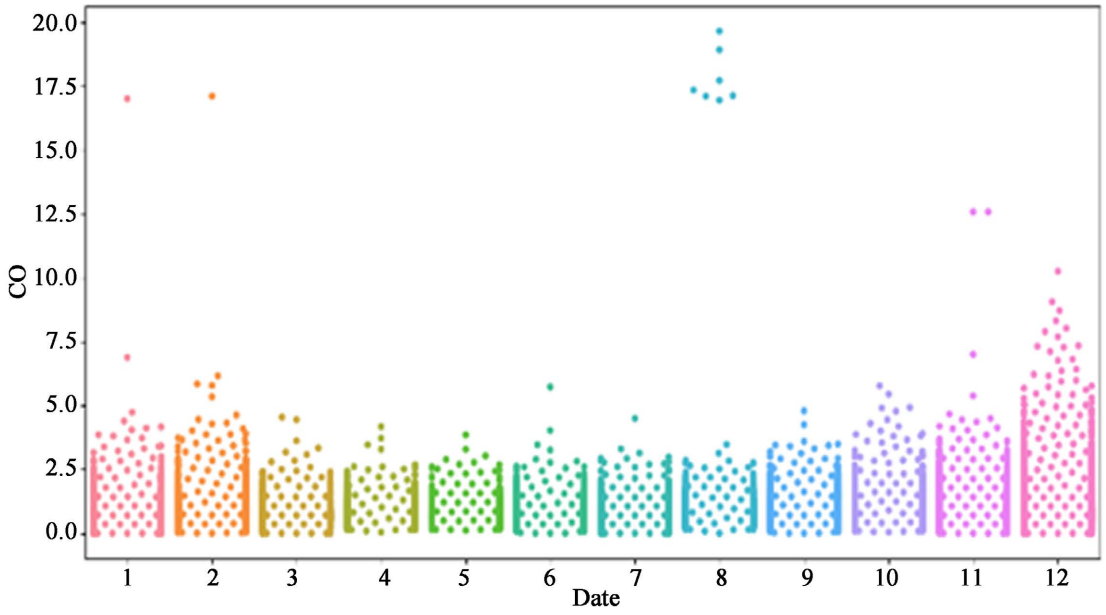


**Figure 14.** Monthly variation of CO.

## 5.4. Environmental Factor Variation

Considering each parameter independently, using the descriptive statistics, it can be noted that the concentrations of NO, $NO_2$, $NO_x$, CO, $PM_{2.5}$, and $PM_{10}$ are extremely high and that $NO_2$, $PM_{2.5}$, and $PM_{10}$ are above the guidelines, whereas the $SO_2$ concentration is almost certainly not above the limits. It is also to be noted that $NO_2$ concentration is highest during the summer season. In order to

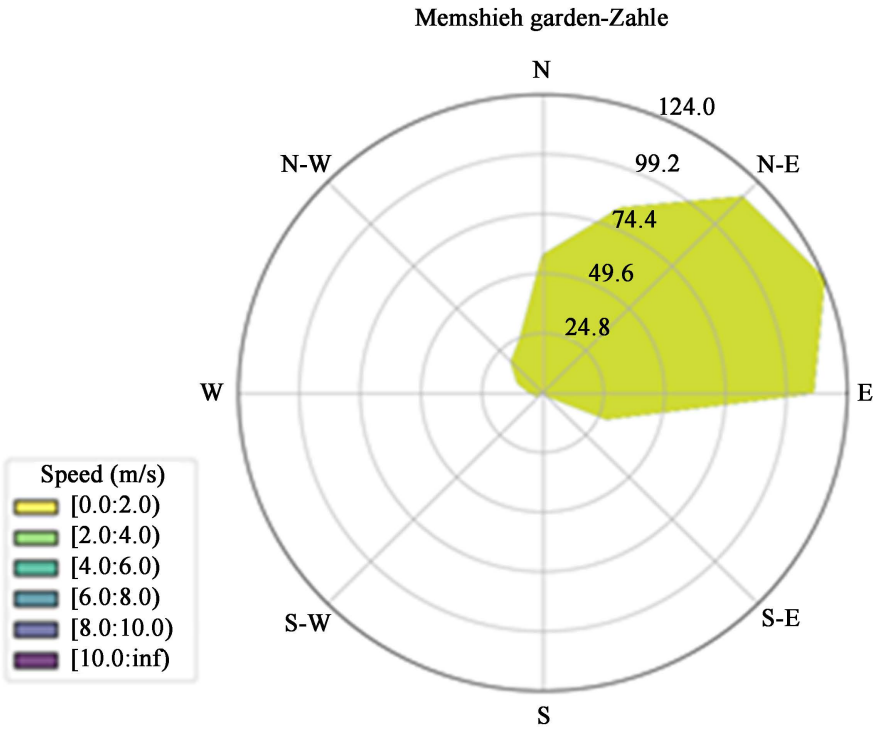


**Figure 15.** Observed wind speed and direction at Zahleh 06-2017/2018.

justify these levels, it should be mentioned the presence of an electric power station in Zahleh, 5 km away from the monitoring station, which causes 40% of the $NO_x$ emissions. Regarding the elevated CO levels, these are the consequence of heating for long periods, since the temperature is usually low and significantly decreases during winter in this city.

One further factor which can affect the level of pollution is the wind direction, as can be shown in Figure 15, where the wind is directed to the northeast, which means that it is moved towards Zahleh, and as this city is located between the Lebanese mountain chains to the east and the west, rather than being an entirely wide area, the pollution transmitted by the wind from other cities can be captured at Zahleh for longer time periods, which explains the high pollution levels around the year. Furthermore, the temperature in Zahleh is much warmer than in other cities like Beirut all year round, leading to greater usage of gas heaters, fireplaces, and ovens in Zahleh, thus leading to increased CO levels.

### 5.5. Pollutants Concentration Correlation

The Pearson seasonal correlation was used to identify a correlation between the air pollution parameters and the meteorological components. Firstly, the data were divided into two seasons: the summer season including the months from March to August, and the winter season including the months from September to February. Results are shown in Table 2 and Table 3.

Table 2. Winter correlation matrix in Zahleh.

| | NO | $NO_2$ | $NO_x$ | CO | $SO_2$ | $PM_{2.5}$ | $PM_{10}$ | Temp | Hum | WD | WS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NO | 1 | 0.62 | 0.97 | 0.89 | 0.56 | 0.21 | 0.23 | −0.20 | −0.07 | 0.29 | 0.00 |
| $NO_2$ | | 1 | 0.53 | 0.75 | 0.62 | 0.27 | 0.29 | 0.46 | −0.59 | 0.29 | −0.16 |
| $NO_x$ | | | 1 | 0.89 | 0.55 | 0.25 | 0.27 | −0.08 | −0.20 | 0.32 | −0.03 |
| CO | | | | 1 | 0.59 | 0.25 | 0.26 | −0.12 | −0.18 | 0.33 | −0.02 |
| $SO_2$ | | | | | 1 | 0.29 | 0.31 | 0.32 | −0.57 | 0.41 | −0.05 |
| $PM_{2.5}$ | | | | | | 1 | 1.00 | 0.23 | −0.15 | 0.16 | −0.17 |
| $PM_{10}$ | | | | | | | 1 | 0.24 | −0.16 | 0.18 | −0.18 |
| Temp | | | | | | | | 1 | −0.74 | 0.11 | −0.39 |
| Hum | | | | | | | | | 1 | −0.27 | 0.16 |
| WD | | | | | | | | | | 1 | 0.14 |
| WS | | | | | | | | | | | 1 |

Table 3. Summer correlation matrix in Zahleh.

| | NO | $NO_2$ | $NO_x$ | CO | $SO_2$ | $PM_{2.5}$ | $PM_{10}$ | Temp | Hum | WD | WS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NO | 1 | 0.64 | 0.77 | 0.38 | 0.66 | 0.15 | 0.12 | −0.45 | −0.08 | 0.21 | 0.27 |
| $NO_2$ | | 1 | 0.53 | 0.44 | 0.67 | 0.11 | 0.14 | 0.60 | −0.58 | 0.36 | −0.24 |

**Continued**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $NO_x$ | 1 | 0.41 | 0.62 | 0.18 | 0.18 | 0.02 | −0.43 | 0.41 | 0.06 |
| CO | | 1 | 0.31 | 0.22 | 0.23 | 0.15 | −0.25 | 0.21 | −0.09 |
| $SO_2$ | | | 1 | 0.18 | 0.16 | −0.16 | −0.27 | 0.26 | 0.27 |
| $PM_{2.5}$ | | | | 1 | 1 | 0.11 | −0.23 | 0.11 | 0.09 |
| $PM_{10}$ | | | | | 1 | 0.16 | −0.24 | 0.12 | 0.07 |
| Temp | | | | | | 1 | −0.58 | 0.11 | −0.45 |
| Hum | | | | | | | 1 | −0.33 | 0.00 |
| WD | | | | | | | | 1 | 0.32 |
| WS | | | | | | | | | 1 |

These Pearson correlation matrices were built and found that, during winter, the temperature has a moderate and positive correlation with $NO_2$ and $SO_2$ at Zahleh, and weak correlation with the remaining pollutants. The humidity correlates strongly negatively with $NO_2$ and $SO_2$ at Zahleh, and correlates negatively with $NO_2$, $NO_x$, and $SO_2$ moderately. The wind direction shows a moderate positive correlation with $NO_x$, CO and $SO_2$ at Zahleh. The wind speed has a moderate negative correlation with NO, $NO_x$ and CO, and a strong negative correlation with $NO_2$.

During summer, the temperature has a high positive correlation with $NO_2$, but a moderate negative correlation with NO at Zahleh. Humidity correlates negatively and strongly with $NO_2$ but moderately with $NO_x$ at Zahleh, negatively correlates moderately with NO and $SO_2$, and negatively correlates strongly with $NO_2$ and $NO_x$. The wind direction has a positive and medium correlation with $NO_2$ and $NO_x$. The wind speed has a weak correlation with all the air pollutants during the entire year at Zahleh. This is attributed to Zahleh being enclosed in two hills which reduces the wind speed. As all the pollutants share some common sources and some pollutants are in the same cluster, we could identify some correlations between them. NO and $NO_2$ belong to the $NO_x$ group, and they strongly correlated and similarly correlate to other pollutants. This is also the case for $PM_{2.5}$ and $PM_{10}$. During the winter, $NO_x$ and their group are strongly positively correlated with CO and $SO_2$, and are weakly positively correlated with PM. PM is weakly positively correlated with all the pollutants. $SO_2$ and CO are strongly positively correlated. During the summer season, $NO_x$ and their group show a high, moderate, and low positive correlation with $SO_2$, CO, and PM, respectively. PM has a weak positive correlation with the other pollutants. CO and $SO_2$ have a moderate and positive correlation. It is noted that the correlation between the different pollutants, and between the various pollutants and meteorological factors, has not significantly varied among the different seasons. Therefore, in the following process, the seasonal variation is not considered and the results will be treated annually.

## 6. Application of AQI

The principal purpose of AQI is to evaluate the concentration of atmospheric pollutants to study air quality. AQI is a dimension-free measure. Firstly, the sub-AQI of the six principal pollutants ($PM_{2.5}$, $PM_{10}$, $SO_2$, CO, $NO_2$, and $O_3$) were computed using the observed concentrations. Secondly, AQI is derived based on the maxima of the sub-IAQs of all pollutants, as illustrated in equation 2. It should be noted that once the AQI is greater than 50, the maximum sub-AQI pollutant is defined as the primary pollutant occurring that day. The higher AQI indicates that air pollution is serious and produces major health damage. The AQI formula is shown as follows:

$$\mathrm{IAQI_P} = \frac{I_{High} - I_{Low}}{C_{High} - C_{Low}}\left(C_p - C_{Low}\right) + I_{Low} \tag{1}$$

$$\mathrm{AQI} = \max\left(\mathrm{IAQI}_1, \mathrm{IAQI}_2, \cdots, \mathrm{IAQI}_n\right) \tag{2}$$

The AQI can be divided into six categories of air quality assessment, presented in Table 4. When the values of the AQI are below 100, the quality of air is adequate. Where the AQI value is almost 100, pollutant measurements are within legal guidelines. In contrast, when the AQI values exceed 100, the air quality deteriorates. The U.S. Environmental Protection Agency (EPA) has proven the existence of one national standard for the quality of the air in order the protect public health [4].

Table 4. EPA's AQI values.

| $PM_{2.5}$ (µg\|m³) | $PM_{10}$ (µg\|m³) | CO | $SO_2$ | $NO_2$ | $NO_x$ | AQI | AQI Category |
|---|---|---|---|---|---|---|---|
| $C_{low}$ - $C_{high}$ (24 hr) | $C_{low}$ - $C_{high}$ (24 hr) | $C_{low}$ - $C_{high}$ (24 hr) | $C_{low}$ - $C_{high}$ (24 hr) | $C_{low}$ - $C_{high}$ (24 hr) | $C_{low}$ - $C_{high}$ (24 hr) | $I_{low}$ - $I_{high}$ | |
| 0.0 - 12.0 | 0 - 54 | 0.0 - 4.4 | 0 - 35 | 0 - 53 | 0 - 40 | 0 - 50 | Good |
| 12.1 - 35.4 | 55 - 154 | 4.5 - 9.4 | 36 - 75 | 54 - 100 | 81 - 180 | 51 - 100 | Moderate |
| 35.5 - 55.4 | 155 - 254 | 9.5 - 12.4 | 76 - 185 | 101 - 360 | 41 - 80 | 101 - 150 | Unhealthy for Sensitive Groups |
| 55.5 - 150.4 | 255 - 354 | 12.5 - 15.4 | 186 - 304 | 361 - 649 | 181 - 280 | 151 - 200 | Unhealthy |
| 150.5 - 250.4 | 355 - 424 | 15.5 - 30.4 | 305 - 604 | 650 - 1249 | 281 - 400 | 201 - 300 | Very Unhealthy |
| 250.5 - 350.4 | 425 - 504 | 30.5 - 40.4 | 605 - 804 | 1250 - 1649 | 400 | 300 | Hazardous |

All results obtained are shown and plotted below.

Table 5. Descriptive data of measured air quality.

| Mean | Std | Min | $Q_1$ | $Q_2$ | $Q_3$ | Max |
|---|---|---|---|---|---|---|
| 139.13 | 53.89 | 52 | 110 | 129 | 151 | 429 |

Table 5 shows some characteristics of AQI from June 2017 to 31st December 2018. The average AQI score for the Zahleh cities is equal to 139.13. Although Memchiyeh monitoring recorded the highest average AQI score and the most

dispersed measure of air pollution concentration (SD = 53.89; range = 377). These results ensure our previous ones, and that Zahleh is affected by air pollution and suffers from serious health problems.

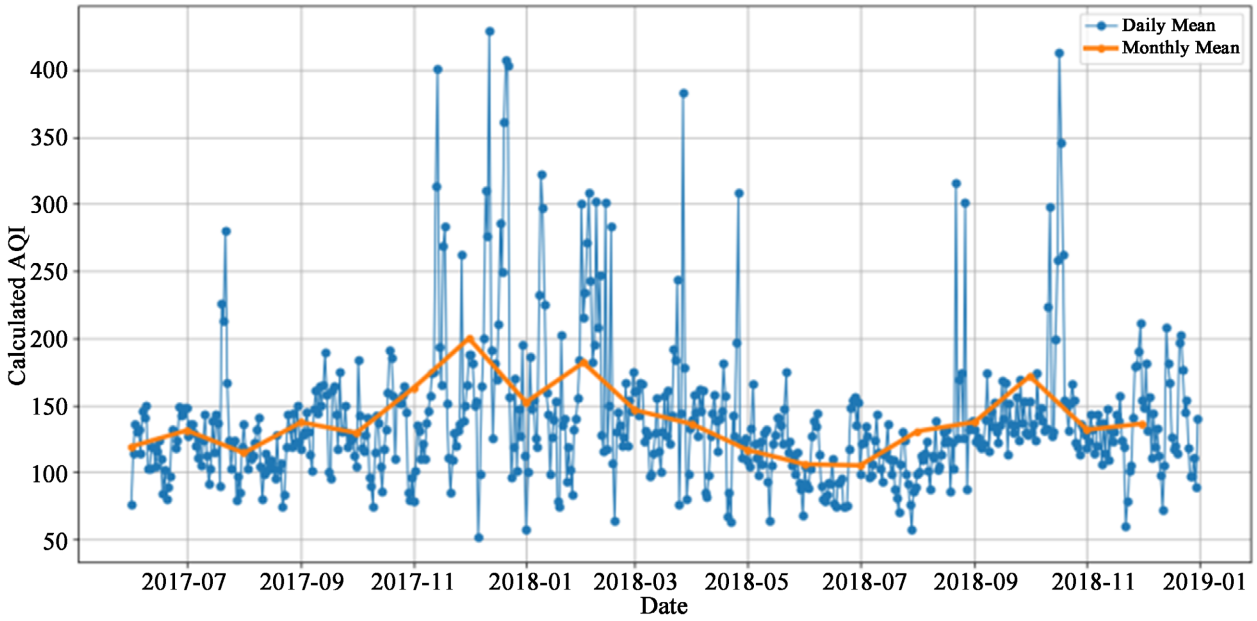


**Figure 16.** Daily and Monthly average of calculated AQI in Zahleh.

Figure 16 shows the average of daily and monthly calculated AQI. The monthly AQI reaches its first maxima (199.839) in December and the second one (181.75) in February, then reaches its minima (105.161) in July and the second one (106.033) in June. These results ensure our previous analysis, and that the pollution increase in the winter and decreases in the summer.

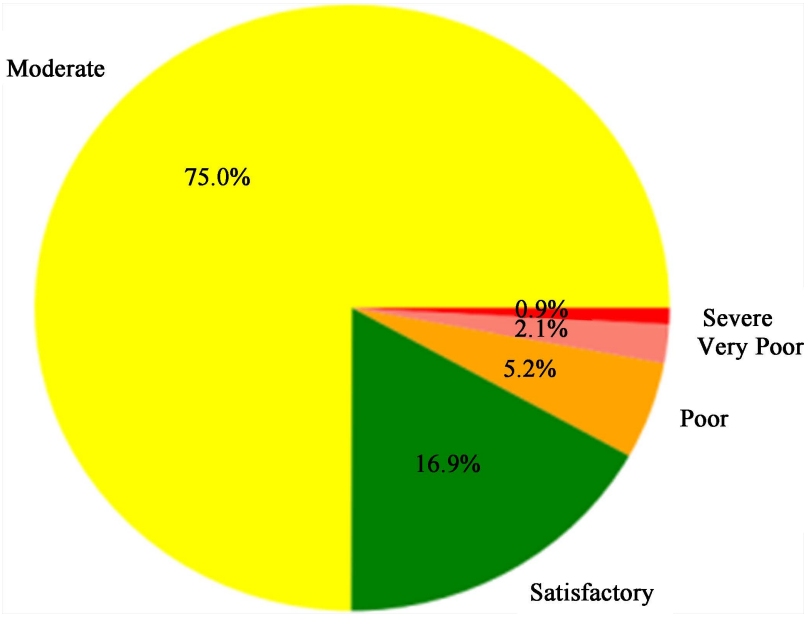


**Figure 17.** Daily AQI in Zahleh 06-2017 to 12-2018.

We notice that 98 days (16.9%) out of total days are leveled as satisfactory, 434 (75%) leveled as moderate, 30 (5.2%) as poor, and 12 (2.1%) as very poor, and 5 (0.9%) as severe (Figure 17). This means that the people suffering from lungs, asthma, and heart disease are breathing uncomfortably 75% of the days of the year in Zahleh, 5.2% of the days, people exposed to air pollution will breathe discomfort, 2.1% of the days, exposed people will start developing respiratory illnesses, and 0.9% of the days, healthy people will be affected by the pollution, and people with diseases will be subjected to serious impacts. Generally, Zahleh is affected by pollution, and the quality of the air in this region is generally not satisfactory, which causes many health problems and diseases.

## 7. AQI Prediction and Seasonal Decomposition

Prior to beginning the predictive models, we should first decompose the univariate time series. After testing both additive and multiplicative air quality models by using the seasonal decompose function, we noticed that the multiplicative model fitted significantly more efficiently to the data set. The seasonally decompose function yields the following results.

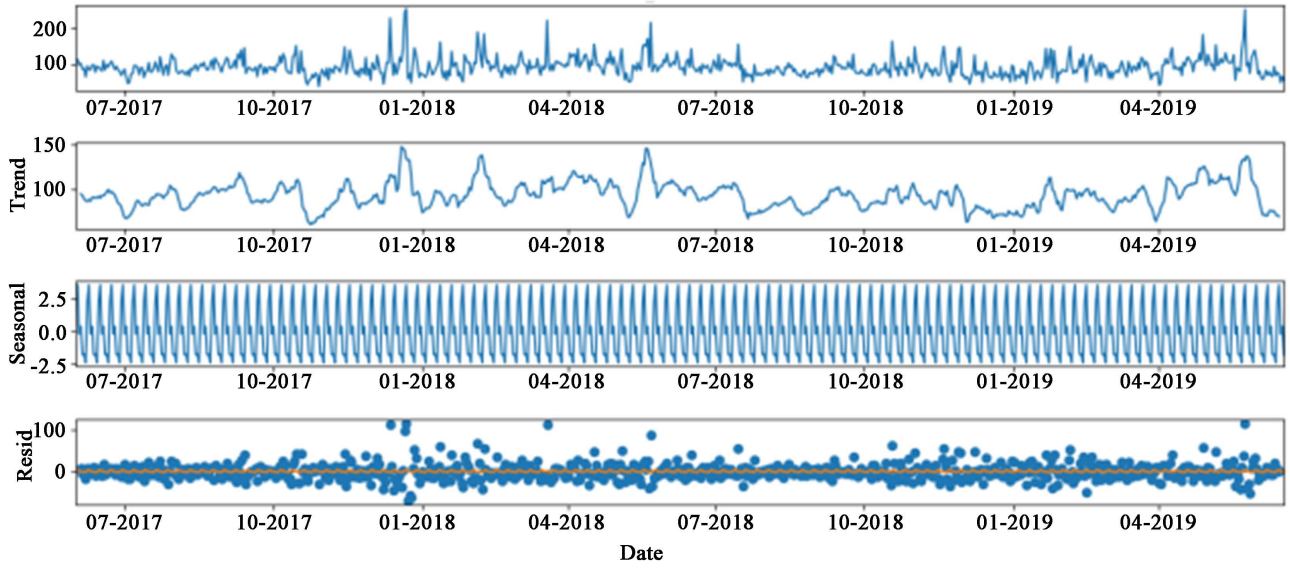


Figure 18. Daily AQI in Zahleh 06-2017 to 12-2018.

Figure 18 shows a variation of the data over time (weekly). In addition, the residuals vary around zero, which verifies the reliability of the predictions and of the fitted model.

### Univariate Time Series Forecasting

Air quality forecasting models have two main purposes. First, they validate theoretical knowledge of the atmospheric mechanisms that govern the evolution of air quality and are therefore necessary for research. They also help authorities to monitor the state of the air to which the population and the natural environ-

ment are exposed, for the protection of public health and the environment. The predictive models provide the authorities with the necessary information in time to take punctual measures to limit or mitigate pollution peaks in the short term. These measures mainly consist in limiting pollutant emissions through restrictions on traffic (automotive, maritime, airport), industrial activities (production of goods, energy...), or domestic activities (notably heating). Forecasting models also allow permanent monitoring that can guide public policies in terms of land use planning to consider the air quality and improve it. New development projects are being designed to avoid localized concentrations of pollutants. The predictive models can provide an efficient tool to evaluate various scenarios. The Naïve, ETS, TBATS, and SARIMA models have been evaluated using the above measures: RMSE, MAE, and RE (Table 6).

Table 6. Summary univariate forecasting models.

| Measurements | RMSE | MAE | RE |
|---|---|---|---|
| Naïve model | 40.44 | 26.81 | 19% |
| Exponential Smoothing | 67.80 | 49.95 | 36% |
| TBATS | 45.27 | 24.54 | 18% |
| SARIMA | 38.04 | 22.52 | 16% |

In order to evaluate the accuracy, we divided the data into two sets: Training set and test set. Then the test set has been compared with the predicted set, and Figure 19 has been obtained.

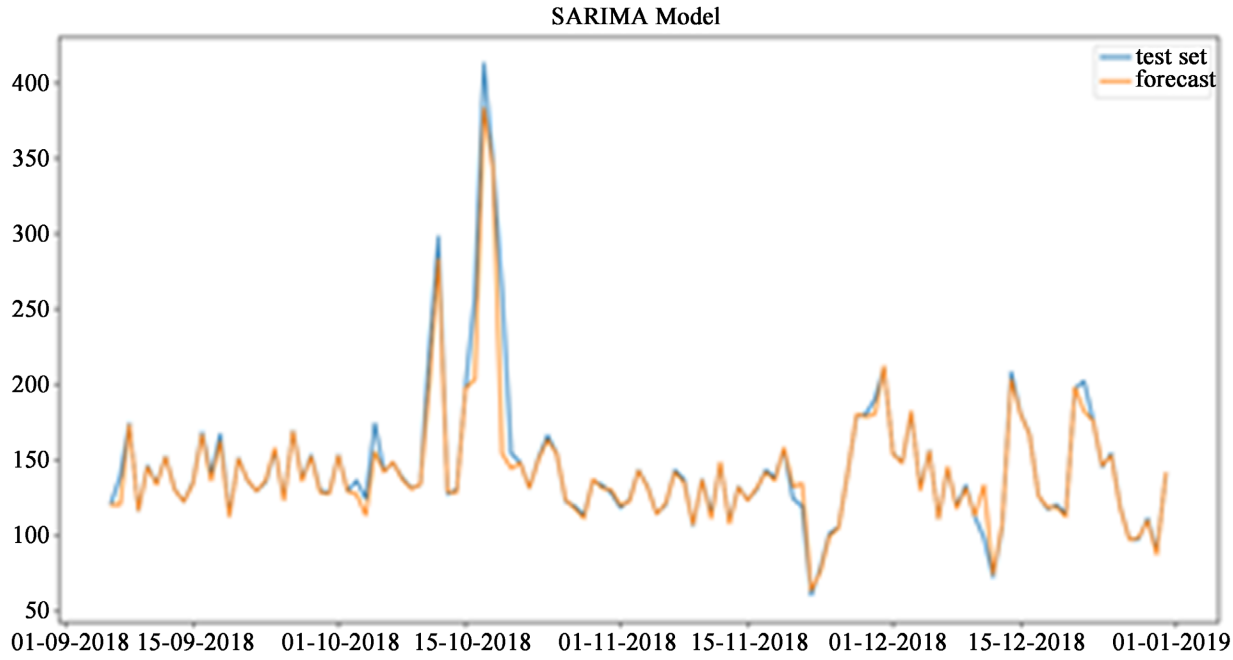


Figure 19. SARIMA model (test set) in Memchiyeh station.

Altogether, the SARIMA model has been shown to be a very appropriate

model to study the air pollution level and predict air quality. To construct the SARIMA model, it is necessary to select the hyperparameters for trend and seasonal components of the time series. The identification of hyperparameters can be done either by selecting the parameters directly or by testing several parameters and then selecting the one with the lowest AIC. Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots can be analyzed to find the correlations between the different lag times. An alternative approach for selecting the parameters is to make a loop using and varying $p$, $d$, $q$, $P$, $D$, and $Q$ between two numbers to get the hyperparameters that have the lowest AIC. In this article, we have used the second approach, which turned out to have more accuracy than the former. Also, the results obtained have been SARIMA (1, 1, 1) $(1, 1, 1)_{52}$ for two sets of data. Finally, the AQI is predicted by using the SARIMA model, and the results have now been plotted and summarized in Figure 20.

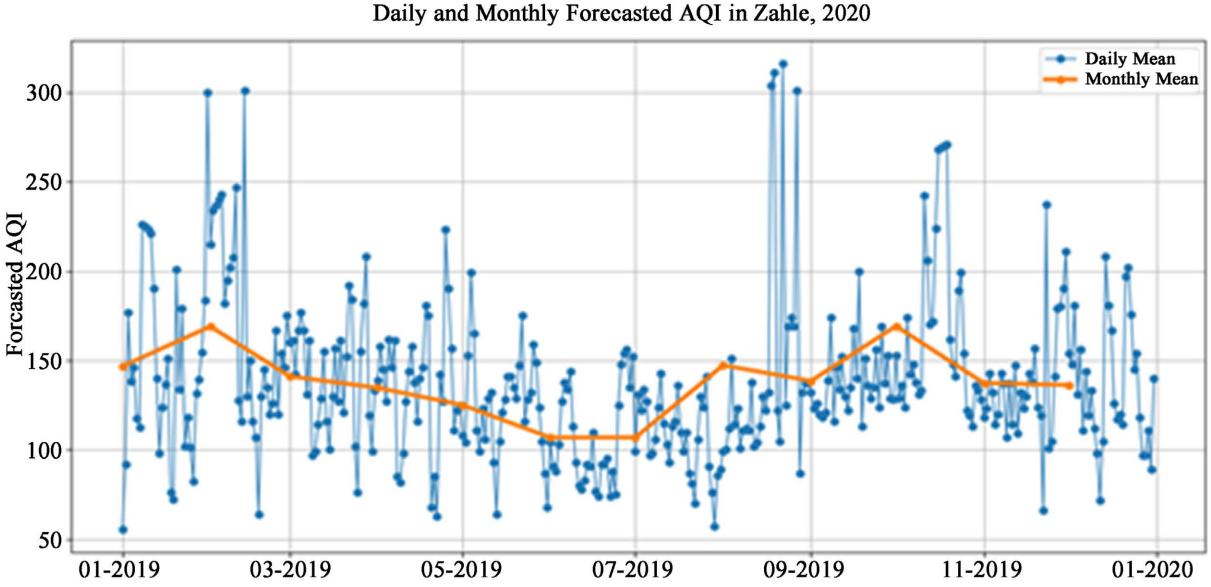


Figure 20. Daily and Monthly forecasted AQI in Memchiyeh station.

For the forecasted Air Quality Index, the results are shown in Table 7.

Table 7. Data descriptive of forecasted air quality.

| Statistics | Mean | Std | Min | Q1 | Q2 | Q3 | Max |
|---|---|---|---|---|---|---|---|
| Value | 138.07 | 44.52 | 55.79 | 111 | 131 | 154 | 316 |

It can also be noted that the average air quality value is 138.07 in Zahleh. The lowest value of AQI is 55.79 and the highest is 316 respectively. These findings confirm the earlier ones that Zahleh is affected by air pollution and has serious problems of health.

## 8. Conclusion

The time series approach was implemented for the air pollutants in Zahleh. The

aim was to investigate the correlation between air pollution, pollutants, and certain meteorological conditions, during a specific period. Then the Air Quality Index was calculated and a prediction of the air pollutants and AQI was performed by using the most suitable multivariate time series prediction methods. The findings indicated that the air pollutants frequently exceed the guideline levels, particularly NO, $NO_2$, and $NO_x$, which reach their maximum at peak periods. Hence, Zahleh is considered to be a much polluted city and it is generally most highly polluted during winter when NO, $NO_x$, $SO_2$, and CO concentrations are higher than during summer. In terms of correlation with the meteorological components, the Pearson coefficients revealed a strong, moderate, and weak correlation with atmospheric pollution. The Air Quality Index showed that air quality in Zahleh is unhealthy, becoming extremely worse during winter. To forecast future air quality values, numerous univariate forecasts and time series models were tested. The SARIMA model was shown to have the most accuracy for the data available. Note that several factors affect air pollution, including human activities, weather, and topography. Therefore, the predictions can provide some basic information about the air quality in this region but can be susceptible to some changes. However, the results showed that there is an urgency to make some decisions concerning the air quality in this city. To begin the process, the removal of manufacturing and burning installations from the residential area could reduce pollution levels and exposure to a highly polluted atmosphere.

## Acknowledgements

The authors would like to thank the personnel of the Air Quality Department, Ministry of Environment (MoE), Lebanon for providing information and the relevant data.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Kahraman, C. (2009) Risk Analysis and Crisis Response. *Stochastic Environmental Research and Risk Assessment*, **23**, 413-414. https://doi.org/10.1007/s00477-008-0230-x

[2] Hamanaka, R.B. and Mutlu, G.M. (2018) Particulate Matter Air Pollution: Effects on the Cardiovascular System. *Frontiers in endocrinology*, **9**, Article 680. https://doi.org/10.3389/fendo.2018.00680

[3] World Health Organization (2021) Ambient (Outdoor) Air Pollution. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[4] Lyon, F. (1994) IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. *Some Industrial Chemicals*, **60**, 389-433.

[5] Lelieveld, J., Berresheim, H., Borrmann, S., Crutzen, P.J., Dentener, F.J., Fischer, H.,

Feichter, J., Flatau, P.J., Heland, J., Holzinger, R., Korrmann, R., Lawrence, M.G., Levin, Z., Markowicz, K.M., Mihalopoulos, N., Minikin, A., Ramanathan, V., de Reus, M., Roelofs, G.J., Scheeren, H.A., Sciare, J., Schlager, H., Schultz, M., Siegmund, P., Steil, B., Stephanou, E.G., Stier, P., Traub, M., Warneke, C., Williams, J. and Ziereis, H. (2002) Global Air Pollution Crossroads over the Mediterranean. *Science*, **298**, 794-799. https://doi.org/10.1126/science.1075457

[6] AbuKhalil, A. (1989) Geography. In: Collelo, T., Ed., *Lebanon: A Country Study*, Federal Research Division, Library of Congress, Washington DC, 42-48.

[7] EPA (1999) Guideline for Reporting of Daily Air Quality—Air Quality Index (AQI). United States Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, Washington DC.

[8] Daher, N., Saliba, N.A., Shihadeh, A.L., Jaafar, M., Baalbaki, R. and Sioutas, C. (2013) Chemical Composition of size-Resolved Particulate Matter at Near-Freeway and Urban Background Sites in the Greater Beirut Area. *Atmospheric Environment*, **80**, 96-106. https://doi.org/10.1016/j.atmosenv.2013.08.004

[9] Baalbaki, R., El Hage, R., Nassar, J., Gerard, J., Sabila, N.B., Zaarour, R., Abboud, M., Fara, W., Khalaf, L.K., Shihadeh, A.L. and Saliba, N.A. (2016) Exposure to Atmospheric PMs, PAHs, PCDD/Fs and Metals Near an Open Air Waste Burning Site in Beirut. *Lebanese Science Journal*, **17**, 91-103. https://doi.org/10.22453/LSJ-017.2.091103

[10] Arnaudo, E., Farasin, A. and Rossi, C. (2020) A Comparative Analysis for Air Quality Estimation from Traffic and Meteorological Data. *Applied Sciences*, **10**, Article No. 4587. https://doi.org/10.3390/app10134587

[11] Ghanem, D.A. (2018) Energy, the City and Everyday Life: Living with Power Outages in Post-War Lebanon. *Energy Research & Social Science*, **36**, 36-43. https://doi.org/10.1016/j.erss.2017.11.012

[12] Farah, W., Nakhlé, M.M., Abboud, M., Annesi-Maesano, I., Zaarour, R., Saliba, N., Germanos, G. and Gerard, J. (2014) Time Series Analysis of Air Pollutants in Beirut, Lebanon. *Environmental Monitoring and Assessment*, **186**, 8203-8213. https://doi.org/10.1007/s10661-014-3998-9

[13] Atoui, A., Sami, A.A., Slim, K., Moilleron, R. and Khraibani, Z. (2023) Prediction and Analysis of the Extreme and Records Values of Air Pollution Data in Bekaa Valley in Lebanon. *International Journal of Environmental Science and Development*, **14**. https://doi.org/10.18178/IJESD

[14] Kumar, A. and Goyal, P. (2011) Forecasting of Daily Air Quality Index in Delhi. *Science of the Total Environment*, **409**, 5517-5523. https://doi.org/10.1016/j.scitotenv.2011.08.069

[15] Xia, X., Qi, Q., Liang, H., Zhang, A., Jiang, L., Ye, Y., Liu, C. and Huang, Y. (2017) Pattern of Spatial Distribution and Temporal Variation of Atmospheric Pollutants during 2013 in Shenzhen, China. *ISPRS International Journal of Geo-Information*, **6**, Article No. 2. https://doi.org/10.3390/ijgi6010002

[16] Wang, L., Wang, J., Tan, X. and Fang, C. (2020) Analysis of $NO_x$ Pollution Characteristics in the Atmospheric Environment in Changchun City. *Atmosphere*, **11**, Article No. 30. https://doi.org/10.3390/atmos11010030

[17] Saliba, N.A., Moussa, S., Salame, H. and El-Fadel, M. (2006) Variation of Selected Air Quality Indicators over the City of Beirut, Lebanon: Assessment of Emission Sources. *Atmospheric Environment*, **40**, 3263-3268. https://doi.org/10.1016/j.atmosenv.2006.01.054

[18] Lee, C.K. (2002) Multifractal Characteristics in Air Pollutant Concentration Time

Series. *Water, Air, and Soil Pollution*, **135**, 389-409.
https://doi.org/10.1023/A:1014768632318

[19] Voigt, K., Welzl, G. and Bruggemann, R. (2004) Data Analysis of Environmental Air Pollutant Monitoring Systems in Europe. *Environmetrics*, **15**, 577-596.
https://doi.org/10.1002/env.653

[20] Kamińska, J.A. (2018) The Use of Random Forests in Modelling Short-Term Air Pollution Effects Based on Traffic and Meteorological Conditions: A Case Study in Wrocław. *Journal of Environmental Management*, **217**, 164-174.
https://doi.org/10.1016/j.jenvman.2018.03.094

[21] Kumar, T.S., Das, H.S., Choudhary, U., Dutta, P.E., Guha, D. and Laskar, Y. (2021) Analysis and Prediction of Air Pollution in Assam Using ARIMA/SARIMA and Machine Learning. In: Muthukumar, P., Sarkar, D.K., De, D., De, C.K., Eds., *Innovations in Sustainable Energy and Technology. Advances in Sustainability Science and Technology*, Springer, Singapore, 317-330.
https://doi.org/10.1007/978-981-16-1119-3_28

[22] Kayes, I., Shahriar, S.A., Hasan, K., Akhter, M., Kabir, M.M. and Salam, M.A. (2019) The Relationships between Meteorological Parameters and Air Pollutants in an Urban Environment. *Global Journal of Environmental Science and Management*, **5**, 265-278.

[23] Van Buuren, S. and Groothuis-Oudshoorn, K. (2011) Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1-67.
https://doi.org/10.18637/jss.v045.i03

[24] Lasheras, S.F., García Nieto, P.J., Gonzalo, G.E., *et al.* (2010) Evolution and Forecasting of PM10 Concentration at the Port of Gijon (Spain). *Scientific Reports*, **10**, Article No. 11716. https://doi.org/10.1038/s41598-020-68636-5

[25] MoE/EU/UNDP (2014) Lebanon Environmental Assessment of the Syrian Conflict & Priority Interventions. MoE/EU/UNDP, Lebanese Ministry of Environment, Beirut.

[26] Saroufim, A. and Otayek, E. (2019) Analysis and Interpret Road Traffic Congestion Costs in Lebanon. *MATEC Web of Conferences*, **295**, Article No. 02007.
https://doi.org/10.1051/matecconf/201929502007