Scientific Research Publishing

Open Journal of Air Pollution, 2023, 12, 142-159
https://www.scirp.org/journal/ojap
ISSN Online: 2169-2661
ISSN Print: 2169-2653

Unveiling the Predictive Capabilities of Machine Learning in Air Quality Data Analysis: A Comparative Evaluation of Different Regression Models

Mosammat Mustari Khanaum[1*], Md Saidul Borhan[2], Farzana Ferdoush[3], Mohammed Ali Nause Russel[4], Mustafa Murshed[5]

[1]Department of Civil, Construction and Environmental Engineering, North Dakota State University, Fargo, USA
[2]Texas Department of Transportation, Maintenance Division, Austin, USA
[3]Department of Arts and Sciences, Ahsanullah University of Science and Technology, Dhaka, Bangladesh
[4]School of Social & Environmental Sustainability, University of Glasgow, Glasgow, UK
[5]Department of Economics, Andrew Young School of Policy Studies, Georgia State University, Atlanta, USA
Email: *mosammat.khanaum@ndsu.edu

How to cite this paper: Khanaum, M.M., Borhan, M.S., Ferdoush, F., Russel, M.A.N. and Murshed, M. (2023) Unveiling the Predictive Capabilities of Machine Learning in Air Quality Data Analysis: A Comparative Evaluation of Different Regression Models. *Open Journal of Air Pollution*, **12**, 142-159.
https://doi.org/10.4236/ojap.2023.124009

Received: October 22, 2023
Accepted: December 8, 2023
Published: December 11, 2023

Abstract

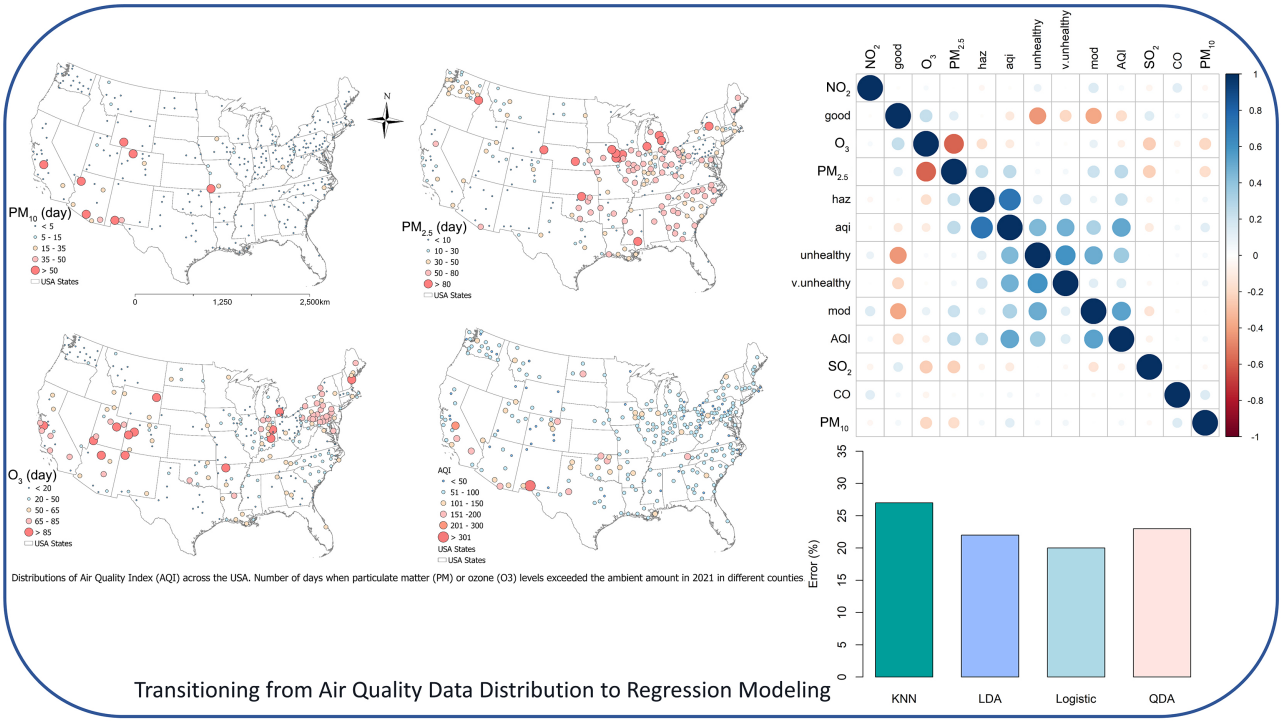
Air quality is a critical concern for public health and environmental regulation. The Air Quality Index (AQI), a widely adopted index by the US Environmental Protection Agency (EPA), serves as a crucial metric for reporting site-specific air pollution levels. Accurately predicting air quality, as measured by the AQI, is essential for effective air pollution management. In this study, we aim to identify the most reliable regression model among linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression, and K-nearest neighbors (KNN). We conducted four different regression analyses using a machine learning approach to determine the model with the best performance. By employing the confusion matrix and error percentages, we selected the best-performing model, which yielded prediction error rates of 22%, 23%, 20%, and 27%, respectively, for LDA, QDA, logistic regression, and KNN models. The logistic regression model outperformed the other three statistical models in predicting AQI. Understanding these models' performance can help address an existing gap in air quality research and contribute to the integration of regression techniques in AQI studies, ultimately benefiting stakeholders like environmental regulators, healthcare professionals, urban planners, and researchers.

Keywords

Regression Analysis, Air Quality Index, Linear Discriminant Analysis,

Quadratic Discriminant Analysis, Logistic Regression, K-Nearest Neighbors, Machine Learning, Big Data Analysis

Graphical Abstract



1. Introduction

Air pollution poses significant health and environmental challenges, demanding effective air quality prediction models. The Air Quality Index (AQI) is a critical metric for reporting pollution levels, but gaps exist in predicting rapid AQI changes. The AQI is a robust tool for communicating a specific geographical area's daily air quality status [1]. Understanding the relationship between pollutant levels and AQI values is crucial for assessing air quality and its impact on public health. The US EPA employs a system of color-coded categories and corresponding statements to depict the air quality in a given location. As mandated by the Federal Clean Air Act, the EPA continually monitors six major air pollutants: carbon monoxide, lead, particulate matter, ozone, nitrogen dioxide, and sulfur dioxide [2]. The AQI is entirely dependent on the concentrations of these six pollutants. It alerts individuals regarding the potential health impacts and enables them to adopt measures to mitigate their exposure to air pollution. Since 1976, the US EPA has been providing easily comprehensible daily reports on air quality, although formats vary across states. The US EPA utilizes AQI as the foundation for air quality forecasting and ongoing air quality reporting [3].

The levels of pollutants present in the air influence the AQI value. Six detri-

mental pollutants responsible for elevating the AQI and compromising air quality are ozone (O_3), particulate matter of two specific sizes (those with a diameter of 2.5 microns or less, $PM_{2.5}$, and those with a diameter of 10 microns or less, PM_{10}), carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). These pollutants are subject to a national air quality standard set by the US EPA to safeguard public health [4]. Notably, these pollutants pose varying degrees of harm to different segments of the population.

Different pollutants have varying harmful effects on sensitive groups. For example, O_3 poses a risk to people with lung disease, children, older adults, outdoor workers, individuals with certain genetic variants, and those with nutrient-restricted diets. $PM_{2.5}$ and PM_{10} primarily affect people with heart or lung disease, older adults, children, and individuals with lower socioeconomic status. CO poses the highest risk to people with heart disease, while NO_2 and SO_2 primarily impact individuals with asthma, children, and older adults [5]. **Table 1** provides an overview of the pollutants, their respective measuring units, and the associated adverse effects.

The AQI is categorized into six levels, each denoted by a specific color and corresponding to a distinct level of health concern. The severity of air pollution and the associated health risks rise with an increase in the AQI value (**Table 2**). For example, an AQI value of ≤50 indicates good air quality, whereas a value exceeding 300 signifies hazardous air quality [5].

AQI has been widely studied for assessing air quality and its impact on public health. Numerous studies have focused on analyzing AQI data and understanding its relationship with various pollutants and health outcomes. However, there

Table 1. Pollutants and their harmful effect (Source: US EPA, 2018).

Pollutants with Unit	Harmful Effect on Sensitive Groups
O_3 (ppm)	The group of people at the highest risk include individuals with lung disease, children, older adults, outdoor workers, those with specific genetic variants, and individuals with diets deficient in certain nutrients.
$PM_{2.5}$ (μg/m^3)	The group of people most at risk are individuals with heart or lung disease, older adults, children, and those with lower socioeconomic status.
PM_{10} (μg/m^3)	The group of people most at risk are individuals with heart or lung disease, older adults, children, and those with lower socioeconomic status.
CO (ppm)	The group of people at the highest risk consists of individuals with heart disease.
NO_2 (ppb)	The group of people at the greatest risk include individuals with asthma, children, and elderly adults.
SO_2 (ppb)	The group of people at the greatest risk include individuals with asthma, children, and elderly adults.

Table 2. Color coding, level of concern, and description for the six AQI categories (Source: AirNow, 2018).

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

is a noticeable gap in the literature regarding machine learning-driven regression analysis, particularly in the context of AQI-related research. Several studies have explored regression analysis to examine the relationship between air pollutants and AQI values [6] [7] [8]. For instance, Leung *et al.* [9] conducted a principal component regression analysis to investigate the impact of particulate matter (PM) on AQI in a metropolitan area. Similarly, Zezhou and Xiaoping [10] utilized regression models to assess the effects of multiple pollutants on AQI in urban regions.

Different regression models have proven to be highly useful in air quality studies. Linear regression models are commonly employed to examine the relationships between air pollutants and various predictors, such as toxicity data [11]. They provide valuable insights into the linear associations between variables and can help identify key factors influencing air quality. Additionally, non-linear regression models, such as polynomial regression or generalized additive models, allow for capturing more complex relationships and non-linear trends that may exist in air pollution data. These models enable a better understanding of air quality data's underlying patterns and dynamics.

Furthermore, machine learning-based regression models, including support

vector regression, random forests, and neural networks, have gained popularity due to their ability to handle high-dimensional data and capture intricate relationships [12]. These models can handle nonlinearities, interactions, and complex dependencies, enhancing the accuracy of air quality predictions [13]. Overall, the variety of regression models available provides researchers with powerful tools to analyze and predict air quality, contributing to advancing our understanding and management of air pollution.

While studies have made valuable contributions, there is a lack of research focusing on providing guidance and a machine learning code example or template for emerging practitioners interested in performing regression analysis with AQI or related datasets. This gap in the literature limits the accessibility and ease of conducting regression analyses in air quality research.

Therefore, this study aims to fill this gap by comprehensively analyzing regression methods for AQI datasets using R software. By demonstrating practical examples and providing code snippets, this research will serve as an effective resource to perform regression analysis on AQI or related data. The study will evaluate different regression models and compare their performance in predicting AQI values, thereby assisting researchers in selecting the most appropriate approach for their analyses.

Therefore, the overall objective of the study is to perform different regression analyses with the AQI dataset to find out the best regression model and then compare their results. Additionally, the research will include the R-script as an appendix to the paper.

Hence, the research question is:

- Which regression method performs the best, producing the lowest error rate for the AQI dataset?

2. Methodology

2.1. Description of Dataset

Air quality data were obtained from the website of the US EPA [14]. After the QA/QC process, a partial dataset is illustrated in **Table 3**, consisting of 520 rows and 14 columns. The dataset primarily encompassed air quality statistics for 520 cities in the United States for 2021. The 14 columns included the following information (**Table 3**): the name of the city along with the state (CBSA), the year of observation (Year), the number of days classified as "good" air quality in 2021 (good), the number of days classified as "moderate" air quality in 2021 (mod), the number of days classified as "unhealthy" air quality in 2021 (unhealthy), the number of days classified as "very unhealthy" air quality in 2021 (v.unhealthy), the number of days classified as "hazardous" air quality in 2021 (haz), the number of days when carbon monoxide (CO) levels exceeded the ambient amount in 2021 (CO), the number of days when nitrogen dioxide (NO_2) levels exceeded the ambient amount in 2021 (NO_2), the number of days when ozone (O_3) levels exceeded the ambient amount in 2021 (O_3), the number of days when sulfur dioxide

Table 3. Dataset (partial) used in the regression analysis (Source: US EPA, 2021).

CBSA	Year	Good	Mod	Unhealthy	v.unhealthy	O_3	SO_2	$PM_{2.5}$	PM_{10}	aqi
Adjuntas, PR	2021	53	4	0	0	0	0	57	0	58
Adrian, MI	2021	296	68	2	0	199	0	167	0	122
Akron, OH	2021	275	86	4	0	141	4	220	0	114
Albany, GA	2021	276	87	2	0	0	0	365	0	114
Albany-Schenectady-Troy, NY	2021	306	57	3	0	197	1	168	0	143
Albuquerque, NM	2021	131	225	9	1	202	0	81	83	214
Alexandria, LA	2021	111	9	1	0	0	0	121	0	103
Allentown-Bethlehem-Easton, PA-NJ	2021	299	66	1	0	217	0	143	0	117
Altoona, PA	2021	310	51	0	0	229	1	131	0	87
Amarillo, TX	2021	245	107	14	0	288	57	21	0	138
Americus, GA	2021	237	3	0	0	240	0	0	0	64
Anchorage, AK	2021	326	39	1	0	0	0	247	115	103

(SO_2) levels exceeded the ambient amount in 2021 (SO_2), the number of days when particulate matter $PM_{2.5}$ levels exceeded the ambient amount in 2021 ($PM_{2.5}$), the number of days when particulate matter PM_{10} levels exceeded the ambient amount in 2021 (PM_{10}), and the annual average Air Quality Index (AQI) for each city (aqi). Most of the days, in 2021, the levels of haz, CO, and NO_2 did not exceed the ambient levels, so they were not included in Table 3.

2.2. Machine Learning Algorithms with Different Regression Models

There are many applications of machine learning algorithms in the fields of big data analysis, machine learning, and classification. Each model has its advantages and assumptions. Below is a summary of the regression models utilized in this study to facilitate machine learning algorithms.

In order to distinguish between various classes, the classification technique known as linear discriminant analysis (LDA) seeks to discover a linear combination of characteristics. It assumes that each class has its covariance matrix and that the input data is usually distributed. LDA aims to project the data into a lower-dimensional space to maximize class separability and this analysis is not suitable for nonlinear class boundaries. Finding a straight line that optimizes the distance between the average of each class while minimizing the total error is the goal when categorizing the given data into the number of classes [15]. This supervised machine learning technique seeks to find linear combinations of features that maximize the separation between multiple classes, making it a valuable tool for pattern recognition and data preprocessing. LDA aims to reduce dimensionality while preserving class discrimination as much as possible, resulting in

more efficient and accurate classification models. LDA defines the discriminant function as follows [16]:

$$\delta_k(x) = x^{\mathrm{T}}\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^{\mathrm{T}}\Sigma_k^{-1}\mu_k + \log \pi_k \tag{1}$$

where, k is the class, x is the set of measurements, μ_k is the mean vector, π_k is the prior probability and $\sum_k$ is the covariance matrix. $\delta_k(x)$ is a linear function in x.

Quadratic discriminant analysis (QDA) is similar to LDA but relaxes the assumption of equal covariance matrices across classes [15]. QDA, on the other hand, permits each class to have a unique covariance matrix, which may result in more adaptable decision boundaries. When the data cannot be separated linearly, QDA is helpful. In the event of variance-covariance heterogeneity, QDA is a more suitable approach [17]. QDA is a generalized version of LDA when the feature distributions within the two classes are generally distributed with potentially different covariance matrices. In QDA, a subspace is determined, and within this subspace, quadratic separating surfaces are employed to effectively separate the classes. The discriminant function for QDA is formulated as follows [16]:

$$\delta_k(x) = -\frac{1}{2}x^{\mathrm{T}}\Sigma_k^{-1}x + x^{\mathrm{T}}\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^{\mathrm{T}}\Sigma_k^{-1}\mu_k + \log \pi_k \tag{2}$$

Logistic regression is a classification algorithm used for establishing the relationship between a dependent variable (binary) and one or more independent variables. It is a statistical technique that deals with multiple variables, where the outcome (response) variable is either categorical or ordinal, and the independent variables can be of various types, such as continuous, discrete, categorical, or ordinal [18]. It is not suitable for analyzing the complex nonlinear relationships. It uses a function to calculate the likelihood of dependent variables being part of a specific class. Logistic regression is widely applied in various fields, and its coefficients provide insights into the importance of each independent variable. The function for logistic regression as follows [16]:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{3}$$

where β_0 and β_1 are the coefficients.

K-Nearest Neighbors (KNN) is a non-parametric classification algorithm that classifies objects based on their proximity to the training data [19]. It is sensitive to outliers and is not suitable for high-dimensional data. It assigns a new data point to the most common class among its k nearest neighbors. Though it has a simple algorithm, the k parameter and distance measure choice can impact how well it performs. Small K values might result in overfitting, which means that the model overemphasizes the unique properties of the data in each region. However, if K is set to a very high value, the model may become overly regularized and underfit, unable to recognize the underlying patterns [20]. The value of the k nearest neighbors to predict the value of the data point uses the following func-

tion [16]:

$$\underset{C_1,\cdots,C_K}{\text{minimize}}\left\{\sum_{k=1}^{K} W\left(C_k\right)\right\} \tag{4}$$

where C_k is cluster, K is the number of clusters, $W\left(C_k\right)$ is the measure of C_k.

2.3. Regression Analysis with R Studio

Different regression models were run and compared to determine the most effective regression model with the lowest error rate. The "aqi" (Air Quality Index) was used as the response variable, and 12 parameters (excluding city name and year) were utilized as predictor variables to identify the optimal regression model for this dataset. Subsequently, the results of the selected models were compared.

The necessary R packages, MASS [21] and ISLR [22], required for the analyses, were then installed/called on the R platform. A new binary variable, AQI, was created based on the existing air quality index (aqi) data. The procedure adopted for that purpose is shown in Equations (1) and (2): assigning a value of 1 to AQI when it exceeded the median AQI and 0 when it was less than or equal to the median AQI. The dataset was then imported into R Studio, and basic analyses were conducted to determine the number of days the air quality was categorized as good, unhealthy, very unhealthy, or hazardous.

$$\text{AQI} = 1 \text{ when aqi} > \text{aqi}_{\text{median}} \tag{5}$$

$$\text{AQI} = 0 \text{ when aqi} \le \text{aqi}_{\text{median}} \tag{6}$$

The subsequent step involved regression modeling, starting with a full linear regression analysis and summarizing the fitted model. The correlation between the AQI (response variable) and other significant predictors was examined to identify the most closely associated variable in the model. A test was conducted to assess whether the variables followed a normal distribution and whether they exhibited a linear or quadratic distribution.

At this point, the dataset was divided into two parts based on the presence of ozone in the air. The training data consisted of observations with an ozone concentration of less than 200 DU, while the remaining data were designated as testing data. Four different regression analyses, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression, and K-Nearest Neighbors (KNN), were conducted on the dataset. Subsequently, a comparative analysis assessed the performance of these models based on test error or mean squared error (MSE).

The dataset was divided into two subsets to conduct different regression modeling analyses. The training data consisted of ozone concentrations below 200 DU, while the remaining data was utilized as the testing set. The ozone concentration of 200 DU has been used as the dividing point for testing and training because it is a big difference from the normal ozone levels in the air, which are about 300 DU, and the lower levels in places like the Antarctic Ozone "Hole",

where the level is about 100 DU on average [23].

R Studio [24] provides a convenient and visually appealing platform for data visualization. It facilitates the effortless calculation of critical values, R^2 values, t-values, F-values, regression analysis, test errors, and other essential model performance metrics. With its simplicity and affordability, R Studio is an efficient tool for regression modeling. Moreover, it requires only a few packages, which are freely available and easily installable within the platform. R Studio seamlessly generates correlation matrices, confusion matrices, summary statistics for regression analysis, and prediction rates and error rates. Its user-friendly interface simplifies implementing various regression methods, ultimately saving significant analysis and computation time. Considering these advantages, the study employed the R platform (R Studio/2023.03.0+386) to analyze the air quality index dataset.

3. Results

3.1. Information on Air Quality across Counties

In the year 2021, a total of only 20 counties in the United States were observed as Good, with an air quality index equal to or less than 50 (AQI ≤ 50). On the other hand, 195 counties had a Moderate, AQI ranging from 51 to 100 (**Figure 1**). Furthermore, 147 counties with an AQI categorized as Unhealthy for Sensitive Groups ranged from 101 to 150. Additionally, 92 counties had an Unhealthy AQI falling within the range of 151 to 200. Moreover, 27 counties were classified as having a Very Unhealthy AQI with values ranging from 201 to 300. Lastly, 39 counties had a Hazardous AQI (AQI > 301).

A multiple regression analysis was conducted, and it was determined that four predictors, namely "mod," "haz," "aqi," and "v.unhealthy," exhibited statistical significance. Since the response variable was based on AQI, a decision was made to proceed with a regression model utilizing three of these significant predictors: "mod," "haz," and "v.unhealthy." The following represents the full (Equation (3)) and fitted models (Equation (4)) of the fitted multiple regression analysis (**Table 4**).

$$\begin{aligned}\text{AQI} = &-0.07562 + 0.00361 \times \text{mod} - 0.02604 \times \text{v.unhealthy} \\ &- 0.01119 \times \text{haz} + 0.0024 \times \text{aqi}\end{aligned} \tag{7}$$

$$Y = \beta_0 + \beta_1(\text{mod}) + \beta_2(\text{haz}) + \beta_3(\text{v.unhealthy}) + \varepsilon \tag{8}$$

where Y represents the response variable, β_0 represents the intercept, β_1, β_2, and β_3 represent the coefficients associated with the predictors "mod," "haz," and "v.unhealthy," respectively, and ε represents the error term.

The *cor*() and *plot*() functions were employed to examine the correlation between the AQI and other variables. The correlation analysis revealed noteworthy findings. Specifically, a strong positive correlation was observed between the AQI and the variable representing moderate days (mod). Furthermore, a moderately positive correlation was identified between AQI and the variables

Table 4. Full multiple linear regression model for AQI forecasting.

Variable	Estimate	Standard Error	t-Statistic	p-Value
(Intercept)	−0.0756	0.0785	−0.9630	0.3360
Good	−0.0001	0.0004	−0.2840	0.7770
Mod	0.0036	0.0006	6.4930	0.0000***
Unhealthy	0.0009	0.0016	0.5740	0.5660
v.unhealthy	−0.0260	0.0064	−4.0730	0.0001***
haz	−0.1119	0.0239	−4.6790	0.0000***
CO	0.0004	0.0053	0.0850	0.9330
NO_2	−0.0007	0.0018	−0.4110	0.6810
O_3	0.0000	0.0003	−0.0400	0.9680
SO_2	0.0005	0.0004	1.0940	0.2740
PM2.5	0.0005	0.0003	1.4140	0.1580
PM10	NA	NA	NA	NA
aqi	0.0024	0.0002	10.3140	0.0000***

***level of significance, $p < 0.001$.

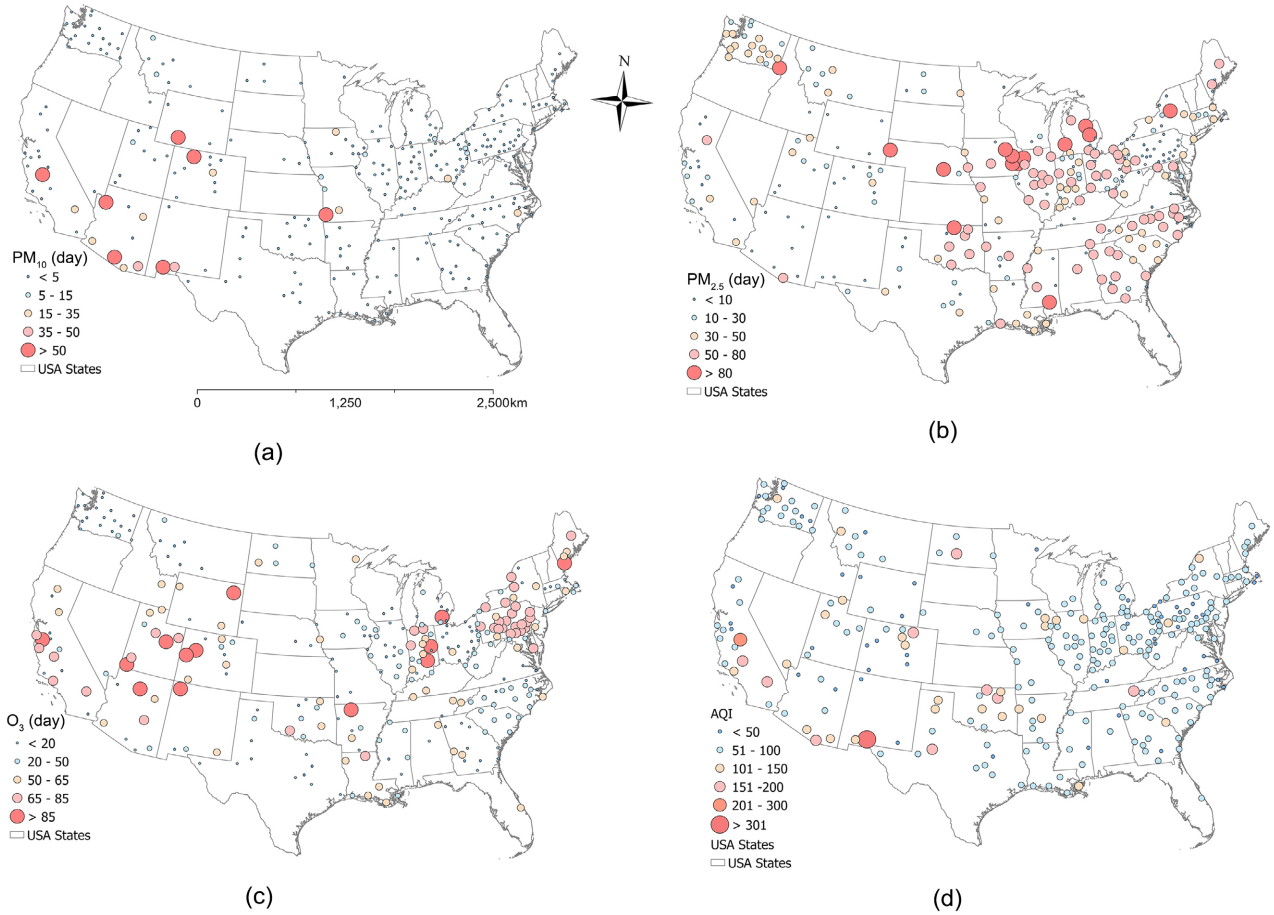


Figure 1. Number of days when (a) particulate matter PM10, (b) particulate matter PM2.5, (c) ozone (O3) levels exceeded the ambient amount in 2021 in different counties, and (d) distributions of Air Quality Index (AQI) across the USA.

representing unhealthy days and $PM_{2.5}$ levels. Conversely, a moderately negative association was found between AQI and the variables representing good air quality, ozone (O_3) levels, and sulfur dioxide (SO_2) levels. These results are summarized in (Figure 2).

A multiple regression analysis was conducted, and it was determined that four predictors, namely "mod", "haz", "aqi" and "v.unhealthy", exhibited statistical significance.

However, the pairwise plot was not informative due to the binary nature of the AQI variable, which only includes values of 0 or 1 (Figure 3). Since the AQI is a binary air quality classification, it represents either "good" or "not good" air quality conditions. Consequently, pairwise plotting the AQI against other variables does not yield meaningful insights or reveal any discernible patterns.

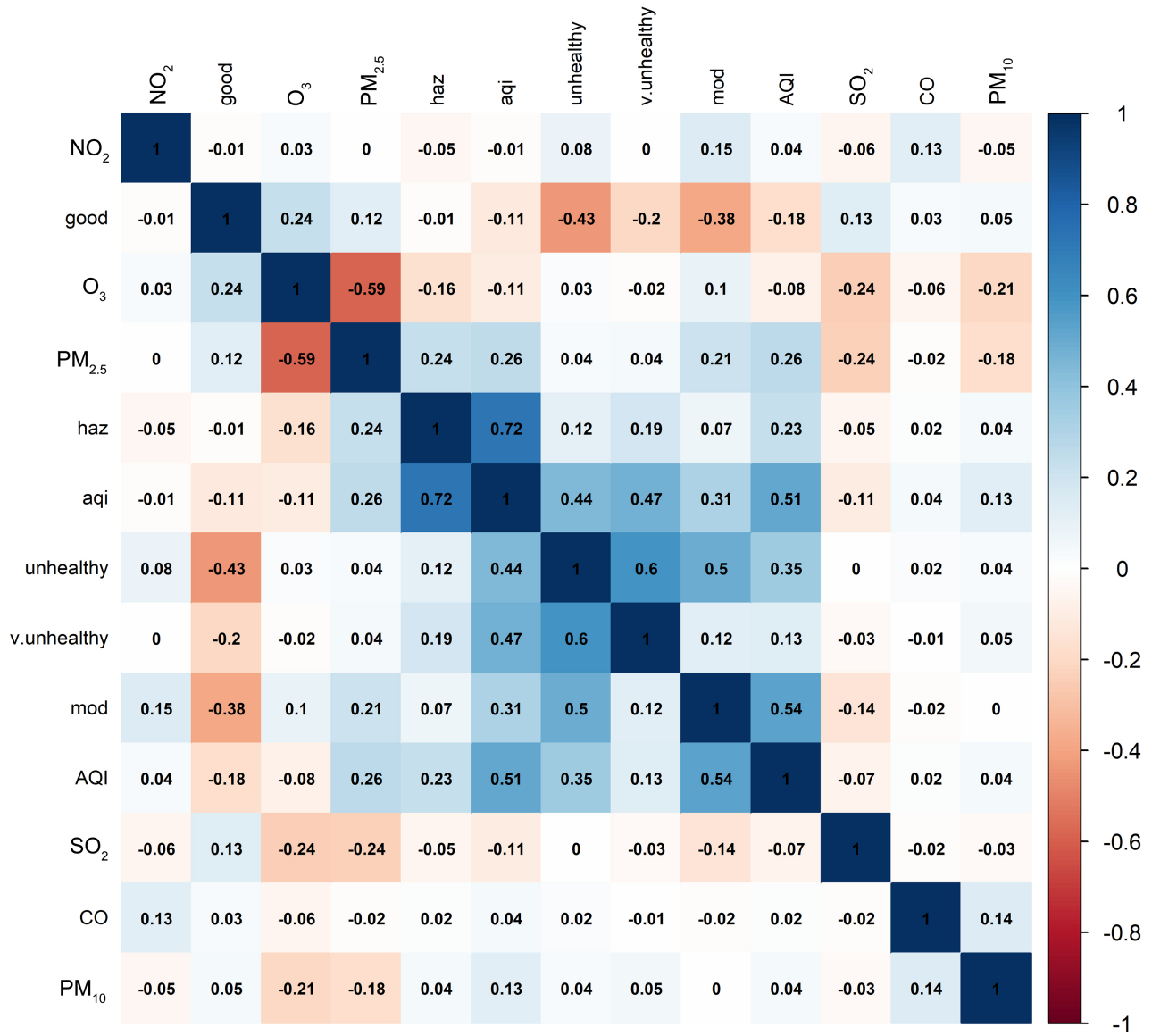


Figure 2. The pearson correlation coefficient between different variables.

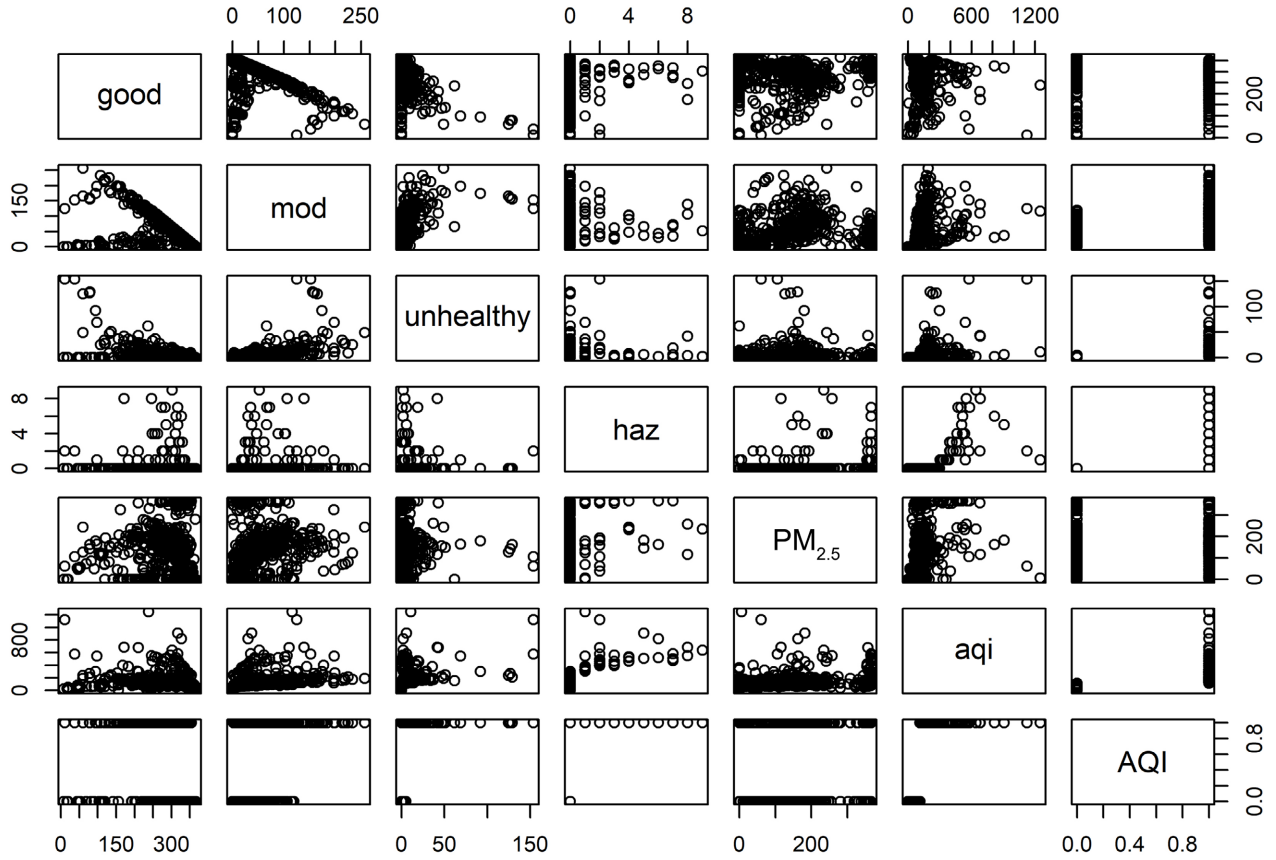


Figure 3. Pairwise plot of different variables.

To better understand the relationship between the AQI and other variables, alternative analytical methods tailored for binary variables, such as logistic regression or categorical analysis, would be more appropriate. These techniques are specifically designed to handle binary response variables and can provide a more comprehensive analysis of the relationships between the AQI and the predictors.

The box plots demonstrate a vivid relationship between the number of unhealthy and moderate days and the AQI. Specifically, as the number of unhealthy and moderate days increases, the AQI also increases. Additionally, there is a positive relationship between AQI and $PM_{2.5}$, with higher AQI values indicating lower $PM_{2.5}$ concentrations (Figure 4). Evidence of a negative association between the response variable and three independent variables is also observed. For instance, an increase in the total number of good days corresponds to lower AQI values. Similarly, there is a negative association between AQI and O_3 as well as AQI and SO_2, it is worth noting that the box and whisker plot reveals numerous outliers for the SO_2 variable (Figure 4).

3.2. Comparing Different Regression Models

1) LDA: Based on the analysis of the associated variables, including mod, v.unhealthy, and haz, these three variables were utilized to predict the response

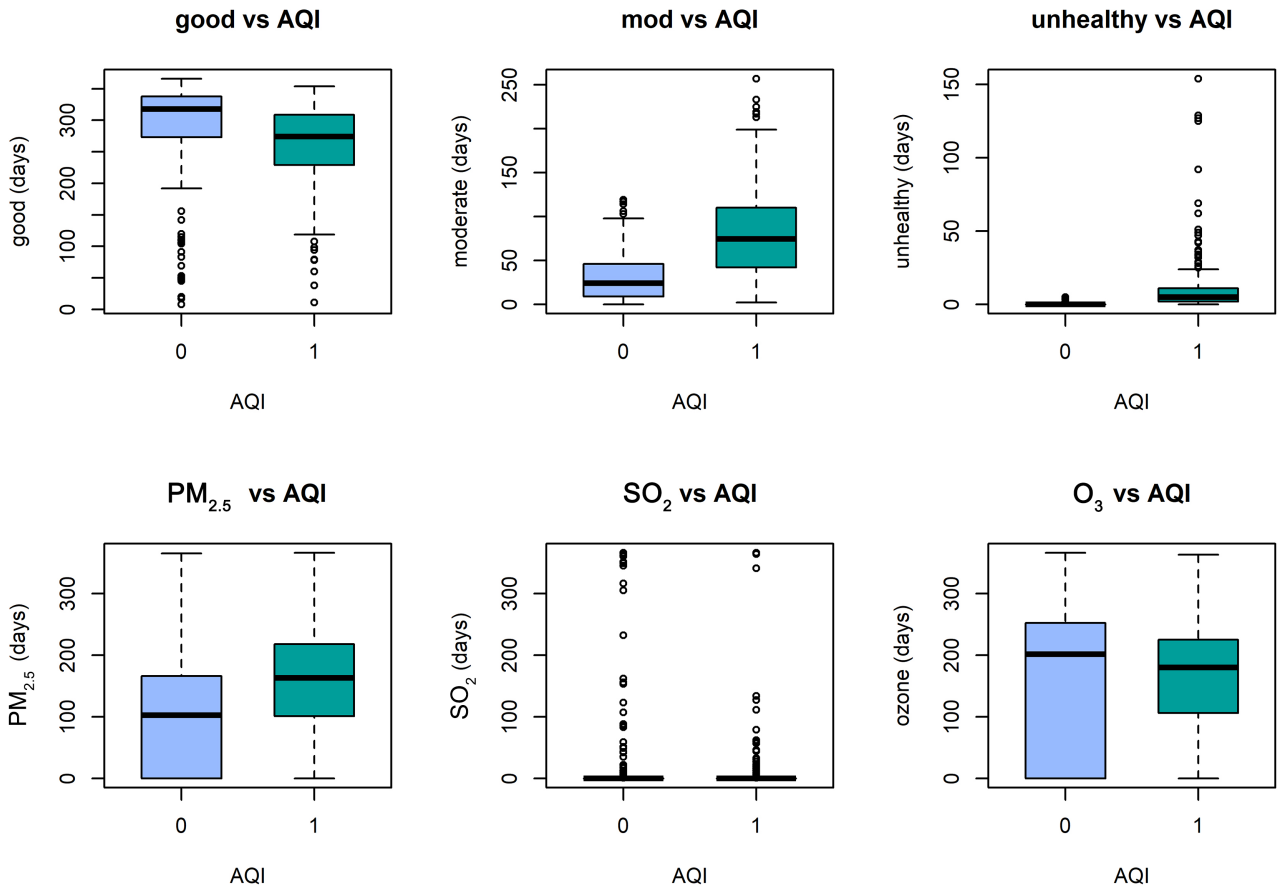


Figure 4. Box and whisker plots of response and different predictors.

variable, AQI. The resulting confusion matrix of the LDA model is presented in Table 5(a). The LDA model exhibits a test error rate of 22%, indicating a relatively low level of misclassification. However, the overall correction rate of 78% suggests a moderate level of accuracy in predicting the AQI based on the selected variables. Alvarez-Guerra *et al.* observed LDA as a promising regression model for predicting toxicity levels as an environmental problem [11].

2) QDA: Table 5(b) displays the confusion matrix for the QDA model. The test error of the QDA model is determined to be 23%, which is slightly higher than the test error rate observed for the LDA model. This indicates that the QDA model has a correction rate of 77%. Compared to the LDA model, the QDA model exhibits a slightly higher level of misclassification in predicting the AQI based on the selected variables. Similar error rates were observed with the QDA model by Alvarez-Guerra *et al.* [11]. The findings indicated that the performance of the QDA model was comparable to other approaches.

3) Logistic regression model: The confusion matrix for the logistic regression model is presented in Table 5(c). The test error of this model is calculated to be 20%, which is lower than the LDA and QDA models. This implies that the correction rate for the logistic regression model is 80%, indicating a relatively good accuracy in predicting the *AQI* based on the selected variables. The findings of

Table 5. Confusion matrix of (a) LDA, (b) QDA, and (d) Logistic regression model.

(a)

lda.class	test.data	
	0	1
0	126	41
1	11	57

(b)

qda.class	test.data	
	0	1
0	125	41
1	12	57

(c)

glm.pred	test.data	
	0	1
0	126	37
1	11	61

this study align with the research conducted by Thach *et al.* [25], who employed a conditional logistic regression model to examine air pollution profiles [23]. The study provided further support for the effectiveness of regression models, particularly logistic regression, in analyzing air quality data and predicting relevant outcomes. In contrast to the LDA and QDA models, the logistic regression model exhibits a slightly lower test error rate, further supporting its effectiveness in this analysis. Similarly, Li *et al.* [26] successfully applied the logistic regression model to extract latent representations of air quality features from air quality data, specifically focusing on capturing non-linear spatial and temporal correlations [24].

4) KNN: **Table 6** illustrates the confusion matrix for the K-nearest neighbors (KNN) model with various k-values. The test error of the KNN model at the optimal k-value, k = 5, is found to be 27%, signifying the highest level of performance compared to other k-values. However, it should be noted that the test error for the KNN model is higher than that of all other regression models under consideration. Dragomir (2010) noted different results than this study, where they noted that KNN is a good model for predicting air quality [13]. The reason for the discrepancy between our results and the other study [13] may be due to differences in the specific dataset used, the selection of input features, or adjustments in the experimental setup, all of which may have affected the KNN model performance assessment and suitability for the air quality data used in this study.

Figure 5 shows the prediction error (%) for each of the four regression models mentioned above, with the KNN model representing the optimal k-value.

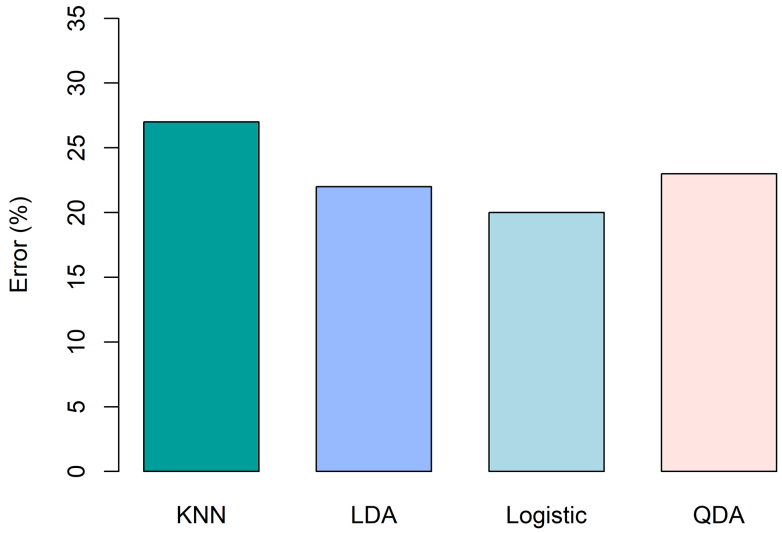


Figure 5. Comparing prediction error (%) for various regression models. The K-nearest neighbors (KNN) model represents the optimal k-value.

Table 6. Confusion matrix of KNN.

K value	Confusion Matrix			Test Error
1		test.data		31%
	knn.pred1	0	1	
	0	101	37	
	1	36	61	
3		test.data		28%
	knn.pred3	0	1	
	0	114	43	
	1	23	55	
5 optimum k		test.data		27%
	knn.pred5	0	1	
	0	113	39	
	1	24	59	
7		test.data		28%
	knn.pred7	0	1	
	0	114	43	
	1	23	55	
9		test.data		30%
	knn.pred9	0	1	
	0	107	41	
	1	30	57	

Based on the observations derived from the figure, it is evident that among the four regression models examined, the logistic regression model demonstrated the highest level of efficiency, exhibiting the lowest error percentage, followed by the LDA, QDA, and KNN regression models.

4. Conclusions

Three statistically significant variables, namely "mod", "haz", and "v.unhealthy", were utilized to build the model for predicting the response variable, AQI. Four different regression models were applied to the dataset, with ozone concentrations below 200 DU serving as the training data, while the remaining data was utilized for testing purposes.

The results of the regression analysis yielded a test error rate of 22% for the LDA model, 23% for the QDA model, and 20% for the logistic regression model; however, for the KNN model with the optimal k-value (k = 5), the error rate was 27%. The logistic regression model demonstrated the best performance among these models, exhibiting the lowest test error rate of 20%. Conversely, the KNN model produced the highest test error rate of 31%.

Overall, the logistic regression model performed better for air quality or other related datasets, as it demonstrated better performance by yielding the lowest test error rate. This research addresses an existing gap in air quality research by providing practical regression analysis tools and contributing to the integration of regression techniques in AQI studies.

The study has identified valuable insights into air quality prediction through regression analysis. As we look toward future studies, several areas are worth exploring. Firstly, one potential study could consider additional environmental variables, and advanced modeling techniques could lead to even more accurate AQI predictions. Furthermore, this study did not consider the temporal and seasonal variations on air quality index, which could be a potential future investigation.

The outcomes of this study will not only assist researchers in conducting their analyses but also promote consistency and reproducibility in forthcoming studies related to AQI. It will offer practical tools for air quality prediction, serving public health and urban development while providing insights into data analysis for stakeholders, including environmental regulators, healthcare professionals, urban planners, and researchers.

Data Availability

The datasets and R code used in the current study are available as supplementary materials.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Cromar, K.R., Ghazipura, M., Gladson, L.A. and Perlmutt, L. (2020) Evaluating the US Air Quality Index as a Risk Communication Tool: Comparing Associations of Index Values with Respiratory Morbidity among Adults in California. *PLOS ONE*, **15**, e0242031. https://doi.org/10.1371/journal.pone.0242031

[2] Declet-Barreto, J., Goldman, G.T., Desikan, A., Berman, E., Goldman, J., Johnson, C., Rosenberg, A.A., *et al.* (2020) Hazardous Air Pollutant Emissions Implications Under 2018 Guidance on US Clean Air Act Requirements for Major Sources. *Journal of the Air & Waste Management Association*, **70**, 481-490. https://doi.org/10.1080/10962247.2020.1735575

[3] United States Environmental Protection Agency (2018) Technical Assistance Document for the Reporting of Daily Air Quality—The Air Quality Index (AQI). Office of Air Quality Planning and Standards Air Quality Assessment Division Research Triangle Park, NC. https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf

[4] Lumb, A., Sharma, T.C., Bibeault, J.F. and Klawunn, P. (2012) A Comparative Study of USA and Canadian Water Quality Index Models. *Water Quality, Exposure and Health*, **3**, 203-216. https://doi.org/10.1007/s12403-011-0056-5

[5] AirNow (2018) Basic Information on the AQI in English. https://www.airnow.gov/aqi/aqi-basics/

[6] Cairncross, E.K., John, J. and 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. https://doi.org/10.1016/j.atmosenv.2007.07.003

[7] Kyrkilis, G., Chaloulakou, A. and Kassomenos, P.A. (2007) Development of an Aggregate Air Quality Index for an Urban Mediterranean Agglomeration: Relation to Potential Health Effects. *Environment International*, **33**, 670-676. https://doi.org/10.1016/j.envint.2007.01.010

[8] Hu, J., Ying, Q., Wang, Y. and Zhang, H. (2015) Characterizing Multi-Pollutant Air Pollution in China: Comparison of Three Air Quality Indices. *Environment International*, **84**, 17-25. https://doi.org/10.1016/j.envint.2015.06.014

[9] Leung, D.M., Tai, A.P., Mickley, L.J., Moch, J.M., van Donkelaar, A., Shen, L. and Martin, R.V. (2018) Synoptic Meteorological Modes of Variability for Fine Particulate Matter ($PM_{2.5}$) Air Quality in Major Metropolitan Regions of China. *Atmospheric Chemistry and Physics*, **18**, 6733-6748. https://doi.org/10.5194/acp-18-6733-2018

[10] Gong, Z.Z. and Zhang, X.P. (2017) Assessment of Urban Air Pollution and Spatial Spillover Effects in China: Cases of 113 Key Environmental Protection Cities. *Journal of Resources and Ecology*, **8**, 584-594. https://doi.org/10.5814/j.issn.1674-764x.2017.06.004

[11] Alvarez-Guerra, M., Ballabio, D., Amigo, J.M., Viguri, J.R. and Bro, R. (2010) A Chemometric Approach to the Environmental Problem of Predicting Toxicity in Contaminated Sediments. *Journal of Chemometrics*, **24**, 379-386. https://doi.org/10.1002/cem.1264

[12] Srivastava, C., Singh, S. and Singh, A.P. (2018) Estimation of Air Pollution in Delhi Using Machine Learning Techniques. 2018 *International Conference on Computing, Power and Communication Technologies* (*GUCON*), Greater Noida, 28-29 September 2018, 304-309. https://doi.org/10.1109/GUCON.2018.8675022

[13] Dragomir, E.G. (2010) Air Quality Index Prediction Using K-Nearest Neighbor Technique. *Bulletin of PG University of Ploiesti*, *Series Mathematics*, *Informatics*, *Physics*, **LXII**, 103-108.

[14] United States Environmental Protection Agency (2021) Air Data: Air Quality Data Collected at Outdoor Monitors Across the US. https://www.epa.gov/outdoor-air-quality-data

[15] Choi, B.G., Rha, S.W., Kim, S.W., Kang, J.H., Park, J.Y. and Noh, Y.K. (2019) Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-Up in Non-Diabetic Patients with Cardiovascular Risks. *Yonsei Medical Journal*, **60**, 191-199. https://doi.org/10.3349/ymj.2019.60.2.191

[16] Witten, D. and James, G. (2013) An Introduction to Statistical Learning: With Applications in R. Springer, New York.

[17] Huberty, C.J. and Olejnik, S. (2006) Applied MANOVA and Discriminant Analysis. John Wiley & Sons, New York. https://doi.org/10.1002/047178947X

[18] Rencher, A.C. and Schimek, M.G. (1997) Methods of Multivariate Analysis. *Computational Statistics*, **12**, 422.

[19] Fix, E. and Hodges, J.L. (1951) Discriminatory Analysis, Non-Parametric Discrimination. *International Statistical Review*, **57**, 238-247. https://doi.org/10.1037/e471672008-001

[20] Ziegel, E.R. (2001) Multivariate Data Reduction and Discrimination with SAS Software. *Technometrics*, **43**, 248-249. https://doi.org/10.1198/tech.2001.s616

[21] Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D. and Ripley, M.B. (2013) Package 'Mass'. *Cran R*, **538**, 113-120.

[22] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017) Data for an Introduction to Statistical Learning with Applications in R. Package 'ISLR'. CRAN.

[23] NASA Ozone Watch (2023) Images, Data, and Information for Atmospheric Ozone. https://ozonewatch.gsfc.nasa.gov/facts/dobson_SH.html#:~:text=The%20average%20amount%20of%20ozone,of%20about%20100%20Dobson%20Units

[24] RStudio Team (2020) RStudio: Integrated Development for R. Boston. http://www.rstudio.com/

[25] Thach, T.Q., Tsang, H., Cao, P. and Ho, L.M. (2018) A Novel Method to Construct an Air Quality Index Based on Air Pollution Profiles. *International Journal of Hygiene and Environmental Health*, **221**, 17-26. https://doi.org/10.1016/j.ijheh.2017.09.012

[26] Li, X., Peng, L., Hu, Y., Shao, J. and Chi, T. (2016) Deep Learning Architecture for Air Quality Predictions. *Environmental Science and Pollution Research*, **23**, 22408-22417. https://doi.org/10.1007/s11356-016-7812-9