Carpathian Journal of Earth and Environmental Sciences, February 2022, Vol. 17, No. 1, p. 119 – 130; DOI:10.26471/cjees/2022/017/206

MODELLING THE AIR QUALITY INDEX FOR BOLU, TURKEY

Mine Tulin ZATEROGLU
Electrical and Energy Department, Vocational School of Technical Sciences, Cukurova University, Turkey
(mtzateroglu@cu.edu.tr)

Abstract: The monthly air quality index (AQI) derived from ground observation stations that obtained daily air pollutants information for 1990- through 2010 was analyzed in this study. AQI was evaluated using the common comparative index method presented by the U.S. Environmental Protection Agency (USEPA), and a statistically based approach was used for predicting the AQI value. With the first method, AQI was predicted using the USEPA subindex formula for different pollutants, such as particulate matter and sulfur dioxide, which contribute the most to air pollution. A combination of the principal component analysis (PCA) and multiple linear regression (MLR) methods were used with the measured values of climate variables obtained from the ground stations for the most effective contributors and a prediction was modelled. The results of these two methods were compared and evaluated for consistency. Two methods were presented for determining the AQI value. According to the findings, the common comparative index method was consistent with the statistical prediction models, and the best results were obtained using PCA models with varimax rotation.

Keywords: Air quality index, pollutants, climate variables, MLR, PCA.

1. INTRODUCTION

Environmental pollutants have gained worldwide attention in recent years. In particular, air pollutants have caused several problems in urban areas, and some, such as particulate matter of smaller than 10 or 2.5 microns (particulate matter PM_{10} or $PM_{2.5}$) and sulfur dioxide (SO_2), reduce air quality and adversely affect human health.

The concentrations of air pollutants are measured at ground monitoring stations. These values are then determined to be index values for reporting the urban AQI, also expressed as the air pollution index (API). These values are calculated using several air pollutants, such as PM_{10} and $PM_{2.5}$, SO_2, ozone (O_3), nitrogen dioxide (NO_2), and carbon monoxide (CO). The sources of air pollutants may be natural or from anthropogenic activities, but the urban air pollutants come mostly from anthropogenic activities, such as fossil fuel combustion, traffic, heating, and a rapid population increases. These actions affect the climate system and degrade the air quality in urban areas. Air pollutants, such as PM_{10}, and SO_2, also affect the quantity of sunshine duration (Zateroglu, 2021a). Environmental variables, such as air pollutants and climate elements, interact within the atmospheric circumference, which is a thin sheet of air that extends from the surface of the Earth to the edge of space (Zateroglu, 2021b). Further, Kori et al., (2019a) studied the ambient AQI for industrial areas in India. Kori et al., (2019b) assessed the ambient AQI and found generally moderate levels of PM_{10}, $PM_{2.5}$, and SO_2 because of the standard limits in India.

In 1976, the U.S. Environmental Protection Agency (USEPA) introduced a Pollutant Standards Index (PSI) to measure air quality within a range between 0 and 500 based on National Ambient Air Quality Standards (NAAQS), then revised and renamed it in 1999 as AQI (Cheng et al., 2007; Monteiro et al., 2017). AQI, also known as the Air Pollution Index (API) (Kanchan et al., 2015), is a very important parameter by which people can determine when the air quality is good or bad, especially for human health. The AQI value is determined using the calculated index values of all pollutant concentrations measured in the ground-based monitoring stations. It is calculated as an index value using the common comparative index method. In addition, new AQI systems were created by researchers (Cairncross et al., 2007; Hu et al., 2015; Li et al., 2018; Trivero et al., 2012).

The amount of air pollutants is measured at ground-based monitoring stations; however, air pollutants cannot be measured at some locations in which monitoring stations are limited or absent. In addition, the USEPA AQI formula considers only the concentrations of pollutants and the pollutant with the biggest index value. Those with small index values together with meteorological factors that have crucial effects on air pollutants (i.e., air quality) are not included for AQI assessment. These constraints require the use of alternative methods to evaluate AQI by considering other factors. In addition, a new approach has been proposed by Gibergans-Baguena et al., (2020) because of USEPA's lack of a standardized scale that disregards the compositional nature of the concentrations of air pollutants.

Air pollution models, in general, use the Pasquill–Gifford–Turner (PGT) protocol to predict horizontal and vertical dispersion of a plume (Venkatram, 1996). These prediction models usually operate the Gaussian models that consider steady-state atmospheric conditions, such as constant wind speed and spatial homogeneity or flat terrain, to determine the dispersion of concentrations by taking into account the sources of the pollutants (USEPA, 1993). This method determines the stability of an atmospheric region by considering the horizontal surface wind, the amount of solar radiation, and the fractional cloud cover. To determine the atmospheric turbulence, the PGT scheme considers six stability classes from A to F, with class A being the highly unstable or most turbulent level, and class F being the extremely stable or least turbulent level and arranged according to wind speed, wind direction, solar radiation and cloud cover. The PGT scheme also considers the vertical temperature gradient; however, statistical models are preferred for predictions over Gaussian models because of requiring detailed information on the parameters (e.g., pollutant sources and the other variables).

Different statistical analysis methods are used for predicting AQI. The most commonly preferred ones by researchers are the principal component analysis (PCA) and multiple linear regression (MLR) methods and a combination of the two (Abdul-Wahab et al., 2005; Al-Alawi et al., 2008; Rajab et al., 2013; Sousa et al., 2007; Statheropoulos et al., 1998). In some studies, the nueral network approach was used to predict air pollutant concentrations (Boznar et al., 1993; Slini et al., 2006). Cotta et al., (2020) used the PCA method to identify unnecessary air quality monitoring stations. In addition, Lu et al., (2011) also used PCA to evaluate the performance of air quality monitoring networks, and Ibarra-Berastegi et al., (2009) and Sanchez et al., (1986) used PCA to assess the variability of SO_2. Comrie (1997) examined the neural network and regression analysis methods to predict ozone concentrations. Similarly, Finzi & Tebaldi (1982) constructed the mathematical models to predict pollutant concentrations. In addition, the meteorological variables were taken into account to estimate air pollutants using the statistical models (Cogliani, 2001; Sanchez et al., 1990; Ziomas et al., 1995).

MLR is one of the most commonly studied methods in climatic and atmospheric studies by determining a dependent variable using contributing independent variables. To avoid multicollinearity between independent variables, PCA is used, which is the preferred method in prediction studies. In many cases, the use of both methods is more convenient for more reliable results.

AQI within a specified region is influenced by climate elements and their interactions among air pollutants, transportation and deposits of pollutant concentrations. Wind speed, atmospheric pressure, relative humidity, precipitation, and air temperature affect the distribution and quantity of pollutant concentrations in the air. Wind speed and direction are considerable factors that help disperse and transport air pollutants into or out of the area based on prevailing winds and topography. Low- and high- pressure also affect pollutants. In low-pressure (cyclone) systems, air moves vertically to disperse pollutants; whereas, in high-pressure (anticyclone) systems, air moves downward, which does not disperse pollutants. Relative humidity facilitates the formation of precipitation and atmospheric heat absorption, which warms lower areas; however, under high relative humidity conditions, SO_2 reacts with water droplets in the atmosphere to form sulfuric acid (H_2SO_4), which degrades the environment. Under conditions of relative humidity, high pressure and warming, evaporation (water vapor) is slow, which changes with latitude (i.e., slow in the North; high in the South). Precipitation is a crucial parameter in air pollution. Raindrops that uptake solid and gaseous pollutants in the air and bring them down to earth wash out the atmosphere. Air temperature is a factor in burning combustible materials and in using heat in various buildings, which leads to an increase in the amount of air pollutant concentrations. In addition, air pollutants scatter and absorb the incoming solar radiation and reduce the direct solar radiation necessary to measure the sunshine duration. Tropospheric aerosols, which have a short lifetime of days or weeks, behave similarly to cloud condensation nuclei and contribute to the formation of cloud droplets that may fall as snow or raindrops.

Although not a source of air pollutants, topographical structures affect their levels and duration within the environment. Air Pollution is more concentrated in bowl-shaped topographic areas or in areas that form of a perpendicular groove that extends to the prevailing wind direction. The concave area surrounded by hills or mountains prevents air circulation and, thus, dispersion of air pollutants.

The main purpose of the present study was to present an alternative to the USEPA method for calculating AQI using climate elements and air pollutants. Climate elements measured in a meteorological station, such as sunshine duration (SD), cloudiness (CLD), relative humidity (RH), wind speed (WS), precipitation (PREC), air temperature maximum (TMAX) and minimum (TMIN), evaporation (EVP), and atmospheric pressure (PRES) were used to construct the statistical models to predict AQI. Air pollutants were selected based on accessibility and persistence of data at the station in which the measurements were recorded, and PM_{10} and SO_2 levels were the parameters of concern.

In the present study, the AQI value was calculated using the USEPA formula, after which two estimation methods were applied to the dataset. First, the stepwise linear regression analysis was evaluated for AQI. Second, PCA was studied to reduce the number of independent variables and determine the ones that were the most significant. Using PCA with regression models minimizes the collinearity of the datasets, which can lead to worst estimations and also defines the appropriate explanatory variables for the estimation of pollutant concentrations (Sousa et al., 2007).

To verify effectiveness of simulation, the results were compared for two estimation approaches. Thereafter, the findings were elicited from interpreting the obtained values.

2. STUDY AREA AND DATA

Bolu is in northern Turkey (latitude between 40º06ꞌ and 41º01ꞌN, longitude between 30º32ꞌ and 32º36ꞌE) (Fig. 1) and encompasses an area of 8323.39 km^2. The land comprises mountains spanning north to south and to the east and plains from west to east. The mountainous region is 56% of the total area; the plains are 8%. There are several lakes and streams within the area. The climate in Bolu changes based on the topographical structure. These different climate characteristics are known as a "continental temperate climate". The temperate west Black Sea climate, with rains and lower temperatures than the other Bolu areas, influences the northern region. The southern region is dominated by the Anatolia climate in central Anatolia, which is continental, with cold winters, hot and dry summers, and less rainfall than in other areas. Between the northern and southern regions, a subregion of the Black Sea climate, or a transitional climate type, dominates the area and encompasses the characteristics of both climates. In addition, the climate in Bolu is classified as C2, Bꞌ1, s, and bꞌ3 according to the Thornthwaite climate classification. Within the province, the dominant wind directions are west and south. The annual average WS is 1.4 m/s with a maximum WS of 24.4 m/s. Most of the territory is humid, with the remaining areas subhumid. The area receives frost during winter months and the highest temperatures during July (39.3ºC) and August (39.8ºC). The mean high temperature is in August (27.9ºC) and the mean low temperature in January (-3.6ºC). Annual average mean temperature, TMAX and TMIN are 10.5ºC, 17.1ºC, and 4.7ºC, respectively. The amount of PREC is high in December, January, and May and low in July and August. Annual average RH is ~72%. The lowest RH is in July and August with the highest in January, February, and November. In winter, SD is less than in summer. Annual total SD is 2250 h in the north and 2500 h in the south with an anuual average of 5.5 h. Average SD is low in December and January (both 2.1 h) and high in July (9.2 h) and August (8.9 h). Average annual total rainfall in Bolu is ~550 mm, with 30% of annual rainfall in winter, 29% in spring, 20% in summer and 21% in autumn. There is an average of ~138 rainy days. Annual average PRES is 930 hPa, with maximum and minimum PRES of 943.5 and 913.5 hPa, respectively.

According to the Turkish Statistical Institution (2020), Bolu has a population of 314,802 of whom 227,724, (~72%) live in urban areas and 87,078 (~28%) in rural areas. There are 118,375 vehicles within the province of which 57,743 are cars, 1,523 are minibuses, 934 are buses, 17,122 are small trucks, 6,034 are trucks, 12,092 are motorcycles, 511 are special-purpose vehicles, 22,416 are tractors. The emissions from these vehicles contribute to poor urban air quality.

Together, all state and provincial roads within the research area extend 679 km (General Directorate of Highways of Turkish Republic, 2020). Classified by surface types, there are 666km of asphalt of which 300 and 39 km are in provincial and surface treatment, respectively, and 366 km of which 257 km are in provincial, 2 km are stone block (in provincial), and 11 km are primitive (in provincial). The road density, defined as the ratio of the length of state and provincial roads to the province's land area within the research area is 8.16 km/100 km^2 at 8323.39 km^2. There are no subways or trams in Bolu. The annual average daily traffic (vehicle/d) in Bolu can be determined for motorways with 140,205 vehicle/d and state roads with 107,535 vehicles/d (such as 77,584 cars, 8,065 medium-goods vehicles, 998 buses, 7,804 trucks, and 13,086 articulated trucks) (General Directorate of Highways of Turkish Republic, 2020).

In addition, there are several types of industries in Bolu that mainly produce forest products and furniture, food, metallic goods, the heat glass and tempered glass, and electrical equipment, woven apparel and leather.

Air quality for the province is within the good ranges. Winter is the worst time for air pollution because fossil fuel is burned for heat. The main emission sources originate from increased motor vehicles emissions, topographical structure, and domestic heating. Coal, wood, liquid fuel (diesel), and natural gas are used for domestic heating. Because of the topography in the province, the air pollutants are not removed by air circulation. In addition, mining activities, such as cement and lignite quarries, cause dust emissions.

The ground based meteorological station is located in the center of Bolu (latitude 40º43'N, longitude 31º36'E). From 1990 through 2011 at the meteorological station site, only two pollutants, PM_{10} and SO_2, were measured at the air quality traffic station, which was active until 2017 and operated by the Ministry of Environment and Urbanization. There were residential areas and state roads within the vicinity of the station. Climate data, such as SD, CLD, RH, WS, PREC, TMAX, TMIN, EVP, and PRES, were obtained from the Turkish State Meteorological Service for 1990–2011. For the same period, the values for PM_{10} and SO_2 were obtained from the Turkish Statistical Institution and Ministry of Environment and Urbanism.

3. METHODS

Statistical analyses were conducted based on monthly climate variables. PCA and MLR methods were used to estimate the long-term AQI in Turkey using the meteorological parameters (SD, CLD, RH, WS, PREC, PRES, EVP, TMIN, and TMAX) as predictors. The combined PCA and MLR methods are called the Principal Component Regression (PCR) method.

AQI for a location is determined from five pollutants—PM_{10}, SO_2, O_3, NO_2, and CO—in the USEPA system. AQIs have been separated into the different categories according to their health effects. If an index value is >100, it is unhealthy for sensitive groups that have different diseases, such as respiratory problems and asthma.

AQI for all pollutants is determined using the USEPA method, which is based on the concentrations of the pollutants. For each pollutant, an index value is calculated using the following equation:

$$I_p = \left[\frac{(I_{Hi}-I_{Lo})}{(BP_{Hi}-BP_{Lo})}\right] \left(C_p - BP_{Lo}\right) + I_{Lo} \qquad (1)$$

where I_p, C_p, BP_{Hi}, BP_{Lo}, I_{Hi} and I_{Lo} are expressed as the index value for pollutant p, the concentration value of the pollutant p, the breakpoint that is equal to or bigger than C_p, the breakpoint that is equal to or smaller than C_p, the index value determining to BP_{Hi} and the index value determining to BP_{Lo} respectively. each calculated pollutant index I_p, as determined in Eq. (2).

$$AQI = Max(I_1, I_2, \ldots, I_p),\ p = 1, 2, \ldots, 5 \qquad (2)$$

During 1990–2011, only PM_{10} and SO_2 pollutants were measured at the Bolu station; therefore, the USEPA table for breakpoints has been used for only those two pollutants in Table 1 to determine AQI.

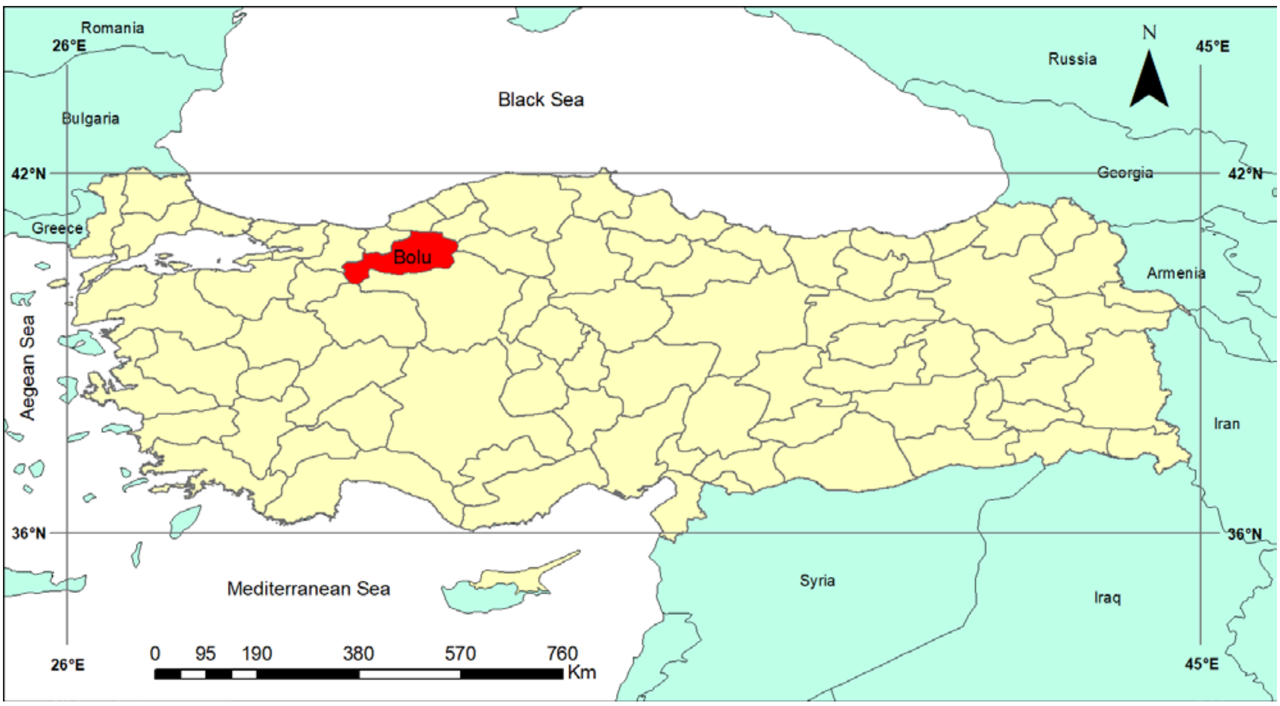

Figure 1. Turkey and the location of Bolu Province

Table 1. United States Environmental Protection Agency breakpoints of air pollutants for the air quality index (AQI) (EPA, 1999)

Breakpoints							AQI	Category
O_3 (ppm) 8-hour	O_3 (ppm) 1-hour[1]	PM_{10} (µg/m³)	$PM_{2.5}$ (µg/m³)	CO (ppm)	SO_2 (ppm)	NO_2 (µg/m³)		
0.000-0.064	-	0-54	0.0-15.4	0.0-4.4	0.000-0.034	(2)	0-50	Good
0.065-0.084	-	55-154	15.5-40.4	4.5-9.4	0.035-0.144	(2)	51-100	Moderate
0.085-0.104	0.125-0.164	155-254	40.5-65.4	9.5-12.4	0.145-0.224	(2)	101-150	Unhealty for sensitive groups
0.105-0.124	0.165-0.204	255-354	65.5-150.4	12.5-15.4	0.225-0.304	(2)	151-200	Unhealty
0.125-0.374	0.205-0.404	355-424	150.5-250.4	15.5-30.4	0.305-0.604	0.65-1.24	201-300	Very unhealty
(3)	0.405-0.504	425-504	250.5-350.4	30.5-40.4	0.605-0.804	1.25-1.64	301-400	Hazardous
(3)	0.505-0.604	505-604	350.5-500.4	40.5-50.4	0.805-1.004	1.65-2.04	401-500	Hazardous

[1] Generally AQI is reported by using 8-hour ozone values. Anyway, for some areas, AQI is evaluated using both 1- and 8-h ozone (O_3) levels and reported as the maximum of each.
[2] NO_2 has no short-term National Ambient Air Quality Standards (NAAQS) and can generate AQI if it is >200.
[3] For higher AQI values (301-400 and 401-500), 1-h O_3 concentrations are used instead of 8-h O_3 concentrations.

All air pollutants were defined as µg/m³ before calculating AQI in the present study. The values for SO_2 were revised to reflect those units as ppm in Table 1. SO_2 breakpoint values in µg/m³ were used in the Cairncross (2007) study to delineate the categories as follows: good, 0–90 µg/m³; moderate, 93–383 µg/m³; unhealthy for sensitive groups, 386–596 µg/m³; unhealthy, 599–809 µg/m³; very unhealthy, 811–1607 µg/m³; hazardous for AQI 301–400, 1609–2139 µg/m³; and hazardous for AQI 401–500, 2141–2671 µg/m³.

In the beginning of the analyses, daily values of pollutant concentrations were transformed into index values and used to define the daily AQI. Because of the missing values in the dataset, daily AQI values were converted into monthly average AQI values. This transformation was necessary for reliable analyses, so that they were compatible with the meteorological parameters.

3.1. Multiple linear regression (MLR)

The MLR method has been generally used by researchers to estimate AQI values. Climate data are fit to normal distribution, which assumes that the data have equal variance and the deviation between observed and estimated values of response variables are independent. MLR analyses can be used as statistical prediction methods in many applications, such as climatological studies, because of the compatibility with normal distribution. In this method, a response variable is predicted using two or more explanatory variables. This relationship is determined using the mathematical model shown in Eq. (3).

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots\dots\dots + a_rX_r + \varepsilon \quad (3)$$

where Y denotes the response variable; $X_1, X_2, \dots, X_r$ are explanatory variables; a_o and $a_1, a_2, \dots, a_r$ determine the constant and coefficients of regression, respectively; and ε identifies the error term that was estimated. To minimize the error term, the least squares method was used to predict the constant and coefficient values of the regression model using the coefficient matrix a in dimension r x 1. Matrix a is defined by $a = (X^TX)^{-1}(X^TY)$, where r is the number of independent variables, n is number of observations, Y is the observed value of the matrix in dimension n x 1 of the response variable, X is the observed value of the matrix in dimension n x r of the explanatory variables, and X^T determines the transpose of X. F distribution and the Student's *t*-test are used to determine the significance levels of the constant and coefficient values. Model consistency is determined using the estimation error and coefficient of determination. All prediction models have been constructed as statistically significant at a 95% confidence interval (CI).

The determination coefficient, R^2, denotes the percent value of the variance in the response variable, which is expressed by the explanatory variables in the model. The multiple determination coefficient value of a regression model that reveals the estimated success is determined using Eq. (4) as follows:

$$R^2 = 1 - \frac{\sum(Y_{e,i}-\bar{Y})^2}{\sum(Y_{o,i}-\bar{Y})^2} \quad (4)$$

where Y_e is the estimated value of Y_i, Y_o is the observed value of Y_i, and $\bar{Y}$ is the mean value of the observed values $Y_i s$. The value of R^2 is between 0 and 1, which indicates that when the value is closest to 1, the model is appropriate with the data.

Standard Error of Estimation (SEE) is expressed as amount of the distinction between the estimated and measured values. It is calculated using the following formula:

$$SEE = \sqrt{\frac{\sum(Y_o-Y_e)^2}{n-2}} \quad (5)$$

where Y_o determines the observed value, Y_e is the estimated value and n is the number of observations.

3.2. Principal component analysis (PCA)

Using SPSS ver. 25 (IBM Corp., Armonk, NY, USA), PCA was evaluated to determine the significance order of the meteorological parameters and reveal the interrelated structures as well as their influence on AQI. The obtained PCA scores of the meteorological variables were applied as explanatory variables (X_i, so that i = 1,2,3….) to MLR analysis to estimate AQI as a response variable (Y).

Bartlett's sphericity test (χ^2 with degrees of freedom equal to the formula k (k-1) / 2) was used to confirm the applicability of PCA to the sequence data (Stevens, 1986). The PCA scores' eigenvalues were acquired using the following equation (Johnson & Wichern, 1982):

$$|D - \lambda I| = 0 \quad (6)$$

where D, λ, and I denote the correlation matrix in the k x k dimension, eigenvalue vector, and identity matrix, respectively. In addition, the standardized weight values of the variables of the principal components were obtained using Eq. (7).

$$(D - \lambda I)V = 0 \quad (7)$$

where V defines the k x k dimension matrix, and the variables of the principal components' standardized weights (v_{ij}) are included using the V matrix. Each variable's weight and eigenvalue were calculated using the D matrix. In the analyses, the factor loadings without rotating were found using eigenvector. Then, with the varimax rotation, the values of the rotated factor loadings, C_{im}, which determine the contributions of the variables as a percentage of the related principal components, were acquired (i is the variable number, m is the principal component number). The variables were classified using the values of loadings for each principal component. Score values were calculated using Eq. (8).

$$s_{mj} = v_{1m}z_{1j} + v_{2m}z_{2j} + \cdots + v_{km}z_{kj} \quad (8)$$

where s_{mj} is the value of the standardized score, j is the observation number (1,2,….,n), k is the explanatory variables' number, z and v are the standardized value and standardized weight of the related variable and observation, respectively. In addition, z was obtained using $z = x_k - \bar{x}/s_x$, where x_k denotes the original values of the variables.

The performance of the prediction models was evaluated using the statistical indicators shown below. The mean biased error (MBE), root mean square error (RMSE), and index of agreement (IOA) were calculated to determine the model suitability defined as an MBE and RMSE value small, ideally closer to zero and IOA value nearer to 1. MBE represents the relevance of the predicted amounts, such as overprediction and underprediction, and is expressed as positive and negative values, respectively. RMSE is a measure of the distinction between the estimated and measured data. IOA is between 0 and 1 and shows that the grade to that estimation is error free.

$$MBE = \frac{\sum_{k=1}^{n}(E_k - M_k)}{n},$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{n}(E_k - M_k)^2}{n}},$$

$$IOA = 1 - \frac{\sum_{k=}^{n}(E_k - M_k)^2}{\sum_{k=1}^{n}(|E_k - \bar{M}_k| + |M_k - \bar{M}_k|)^2}$$

where E_k is estimated value, M_k is measured value, n is number of measurements.

4. RESULTS AND DISCUSSION

Two statistical approaches, MLR and PCR, were used to predict the AQI for 1990–2010. The 2011 AQI values were then used for validation. In the first approach, all meteorological parameters (i.e., SD, CLD, RH, WS, PREC, TMAX, TMIN, EVP, and PRES) were used as independent variables for MLR analyses. The best explanatory variables for the biggest contribution to the statistically significant response variable were selected using stepwise regression analysis. The mathematical models were constructed for April–September, annual, and seasonal terms using Eqs. (9–14).

For October-March;
No statistically significant model

For April-September;

$$AQI = -7877.553+106.02xWS+8.334xPRES \quad , R^2 = 0.624 \quad , SEE = 10.11 \quad (9)$$

For Annual;

$$AQI = -24{,}69+110.54xWS-17.382xCLD \quad , R^2 = 0.567 \quad , SEE = 12.96 \quad (10)$$

For Winter;

$$AQI = 274.671-1.215xSD-22.210xCLD \quad , R^2 = 0.496 \quad , SEE = 16.72 \quad (11)$$

For Spring;

$$AQI = -91{,}406+84.123xWS \quad , R^2 = 0.459 \quad , SEE = 12.007 \quad (12)$$

For Summer;

$$AQI = -172.441+103.228xWS+4.238xTMIN \quad , R^2 = 0.729 \quad , SEE = 8.01 \quad (13)$$

For Autumn;

$$AQI = -9.249+98.868xWS-15.297xCLD \quad , R^2 = 0.523 \quad , SEE = 15.57 \quad (14)$$

In Eqs. (9–14), WS and CLD were obtained as common variables with the MLR method. Air pollutants are transported by winds. Because Bolu is surrounded by hills, southern winds, especially during the heating period, may not disperse and transport air pollutants. In addition, air pollutants in the atmosphere act as nuclei in cloud condensation and clouds. Clouds form when RH is high enough that the atmospheric water vapor condenses into tiny liquid droplets. In addition, cloud condensation nuclei are a subset of hygroscopic aerosol particles that nucleate water droplets at supersaturations <1%; therefore, the two variables are closely related to pollutants. There was no statistically significant model for October–March. WS and PRES were significant variables for April–September. The effect of atmospheric pressure on pollutants changes with low and high PRES areas. In low-PRES systems, air moves upward, which disperses the pollutants; in high-PRES systems, the air moves downward, which does not disperse the pollutants. In winter, SD and CLD were obtained as significant variables with the prediction model. Air pollutants scatter and absorb the incoming solar radiation and reduce the direct solar radiation necessary for the equipment to measure SD. WS was the only parameter in spring. WS and TMIN were the significant parameters in summer. Finally, WS and CLD have effects on AQI for predictions in autumn. The predicted AQI uses the selected variables as ~62% in April-September, 57% in annually, 50% in winter, 46% in spring, 73% in summer, and 52% in autumn. For the models, the higher R^2 values accompany the lower SEE values.

In addition, according to stable atmospheric conditions with the PGT scheme, the dispersion of air pollutant concentrations is closely related to WS and direction, CLD, solar radiation, and vertical air temperature gradient.

Using PCR, the independent variables from MLR were selected based on the PCA components. The original independent variables were changed into principal components over the Eigenvalue variable matrix. Eigenvalues of variables describe most of the aggregate variation in the studied data (Table 2). To determine the best predictors of AQI, the principal components with Eigenvalues>1 were used (Johnson and Wichern, 1982). In Table 2, the Eigenvalues and quantity of variance of either principal component with Eigenvalue only >1 are shown. The other components with Eigenvalues < 1 were disregarded because they expressed a lower variance than each of the other input variables. Table 2 indicates that four principal components with Eigenvalues > 1 and having cumulative variance values were 75.09% for October–March and 78.106% for autumn, and three principal components with Eigenvalues > 1 having cumulative variance values were 77.504% for April–September, 80.175% annually, 77.671% for winter, and 71.383% for spring, and two principal components with Eigenvalues > 1 were 72.74% for summer.

In addition, communalities of the original variables for all time periods were determined for each term (Table 3), as mentioned by using the first four principal components in October-March and autumn, three components in April-September, annual, winter, and spring, two components in summer. According to communalities, the variables > 0.70 were considered (Stevens, 1986). Table 3 indicates the most convenient original variables for PCR. Considering the results of communalities, the primarily appropriate variables for each term of PCR could be ordered as SD in October–March, EVP for April–September, TMAX for annual, PRES for winter, WS for spring, EVP for summer, and TMIN for autumn. WS determines the horizontal transportation and dispersion of pollutants. High WS diffuse pollutant concentrations. Low WS cause haze episodes. However, due to geographical structure and wind direction conditions, high WS may not disperse the pollutants, reversely may contribute the accumulation of pollutant concentrations. In addition, high WS may lead to an increased EVP rate of pollutants and decreased pollutant concentrations. The air pollutant concentrations may decrease surface temperature by reflecting and scattering solar radiation.

Table 2. Eigenvalues and variances of principal components

Terms	Principal Component	Eigenvalue	% of Variance	Cumulative variance %
October-March	1	2.418	26.872	26.872
	2	1.988	22.092	48.963
	3	1.278	14.196	63.159
	4	1.074	11.931	75.09
April-September	1	3.717	41.305	41.305
	2	1.922	21.355	62.66
	3	1.336	14.844	77.504
Annual	1	3.312	36.797	36.797
	2	2.418	26.867	63.664
	3	1.486	16.511	80.175
Winter	1	3.056	38.201	38.201
	2	2.064	25.803	64.003
	3	1.093	13.667	77.671
Spring	1	3.526	39.174	39.174
	2	1.863	20.702	59.877
	3	1.036	11.506	71.383
Summer	1	4.997	55.527	55.527
	2	1.549	17.214	72.74
Autumn	1	2.688	29.866	29.866
	2	1.749	19.436	49.301
	3	1.478	16.426	65.728
	4	1.114	12.379	78.106

Table 3. Communalities of variables for different terms in 1990-2010

Variable	October-March	April-September	Annual	Winter	Spring	Summer	Autumn
SD	0.858	0.803	0.812	0.76	0.804	0.834	0.905
CLD	0.703	0.847	0.829	0.879	0.679	0.851	0.775
RH	0.654	0.844	0.752	0.697	0.752	0.824	0.834
WS	0.665	0.808	0.725	0.743	0.918	0.269	0.706
PREC	0.696	0.314	0.727	0.759	0.621	0.877	0.757
EVP	0.798	0.877	0.864	No data	0.724	0.901	0.753
PRES	0.756	0.789	0.787	0.889	0.704	0.675	0.856
TMIN	0.786	0.817	0.841	0.73	0.693	0.827	0.916
TMAX	0.842	0.875	0.88	0.756	0.529	0.49	0.527

Solar radiation is highly associated with SD because of Ansgtröm-Prescott formula in predicting global solar radiation; thus, pollutant concentrations have also influenced on SD. The reduced surface temperature near the ground causes a low atmospheric movement of pollutants, and hence an increased pollutant accumulation which leads to enhancement in the existing temperature inversion layer. Hence, the existing bidirectional interactions between air pollutant concentrations and climate parameters may conclude in an increase or decrease in air quality for the atmospheric environment.

To obtain the best result for the relationships between the principal components and meteorological variables in PCA, rotating the components is preferred. The principal components were rotated orthogonally and obliquely. The varimax method for orthogonal rotation and the promax method for oblique rotation were used for component rotations for all terms. By using the rotation methods, each of the meteorological variables was related to only one of the principal components having the highest value. After the rotations, factor loadings were obtained that determined the contributions of each variable to the principal

Table 4. Annual results of principal component analysis with varimax rotation for 1990-2010

Variable	Loading of variables for each component			Standardized weight of variables for each component		
	1	2	3	1	2	3
SD	**0.808**	-0.248	-0.312	0.325	-0.043	-0.249
CLD	-0.123	0.164	**-0.887**	0.077	0.156	-0.483
RH	**-0.799**	0.182	-0.282	-0.249	0.069	-0.069
WS	-0.154	0.449	**0.707**	-0.123	0.137	0.362
PREC	-0.386	**0.76**	0.005	-0.101	0.339	-0.021
EVAP	0.645	-0.055	**0.667**	0.15	-0.054	0.286
PRES	-0.02	**-0.881**	0.104	-0.069	-0.429	0.143
TMIN	0.494	**0.689**	0.351	0.176	0.321	0.062
TMAX	**0.909**	0.213	0.096	0.335	0.142	-0.084

Table 5. Annual results of principal component analysis with promax rotation for 1990-2010

Variable	Loading of variables for each component			Standardized weight of variables for each component		
	1	2	3	1	2	3
SD	**0.866**	-0.156	-0.404	0.295	-0.086	-0.204
CLD	0.025	0.287	**-0.94**	0.012	0.096	-0.443
RH	**-0.767**	0.174	-0.215	-0.259	0.072	-0.094
WS	-0.242	0.349	**0.709**	-0.085	0.183	0.355
PREC	-0.348	**0.755**	-0.041	-0.119	0.339	0.008
EVAP	0.555	-0.109	**0.625**	0.185	-0.028	0.295
PRES	-0.095	**-0.921**	0.224	-0.031	-0.406	0.076
TMIN	0.498	**0.69**	0.218	0.166	0.317	0.127
TMAX	**0.935**	0.266	-0.046	0.316	0.116	-0.017

components. In Tables 4 and 5, loadings of meteorological variables for each component as bold and standardized weights of variables are shown for both rotations annually. The results of PCA for the annual term were presented in Table 4 for varimax rotation method and Table 5 for promax rotation method. Three principal components were retained with two methods. For ease of evaluation, only the loadings exceeding 0.5 in absolute values were indicated; the loadings of smaller magnitudes were regarded as insignificant and were removed. The principal components with high loadings of comparable sizes on the identical variables (i.e. with analogous construction) showed up in all analyses; hereby they could be assumed consistent and not sensitive to a specific selection of the rotation technique.

The principal component scores were computed as mentioned in Eq. (8), such as multiplying the standardized value by the standardized weights of the variables. The scores acquired in PCA were used as the independent variables in MLR. In stepwise regression analysis, only statistically significant (95% CI) score variables were selected and nonsignificant values were removed from the mathematical prediction model for AQI. SPSS ver. 25 (IBM Corp.) was used for the statistical approach estimation.

PCR models with varimax rotation were obtained for October–March, April–September, annual, winter, spring, summer, and autumn, respectively, in Eqs. (15–21).

$$AQI = 53.979+9.788*(Score4) \quad , R^2 = 0.232 \quad , SEE = 18.3 \tag{15}$$

$$AQI = 23.667+11.941*(Score3)-6.606*(Score2) \quad , R^2 = 0.765 \quad , SEE = 8.001 \tag{16}$$

$$AQI = 40.679+15.526*(Score3) \quad , R^2 = 0.689 \quad , SEE = 10.7 \tag{17}$$

$$AQI = 61.67+10.322*(Score2) \quad , R^2 = 0.214 \quad , SEE = 20.3 \tag{18}$$

$$AQI = 35.491+12.029*(Score3) \quad , R^2 = 0.605 \quad , SEE = 10.26 \tag{19}$$

$$AQI = 15.311+7.063*(Score1)+6.376*(Score2) \quad , R^2 = 0.384 \quad , SEE = 12.08 \quad (20)$$

$$AQI = 38.546+14.165*(Score2) \quad , R^2 = 0.438 \quad , SEE = 16.46 \quad (21)$$

Because of the PCR results of the varimax rotation, one component (Score4) was found to be statistically significant for October–March, two (Score3 and Score2) for April–September, one (Score3) for annual, one (Score2) for winter, one (Score3) for spring, two (Score1 and Score2) for summer, and one (Score2) for autumn. These equations indicated that the predicted AQI can be explained using the selected variables as ~23% in October-March, 77% in April-September, 69% annually, 21% in winter, 61% in spring, 38% in summer and 44% in autumn.

For the same terms and using the promax rotation, the models were constructed as shown in the equations (22-28) below. According to these equations, the forecasted AQI is expressed by the selected variables as ~20% in October-March, 72% in April-September, 71% annually, 30% in winter, 61% in spring, 38% in summer and 45% in autumn.

$$AQI = 53.979+9.077*(Score4) \quad , R^2 = 0.199 \quad , SEE = 18.69 \quad (22)$$

$$AQI = 23.667+11.359*(Score3)-5.895*(Score2) \quad , R^2 = 0.719 \quad , SEE = 8.74 \quad (23)$$

$$AQI = 40.679+16.087*(Score3)-6.511*(Score1) \quad , R^2 = 0.71 \quad , SEE = 10.61 \quad (24)$$

$$AQI = 61.67-8.671*(Score1)+11.089*(Score2) \quad , R^2 = 0.297 \quad , SEE = 19.76 \quad (25)$$

$$AQI = 35.463+12.065*(Score3) \quad , R^2 = 0.609 \quad , SEE = 10.21 \quad (26)$$

$$AQI = 15.311+9.208*(Score1)+8.864*(Score2) \quad , R^2 = 0.384 \quad , SEE = 12.087 \quad (27)$$

$$AQI = 38.546+14.383*(Score2) \quad , R^2 = 0.452 \quad , SEE = 16.26 \quad (28)$$

As a result of the promax rotation models, the principal components were ensured to be statistically significant for all terms as one component (Score4) for October–March, two (Score3 and Score2) for April–September, two (Score3 and Score1) for annual, two (Score1 and Score2) for winter, one (Score3) for spring, two (Score1 and Score2) for summer, and one (Score2) for autumn and mostly similar to those from the varimax method. As variable selection methods, varimax and promax rotation of principal components were utilized to select the convenient independent parameters for involvement in the eventual regression models. These approaches reduce the influence of multicollinearity on the prediction of the statistical coefficients in regression models.

Table 6. Statistical indicators for prediction models

Indicator	MLR	PCR with varimax	PCR with promax
RMSE	35.67	27.72	28.27
IOA	0.55	0.82	0.79
MBE	11.93	5.09	4.41

Using the three methods, AQI values were predicted for 2011 using mathematical models obtained in MLR and PCR. In addition, statistical indicators were calculated for prediction accuracy (Table 6). The RMSE minimum value was estimated for the varimax rotation approach (in MLR 35.67, PCR with promax 28.27, and PCR with varimax 27.72). Similarly, the IOA values, which determine the best model as having a value close to 1, were 0.55 for MLR, 0.79 for PCR with promax, and 0.82 for PCR with varimax. Finally, MBE had the largest value for MLR with 11.93 and close to the values for PCR models at 5.09 and 4.41 but the least for promax. As seen from the results, the PCR models with varimax and promax performed better than the MLR models. The values from the PCR models were similar; however, the appropriate values for validation were preferable obtained using the varimax method.

5. CONCLUSION

The estimation of the AQI in Bolu, Turkey, was improved using meteorological variables. Prediction models were developed using two statistical approaches with integrated construction, such as PCA and MLR analyses. PCR and MLR were applied to the dataset to estimate the seasonal and annual AQI values. According to the predictions from the statistical processes, PCR and MLR provided similar results. Using PCA with regression models minimizes the collinearity of the datasets, which can lead to inadequate estimations, and also defines the appropriate explanatory variables for the estimation of AQI. Also, the PCR model is simple and clear because of the decrease in the number of variables. Furthermore, if we compare the PCR and MLR models, PCR has better characteristic values so that the determination coefficient values were generally higher and standard error of estimate values were lower than that of MLR.

With this method, AQI can be predicted using climate elements at a specific location. In addition, many factors, such as natural and anthropogenic emissions, industrial activities, fuel types and combustion, population density, traffic, power plants, topography, and land use, affect and degrade urban air quality. Atmospheric conditions, such as North Atlantic Oscillation (NAO), influence the Black Sea area. NAO affects climate elements, such as temperature and rainfall. This phenomenon can be studied in further research.

High levels of PRES and RH and low levels of TEMP, WS, and PREC generate high levels of pollutant concentrations, such that low WS causes an increase in air pollution by not dispersing pollutants. High-PRES prevents the air from the surrounding regions from coming into the area.

According to stabile atmospheric conditions with the PGT scheme, the dispersion of air pollutant concentrations is closely related to CLD, WS, vertical air-temperature gradient and solar radiation. SD is highly related to solar radiation and has also influences on pollutant concentrations. Hence, the existing bidirectional interactions between air pollutant concentrations and climate parameters may conclude in an increase or decrease in air quality for the atmospheric environment.

The topography, such as mountainous structures and land use may influence the climate of Bolu province which is covered by forests and woodlands. Land use is a crucial element that contributes to the production of natural dust. The increased dust concentrations during hot and dry times of year may result from it being transported from deserts, such as Saharan and Arabian. The enhanced emissions can be originated from wind transportation during times of year when heating for buildings is not necessary and from anthropogenic activities during times when heating is necessary.

The principal components with high loadings of comparable sizes on the identical variables (i.e. with analogous construction) showed up in all analyses; hence they could be assumed consistent and not sensitive to a specific selection of the rotation technique.

The accuracy of the methods was analyzed by using statistical indicators for the 2011 data values. All results were similar, but PCR with varimax rotation had the best validation results; therefore, it is more convenient and can be used as an alternative method for predicting AQI values.

Acknowledgements

The author thanks the reviewers for their suggestions that have improved the quality of this paper.

REFERENCES

Abdul-Wahab, S.A., Bakheit, C.S. & Al-Alawi, S.M., 2005. *Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations.* Environmental Modelling and Software, 20, 1263-1271.

Al-Alawi, S.M., Abdul-Wahab, S.A. & Bakheit, C.S., 2008. *Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone.* Environmental Modelling and Software, 23, 396-403.

Boznar, M., Lesjak, M. & Mlakar, P., 1993. *A neural-network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain.* Atmospheric Environment Part B-Urban Atmosphere, 27, 221-230.

Cairncross, E.K., John, J. & Zunckel, M., 2007. *A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants.* Atmospheric Environment, 41, 8442-8454.

Cheng, W-L, Chen, Y-S, Zhang, J., Lyons, T.J., Pai, J-L. & Chang, S-H., 2007. *Comparison of the Revised Air Quality Index with the PSI and AQI indices.* Science of the Total Environment, 382, 191-198.

Cogliani, E., 2001. *Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables.* Atmospheric Environment, 35, 2871-2877.

Comrie, A.C., 1997. *Comparing neural networks and regression models for ozone forecasting.* Journal of Air and Waste Management Association, 47, 653-663.

Cotta, H.H.A., Reisen V.A., Bondon, P. & Filho, P.R.P., 2020. *Identification of Redundant Air Quality Monitoring Stations using Robust Principal Component Analysis.* Environmental Modelling & Assessment, 25, 521-530.

EPA, 1993. *Selection Criteria for Mathematical Models Used in Exposure Assessments: Atmospheric Dispersion Models.* United States Environmental Protection Agency, Office of Research and Development Washington, DC 20460, EPA/600/8-91/038.

Finzi, G. & Tebaldi, G., 1982. *A mathematical model for air pollution forecast and alarm in an urban area.* Atmospheric Environment, 16, 2055-2059.

Gibergans-Baguena, J., Hervada-Sala, J. & Jarauta-Bragulat, E., 2020. *The Quality of Urban Air in Barcelona: A New Approach Applying Compositional Data analysis Methods.* Emerging Science Journal, 4(2), 113-121, doi: http://dx.doi.org/10.28991/esj-2020-01215.

Hu, J., Ying, Q., Wang, Y. & Zhang, H., 2015. *Characterizing multi-pollutant air pollution in China: Comparison of three air quality indices*, Environment International, 84, 17-25.

Ibarra-Berastegi, G., Saenz, J., Ezcurra, A., Ganzedo, U., Argandona, J.D., Errasti, I., Fernandez-Ferrero, A. & Polanco-Martinez, J., 2009. *Assessing spatial variability of SO_2 field as dedected by an air quality network using Self-Organizing Maps, cluster and*

Principal Component Analysis. Atmospheric Environment, 43, 3829-3836.

Johnson, R.A. & Wichern, D.W., 1982. *Applied Multivariete Statistical Analysis.* Prentice-Hall Inc., Englewood Cliffs, USA, 590pp.

Kanchan, K., Gorai, A.K. & Goyal, P., 2015. *A Review on Air Quality Indexing System*. Asian Journal of Atmospheric Environment, 9-2, 101-113.

Kori, R., Saxena, A., Wankhade, H., Baig, A., Kulshreshtha, A. & Sen, S., 2019a. *Ambient Air Quality Index Study of Mandideep Industrial Area, Madhya Pradesh, India.* International Journal of Research, 6(9).

Kori, R., Saxena, A., Upadhyay, N., Shrivastava, P.K., Kulshreshtha, A., Roy, S.A., Mishra, S. & Dubey, P., 2019b. *Recent Status of Ambient Air Quality of Bhopal City, Madhya Pradesh, India.* International Journal of Research & Review, 6(11).

Li, Y., Tang, Y., Fan, Z., Zhou, H. & Yang, Z., 2018. *Assessment and comparison of three different air quality indices in China*. Environ. Eng. Res., 23(1), 21-27.

Lu, W.-Z., He, H.-D. & Dong, L.-Y., 2011. *Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis.* Building and Environment, 46, 577-583.

Monteiro, A., Vieira, M., Gama, C. & Miranda, A.I., 2017. *Towards an improved air quality index.* Air Qual Atmos Health, 10, 447-455.

Rajab, J.M., Matjafri, M.Z. & Lim, H.S., 2013. *Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia.* Atmospheric Environment, 71, 36-43.

Sanchez, M.L., Casanova, J.L., Ramos, M.C. & Sanchez, J.L., 1986. *Studying the spatial and temporal distribution of SO_2 in an urban area by principal component factor analysis*. Atmospheric Research, 20, 53-65.

Sanchez, M.L., Pascual, D., Ramos, C. & Perez, I., 1990. *Forecasting particulate pollutant concentrations in a city from meteorological variables and regional weather patterns.* Atmospheric Environment Part A-General Topics, 24, 1509-1519.

Slini, T., Kaprara, A., Karatzas, K. & Moussiopoulos, N., 2006. *PM_{10} forecasting for Thessaloniki, Greece.* Environmental Modelling and Software, 21, 559-565.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M. & Pereira, M.C., 2007. *Multiple Linear Regression and artificial neural networks based on principal components to predict ozone concentrations.* Environmental Modelling and Software, 22, 97-103.

Statheropoulos, M., Vassiliadis, N. & Pappa, A., 1998. *Principal component and canonical correlation analysis for examining air pollution and meteorological data.* Atmospheric Environment, 32(6), 1087-1095.

Stevens, J., 1986. *Applied Multivariate Statistics for the Social Science*. Hillsdale, New Jersey, USA, 515pp.

The Ministry of Environment and Urbanisation (2020), http://sim.csb.gov.tr/STN/STN_Report/StationDataDownloadNew

The Ministry of Transport and Infrastructure, General Directorate of Highways (2021), https://www.kgm.gov.tr

Trivero, P., Biamino, W., Borasi, M., Cavagnero, M., Musa, M., Rinaudo, C. & Sesia, V., 2012. *An air quality balance index estimating the total amount of air pollutants at ground level,* Environ Monit Assess, 184, 4461-4472.

Turkish Statistical Institute (2020), https://www.tuik.gov.tr

U.S. Environmental Protection Agency (1999). *Guideline for reporting of daily air quality - air quality index (AQI). EPA-454/R-99-010*. Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711.

Venkatram, A., 1996. *An examination of the Pasquill-Gifford-Turner dispersion scheme.* Atmospheric Environment, 30(8), 1283-1290.

Zateroglu, M.T., 2021a. *Assessment of the effects of air pollution parameters on sunshine duration in six cities in Turkey*, Fresenius Environmental Bulletin, 30(02A), 2251-2269.

Zateroglu, M.T., 2021b. *The Role of Climate of Climate Factors on Air Pollutants (PM_{10} and SO_2),* Fresenius Environmental Bulletin, 30(11), 12029-12036.

Ziomas, I.C., Melas, D., Zerefos, C.S., Bais, A.F. & Paliatsos, A.G., 1995. *Forecasting Peak Pollutant Levels From Meteorologival Variables.* Atmospheric Environment, 29(24), 3703-3711.

Received at: 29. 09. 2021
Revised at: 29. 12. 2021
Accepted for publication at: 01. 02. 2022
Published online at: 03. 02. 2022