RESEARCH PAPER

LONG-TERM AIR QUALITY EVALUATION SYSTEM PREDICTION IN CHINA BASED ON MULTINOMIAL LOGISTIC REGRESSION METHOD

Yang. He[1]*, Dongfang. Qi[1], V M. Bure[1,2]

[1]St. Petersburg State University, 7-9, Universitetskaya nab., St Petersburg, 199034, Russian Federation

[2]Agrophysical Research Institute, 14, Grazhdanskiy pr, St Petersburg, 195220, Russian Federation

***Corresponding author:** st082131@student.spbu.ru**

Received: January 18th, 2023 / Accepted: November 14th, 2023 / Published: December 31st, 2023

https://DOI-10.24057/2071-9388-2023-2719

ABSTRACT. The aim of this article evaluate the long-term air quality in China based on the air quality index (AQI) and the air quality composite index (AQCI) though the multinomial logistic regression method. The two developed models employ different dependent variables, AQI and AQCI, while maintaining the same controlled variables gross domestic product (GDP), and a primary pollutant. Explicitly, the primary impurity is associated with one or more contaminants among six pollutant factors: O_3, $PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO. Model quality verification is an integral part of our analysis. The results are illustrate d using real air quality data from China. The developed models were applied to predict AQI and ACQI for the 31 capital cities in China from 2013 to 2019 annually. All calculations and tests are conducted using R-studio. In summary, both models are able to predict China's long-term air quality. A comparison of the AQI and AQCI models using the ROC curve reveals that the AQCI model exhibits greater significance than the AQI model.

KEYWORDS: Multinomial logistic regression, Air Quality Index, Air Quality Composite Index, ROC curve

CITATION: Yang. He, Dongfang. Qi, V M. Bure (2023). Long-Term Air Quality Evaluation System Prediction In China Based On Multinomial Logistic Regression Method. Geography, Environment, Sustainability, 4(16), 164-171
https://DOI-10.24057/2071-9388-2023-2719

ACKNOWLEDGEMENTS: Authors are highly thankful to two anonymous reviewers and the principal editor for their positive constructive advices for the manuscript.

Conflict of interests: The authors reported no potential conflict of interest.

INTRODUCTION

In China, with the continuous improvement in people's quality of life, there is a growing awareness of their living environment (Fann N, et al.2013). In March 2012, the State Environmental Protection Administration of China established the National Ambient Air Quality Standard and introduced a new air quality evaluation standard for public health, known as the Air Quality Index (AQI). The AQI is derived from various pollutants, including particulate matter 2.5 ($PM_{2.5}$), inhalable particulate matter (PM_{10}), ozone (O_3), carbon monoxide (CO), nitrogen oxides (NO_2), sulfur dioxide (SO_2), among others. Government agencies use the AQI to communicate current or forecasted pollution levels to the public (Wang K, et al. 2019). In 2016, the State Environmental Protection Administration introduced another air quality evaluation method, the Comprehensive Air Quality Index (AQCI) (Wang S, et al. 2012). Both indexes have become widely popular for quick air quality assessments in China. Higher AQI or AQCI values are associated with improved public health. The World Health Organization states that 9 out of 10 people worldwide breathe polluted air, which contributes to more than 4.2 million deaths per year, and overall negatively affects the population's health, especially children and elderly (Schachter E N, Moshier E, Habre R, et al. 2016). Therefore, an accurate air pollution prediction system can help government agencies inform the public about air pollution levels (Zaib S, et al. 2022).

To the best of the authors' knowledge, the methods outlined in the referenced papers prove effective in air quality prediction problems, typically categorized into deterministic and uncertainty models. Deterministic models, are usually based on traditional statistic or economic models, such as multiple linear regression, multiple logistic regression, and time series model (Bure V. M. et al. 2007; 2013; 2019, Iakushev V. P. et al. 2020, 2021). Uncertainty models employ the state of the art machine learning techniques, and return the probability of different air quality levels. Data-driven methods, like deep learning, and ensemble learning, generally outperform, deterministic models, as reported by (R. Stern, P. Builtjes, M. Schaap, R. Timmermans, R. Vautard, A. Hodzic, M. Memmesheimer, H. Feldmann, E. Renner, R. Wolke, et al. 2008).

Moreover, (Karimian H, Li Q, Wu C, et al. 2019) explore three models: multiple additive regression trees (MART), deep feed-forward neural network (DFNN), and a new hybrid model based on long-term memory (LSTM), with LSTM emerging as the highest-performing one. (Di Q, Amini H, Shi L, et al. 2019) incorporate the geographic factor as a controlled variable, combining $PM_{2.5}$ to construct a neural network using random forest and gradient boosting. (Li X, Peng L, Hu Y, et al. 2016) employ LSTM, convolutional neural networks (CNN), and one-dimensional convolutional

neural network (ID-CNN), achieving 78% accuracy in air pollution predictions. Additionally, (Li X, Peng L, Hu Y, et al.2016), combine CNN and LSTM to create a CNN-LSTM model for predicting $PM_{2.5}$ concentrations in any capital city in China. (Tong W, Li L, Zhou X, et al. 2019) take a unique approach, transforming air quality data into sequences of images using the Conv-LSTM model to interpolate the predicted air quality data for the entire cities, demonstrated over Seoul City in Korea. (Nadeem I., Ilyas A.M., Uduman P.S) employ an ARMA/ARIMA modelling approach for forecasting Respirable Suspended Particular Matter (RSPM), Sulphur dioxide (SO_2), and Nitrogen dioxide (NO_2) concentrations in Chennai City, India. (Senarathna M, et al.) utilize intelligent sensor technology to detect $PM_{2.5}$ and NO_2 in Kandy City, Sri Lanka.

Furthermore, it is intriguing to determine the model for predicting air quality, especially considering that most research focus on short-term prediction per day or per hour (Stojov V, Koteli N, Lameski P, et al.2018, Tao Q, et at al. 2019, Le V D, et at al. 2020). Therefore, proposing a new model for long-term air quality prediction in China becomes necessary (He Y, et al. 2023). Additionally, there is another significant question: do economic factors, such as GDP, affect air quality? In contrast to previous research, this study considers essential pollution items and combines them with financial aspects. In this paper, two new models based on the multinomial logistic regression algorithm are constructed, classifying 31 capital cities into different GDP states, (high, medium, and low), and combining polluting factors as controlled variables and AQI and AQCI as dependent variables. The dataset contains air quality information on 31 capital cities in China, including six air pollutants, (O_3, $PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO.) per year from 2013-2019. In summary, both models provide essential outcomes that can be used for air quality prediction and assessment.

This contribution is organized as follows: Section 2 details AQI and AQCI calculation. Section 3 focuses on Date and economic model, while Section 4 and 5, present Experiment results and draw conclusions.

AQI AND AQCI CALCULATION

The calculation of AQI in China follows these steps:

$$AQI = max\left\{IAQI_1, IAQI_2, IAQI_3, IAQI_4, IAQI_5, IAQI_6,\right\}$$

Where AQI is the air quality index, $IAQI_i$ is the individual air quality index for each of the pollutants (i from 1 to 6, 1 for SO_2, 2 for NO_2, 3 for PM_{10}, 4 for CO, 5 for O_3, 6 for $PM_{2.5}$). When one of these six pollutants reaches its maximum value, and resulting AQI is above 50, that specific pollutant becomes the primary contributor to pollution.

The individual air quality index (IAQI) for each pollutant

$$IAQI = \frac{IAQI_{h_i} - IAQI_{l_o}}{BP_{h_i} - BP_{l_0}}\left(C_i - BP_{l_o}\right) + IAQI_{l_o}$$

is calculated using the formula:

Where $IAQI_i$ is the Sub Air Quality Index for Pollutant Project i, i represents the pollutant item. C_i is the mass concentration value of the pollutant item i.

BP_{h_i} is the high value of the pollutant concentration limit close to C_i in the concentration limit table (Table 1) corresponding to the air quality sub-index. BP_{l_0} is the low value of the pollution concentration limit close to C_i in the concentration limit table corresponding to the air quality sub-index. $IAQI_{h_i}$ is the air quality sub-index corresponding to BP_{h_i} in the concentration limit table corresponding to the air quality sub-index. $IAQI_{l_o}$ is the air quality sub-index corresponding to BP_{l_o} in the concentration limit table corresponding to the air quality sub-index.

China air quality composite index:

The China Air Quality Composite Index encompasses the pollutants considered in the evaluation during the assessment period. The comprehensive index is determined by summing the individual quality indexes, and a higher value indicates a greater degree of urban air pollution. This index is calculated using the concentrations of six pollutants but with different weighing factors. The key pollutants involved in the air quality composite index assessment are O_3, $PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO.

$$I_i = \frac{C_i}{S_i}$$

Individual China air quality Index:

Where C_i is the concentration value of index i. S_i is the secondary standard value of Index i. The index i can be SO_2, NO_2, PM_{10}, or $PM_{2.5}$, each with its respective secondary standard limit for the annual average concentration; For O_3, S_i represent the biggest 8 hours average secondary standard limit; For CO, S_i signifies the level 2 standard for the quasi-limit of daily average concentration, as detailed in Table 2.

$$I_{sum} = \sum_{i=1}^{6} I_i$$

Where I_{sum} is the China Air Quality Composite Index, and I_i is the individual index for indicator i, covering all

Table 1. Pollution item concentration limit table

IAQI	$PM_{2.5}$	PM_{10}	SO_2	NO_2	CO	O_3
0	0	0	0	0	0	0
50	35	50	50	40	2	100
100	75	150	150	80	4	160
150	115	250	470	180	14	215
200	150	350	800	280	24	265
300	250	420	1600	565	36	800
400	350	500	2100	750	48	1000
500	500	600	2620	940	60	1200

six pollutants (O_3, $PM_{2.5}$, PM_{10}, NO_2, SO_2, and CO). When I_i represents the maximum value among all six pollutants, i is termed the primary pollutant.

DATA AND ECONOMIC MODEL

Data description

All controlled variables were sourced from the National Bureau of Statistics of China (http: //www.stats.gov.cn/tjsj/ndsj/), while the dependent variables AQI and AQCI were calculated using the national air quality calculation platform. Figure 1 illustrates the AQI and the AQCI values for the 31 capital cities in China in 2019. The dataset spans from 2013 to 2019 and includes information on six air pollutants and the economic state. According to technical regulations, AQI values are categorized into six classes: 0–50 (Excellent), 51–100 (Good), 101–150 (Lightly polluted), 151–200 (Moderately Polluted), 201– 300 (Heavily Polluted), and more than 300 (Severely Polluted). AQCI values are classified into six classes as well: 0–2 (Excellent), 2–4 (Good), 4–6 (Light Polluted), 6–10 (Moderately Polluted), 10–12 (Heavily Polluted), and more than 12 (Severely Polluted). Capital cities have their own AQI or AQCI standards aligned with national air quality standards. In Fig.1, the air quality is depicted with a colour scale, where darker shades indicate higher pollution rates. As shown in Fig. 1, the majority of cities exhibit low AQI, i.e. having good or lightly polluted air. As for the AQCI, most cities fall in "Good" and "lightly Polluted" categories, except for those in Hebei, Shanxi and Shandong provinces, which show a significant variation between light and heavy polluted under both standards.

Economic model

Multinomial logistic regression is employed for predicting categorical placement or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be dichotomous (binary) or continuous (interval or ratio in scale). This method is an extension of binary logistic regression, accommodating more than two categories of the dependent or outcome variable. Similar to binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to access the probability of specific membership. Variable selection or model specification methods are similar to those used in standard multiple regression, including sequential or nested logistic regression analysis. These methods are applied when one dependent variable serves as criteria for placement or choice on subsequent dependent variables. The multinomial logistic regression method is employed for predicting China's air quality based on the AQI and AQCI criteria. The economic model is shown below:

$$Pr(Y_k) = \frac{e^{\beta_k \cdot x_i}}{1 + \sum_{j=1}^{k-1} e^{\beta_j \cdot x_i}}$$

Here, $Pr(Y_k)$ is the probability of category k occurring, Y_k is the dependent variable (AQI or AQCI) representing one of the pollution categories observed in actual cases. k signifies the air quality level, including "Excellent", "Good", "Light polluted", "Median polluted", "Heavy" and "Serious polluted". x_i denotes controlled variables, with, x_1

Table 2. Limitation of the secondary standards for each pollution concentration

Pollution items	Time	Concentration limit	
		Level one	Level two
SO_2	Annual average	20	60
NO_2	Annual average	40	40
CO	24 hour average	4	4
O_3	8 hour average	100	160
PM_{10}	Annual average	40	70
$PM_{2.5}$	Annual average	35	75

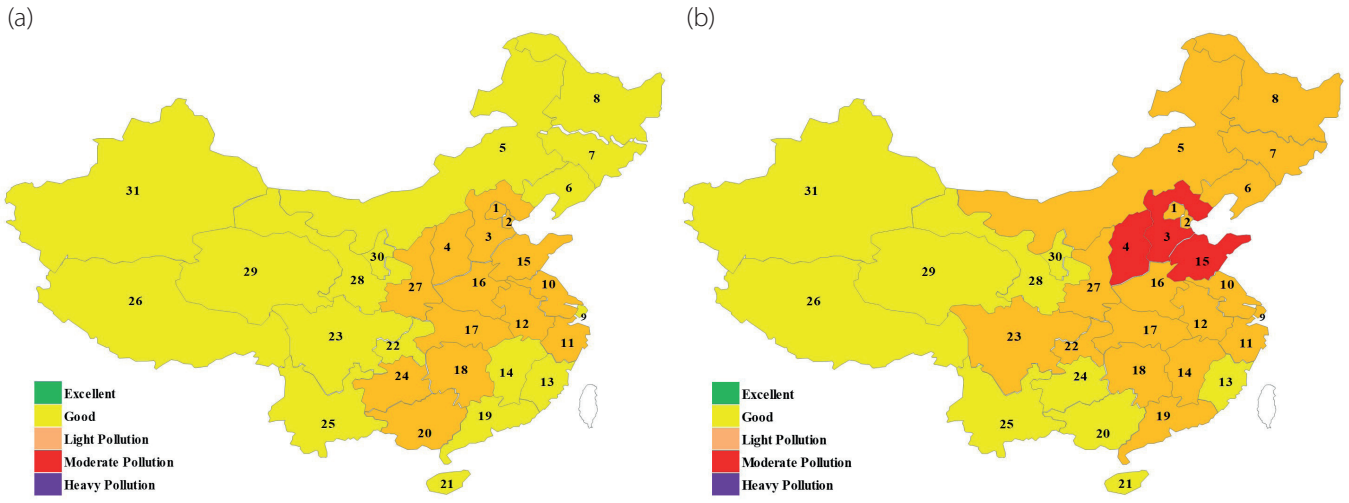

Fig. 1. AQI (a) and AQCI (b) values in Chinese capital cities. 1--Beijing, 2--Tianjin, 3--Shijiazhuang, 4--Taiyuan, 5--Hohhot, 6--Shenyang, 7--Changchun, 8--Harbin, 9--Shanghai, 10--Nanjing, 11--Hangzhou, 12--Hefei, 13--Fuzhou, 14--Nanchang, 15--Jinan, 16--Zhengzhou, 17--Wuhan, 18--Changsha, 19--Guangzhou, 20--Nanning, 21--Haikou, 22-- Chongqing, 23--Chengdu, 24--Guiyang, 25--Kunming, 26--Lhasa, 27--Xi'an, 28--Lanzhou, 29--Xining, 30--Yinchuan, 31--Urumqi

representing the primary pollutant, x_2 indicating low-level GDP status. x_3 representing middle level GDP status, and x_4 denoting high level GDP status. The β_k and β_j are the parameters in the model.

Figures 2 and 3 illustrate the primary pollutants influencing AQI and AQCI during the period 2013-2019, showcasing the proportion of the primary contaminant in the total annual pollution. $PM_{2.5}$ and O_3 were consistently identified as the primary pollutants for AQI and AQCI, respectively, comprising 71.4% and 73.3%. The second primary pollutants for AQI was O_3, one-quarter. Conversely, PM_{10} and CO contributed insignificantly, representing only 6.5% and 0.9%, over the years. As for AQCI, PM_{10} was the second primary pollutant, at 15.2%, followed by O_3 and NO_2 at 8.3% and 3.2%, respectively. The proportion of $PM_{2.5}$ as the primary pollutant in AQCI exhibited fluctuations and an overall decreasing trend. In contrast, the concentration of O_3 as primary pollutant in AQI increased annually, reaching 100% in 2018-2019. The challenging nature of controlling ozone, among six air pollutants, warrants careful consideration.

MODEL RESULTS

The multinomial logistic model, involves choosing one outcome as a "pivot" and deflecting other outcomes relative to the pivot outcome. Similarly, in the AQI model, we use "Heavily polluted" as the pivot. The process is as follows:

$$In\frac{Pr(Y_1)}{Pr(Y_4)} = 10.545 = 0.0525x_a + 10.298x_b + 10.226x_c$$

$$In\frac{Pr(Y_2)}{Pr(Y_4)} = 3.531 + 9.09\cdot 10^{-4}x_a + 8.219x_b + 9.098x_c$$

$$In\frac{Pr(Y_3)}{Pr(Y_4)} = 5.767 - 3.444\cdot 10^{-2}x_a + 4.848\cdot 10^{-1}x_b + 4.742\cdot 10$$

Where $x_a = \log\frac{x_1}{x_4}$, $x_b = \log\frac{x_2}{x_4}$, $x_c = \frac{x_3}{x_4}$

$$RHS_1 = In\frac{Pr(y_1)}{Pr(y_4)},\ RHS_2 = In\frac{Pr(y_2)}{Pr(y_4)},\ RHS_3 = In\frac{Pr(y_3)}{Pr(y_4)}$$

if we exponentiate both sides, and solve for the probabilities, we get

$$Pr(Y_1) = Pr(Y_i = Y_4) * RHS_1$$

$$Pr(Y_2) = Pr(Y_i = Y_2) * RHS_2$$

$$Pr(Y_3) = Pr(Y_i = Y_4) * RHS_3$$

$$Pr(Y_4) = 1 - \sum_{k=1}^{3} Pr(Y_i = Y_k) = \frac{1}{1 + \sum_1^{k-1} e^{RHS_k}}$$

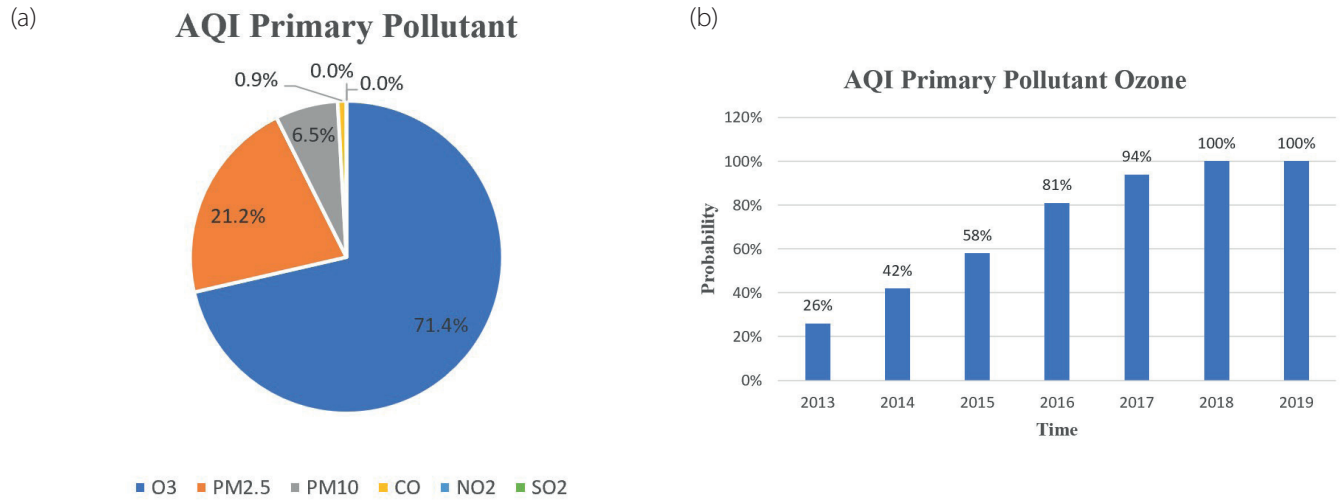

Fig. 2. AQI Pollutant Composition (2013-2019).

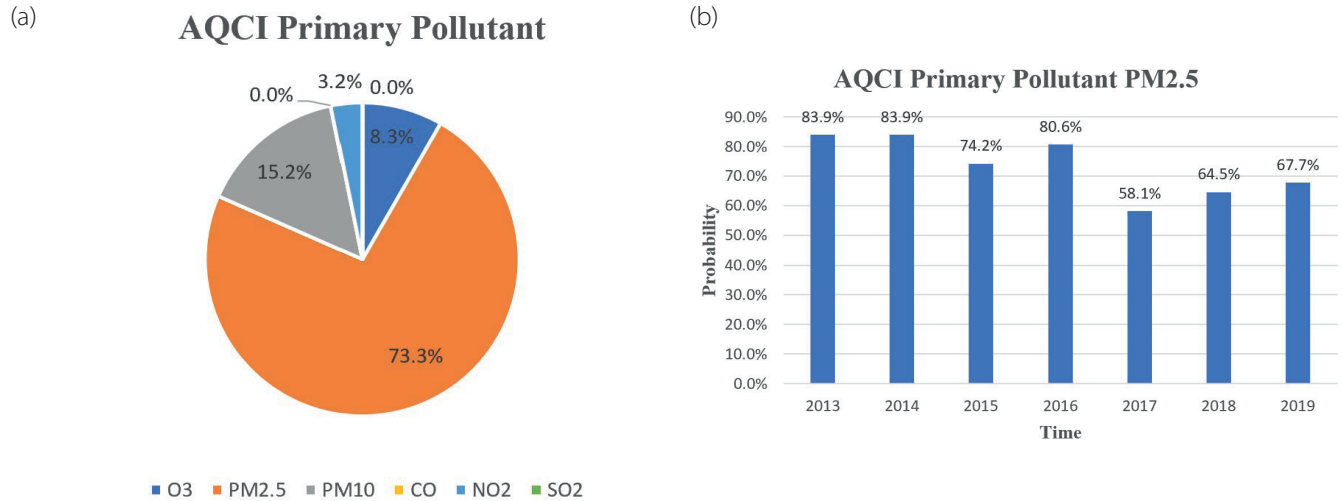

Fig. 3. AQCI pollutant Composition (2013-2019).

In AQI model, we choose the "Serious pollution" as the pivot. The process is as follows:

$$In\frac{Pr(Y_1)}{Pr(Y_4)} = 111.824 - 1.585x_a + 20.369x_b + 69.724x_c$$

$$In\frac{Pr(Y_2)}{Pr(Y_4)} = 72.642 - 0.524x_a + 4.86x_b + 49.127x_c$$

$$In\frac{Pr(Y_3)}{Pr(Y_4)} = 111.897 - 1.085x_a + 4.064x_b + 52792x_c$$

Where $x_a = \log\frac{x_1}{x_4}$, $x_b = \log\frac{x_2}{x_4}$, $x_c = \log\frac{x_3}{x_4}$,

Two tailed tests.

H0: the parameter of coefficient $x_{l,l=a,b,c}$ is significant.

H1: the parameter of coefficient $x_{l,l=a,b,c}$ is not significant.

Two-tailed tests were employed to assess the significance coefficient parameters. The null hypothesis stated that the coefficient was significant, while the alternative hypothesis suggested otherwise. The results, detailed in Tables 3 and 4, indicate that the most absolute z-value exceeded 1.96, corresponding to p-value less than 0.05. Therefore, we could reject the alternative hypothesis, confirming the significance of the coefficients.

The multiple logistic regression models were designed to predict the likelihood of China's air quality categories, including predictions for good air or light, moderate, and heavy polluted. This confusion matrix results, shown cased in Tables 5 and 6 in Appendices, offer a detailed breakdown based on the classification of the four categorical variables.

To assess the model's performance, Receiver Operating Characteristic curve (ROC) combined with Area Under the Curve (AUC) and F-1 score are employed. An AUC value above 0.8 and an F1-score exceeding 0.8 generally indicate high model quality. The AQI model demonstrates an F1-score greater than 0.8, while the AQCI model's F1-score approaches 0.8, affirming their high-quality predictions. Additionally, AUC values exceeding 0.8 in Table 5 and Table 6 further support the models' efficacy. Figure 4 visually presents the ROC curve results for AQI and AQCI models.

In Figure 5(a), the horizontal axis represents ozone concentration, with the maximum concentration value of 211 μg/m³, and the minimum concentration value of 69 μg/m³ adjusted to 65 μg/m³ and 215 μg/m³, respectively. The vertical axis represents the probability of each category, with the high GDP status shown in red. The low GDP status in green, the middle GDP status in blue. Notably, as ozone levels increase, the probabilities of various air quality levels fluctuate, revealing distinct opportunities for additional GDP and air quality levels in different regions. The application of multinomial logistic regression to predict the probability of each pollution level is considered, incorporating pollutant concentration (μg/m³) and GDP state (Dummy) as control variables. The results showed changes in air quality with increase of ozone concentration.

The AQCI model, focusing on $PM_{2.5}$ demonstrated the results similar to the AQI model. The $PM_{2.5}$ concentration

Table 3. Two-tailed test for the AQI model

	Intercept		Ozone/High		Low/High		Middle/High	
	Z-value	P-value	Z-value	P-value	Z-value	P-value	Z-value	P-value
Good/Heavy	3.57	3.5E-4	-2.89	4.0E-3	31.59	0	39.82	0
Light/Heavy	1.19	0.231	0.05	0.959	25.27	0	35.53	0
Moderate/Heavy	1.20	0.228	-1.10	0.271	34.94	0	36.49	0

Table 4. Two-tailed test for the AQCI model

	Intercept		Ozone/High		Low/High		Middle/High	
	Z-value	P-value	Z-value	P-value	Z-value	P-value	Z-value	P-value
Good/Heavy	559.22	0	-4.96	7.2E-7	-3.08	2.1E-3	5.96	0
Light/Heavy	554.99	0	-4.77	1.8E-6	-14.33	0	51.41	0
Moderate/Heavy	57305	0	-2.39	1.7E-2	9.6E+6	0	7.1E+9	0

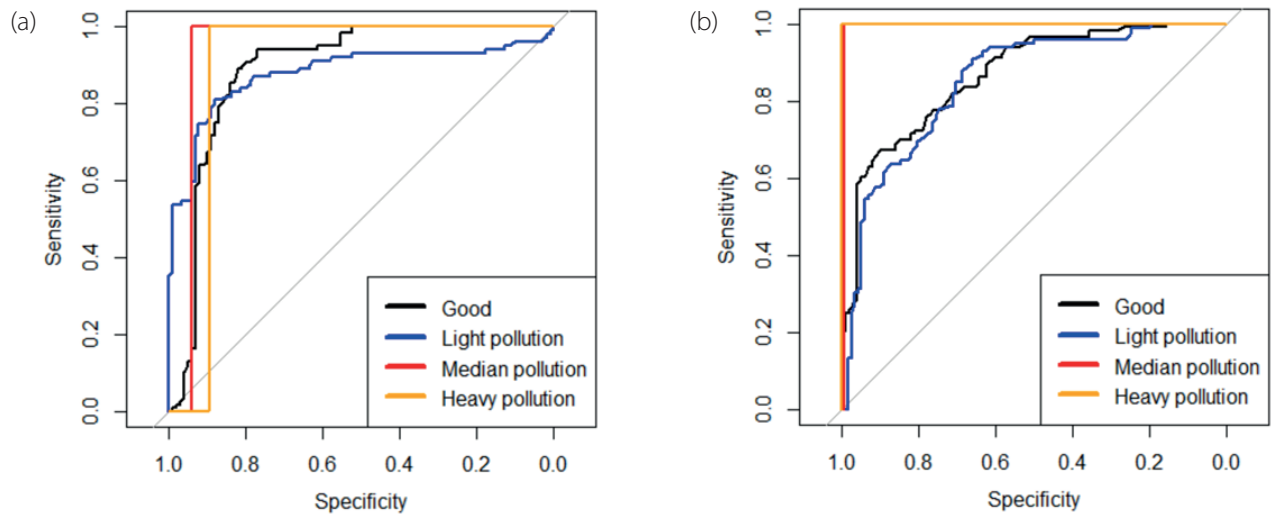

Fig. 4. The AQI model (a) and AQCI (b) model ROC curve results

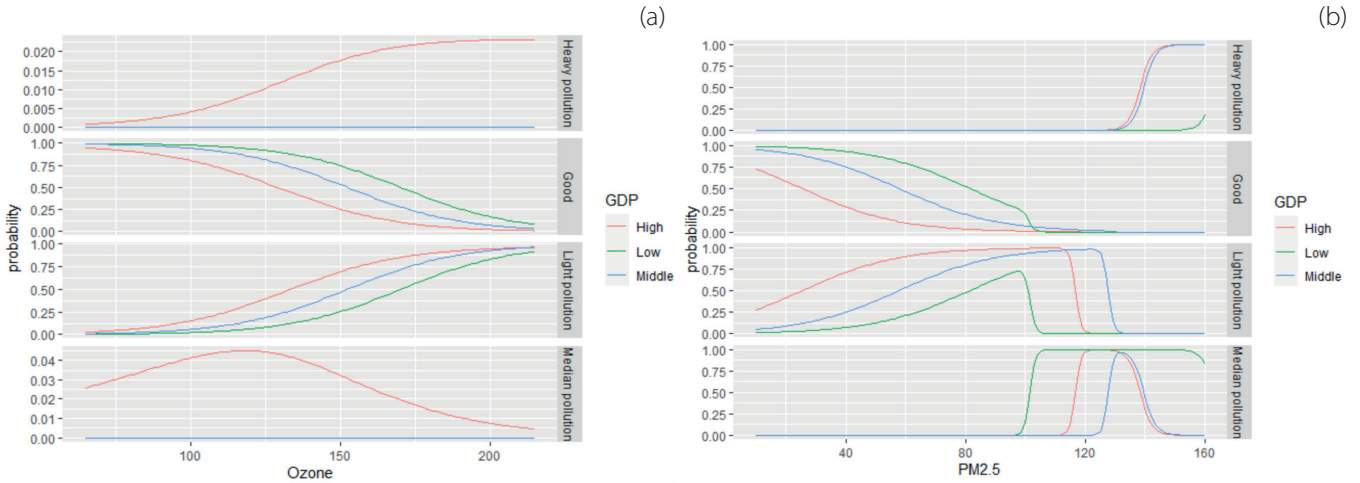

Fig. 5. The AQI model (a) and AQCI model (b) results

range limits of 12 μg/m³ and 154 μg/m³ were adjusted to 5 μg/m³ and 165 μg/m³, respectively. This analysis suggests that, under similar pollution sources, higher GDP status correlates with increased probabilities of air quality pollution across different economic statuses.

CONCLUSIONS AND DISCUSSION

The study's conclusions emphasize the suitability of multinomial logistic regression for predicting air quality in China, especially when considering different evaluation systems. The economic status of a region emerges as a significant determinant of air quality, with regions exhibiting high GDP associated with a higher probability of experiencing light or heavy air pollution rather than excellent and good air quality. This finding highlights the importance of integrating economic factors into air quality assessments.

Furthermore, the result reveals the variability in primary pollutants or influential factors across different air quality evaluation systems, warranting further exploration.

Consequently, this study advocates for the development of diverse criteria for air quality assessment, emphasizing the need for precise performance standards and the application of uncertainty models to evaluate long-term trends comprehensively. By addressing these aspects, more adequate measures are expected to be taken to tackle air pollution issues.

In conclusion, these findings provide valuable insights to future research in air quality assessment. Further investigations should focus on the development of improved performance standards and the exploration of the complex interplay between economic factors, primary pollutants, and air quality evaluation systems.

REFERENCES

Bure V. M, Parilina E. M., (2013). Probability theory and mathematical statistics, 1st ed. St Petersburg, Lan Publ., 416 p. (in Russian).

Bure V. M., Parilina E. M., Sedakov A.A., (2019) Applied statistic methods in R and Excel. 3rd ed. St Petersburg, Lan Publ., 196 p. (in Russian).

Bure V.M., (2007). Methodology for statistical analysis of empirical data. St Petersburg, Lan Publ. (in Russian).

Di Q, Amini H, Shi L, et al. (2019). An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environment international, 2019, 130: 104909, DOI: 10.1016/j.envint.2019.104909.

Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., and Lin, S. (2017). A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-4/W2, 15–22, DOI:10.5194/isprs-annals-IV-4-W2-15-2017.

Fann N, Risley D. (2013). The public health context for PM2. 5 and ozone air quality trends. Air Quality, Atmosphere & Health, 6(1): 1-11, DOI:10.1007/s11869-010-0125-0.

Guo Q, He Z, Li S, et al. (2020). Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. Aerosol and Air Quality Research, 20(6): 1429-1439, DOI: 10.4209/aaqr.2020.03.0097.

He Y., Qi D., Bure V. M. (2023). New application of multiple linear regression method-A case in China air quality. Vestnik of Saint Petersburg University. Applied Mathematics, Computer Sciences, Control Processes, 18(4), 516-526. DOI:10.21638/11701/spbu10.2022.406

Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. K voprosu avtomatizatsii postroeniia variogramm v zadachakh tochnogo zemledeliia [On the issue of semivarigrams constructing automation for precision agriculture problems]. Vestnik of Saint Petersburg University. Applied Mathematics, Computer Sciences, Control Processes, 2020, 16(2), 177-185, (in Russian). DOI:10.21638/11701/spbu10.2020.209.

Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. Theoretical foundations of prababilistic and statistical forecasting of agrometerological risks. Vestnik of Saint Petersburg University. Applied Mathematics, Computer Sciences, Control Processes, 2021, 17(2), 174-182, (in Russian). DOI:10.21638/11701/spbu10.2021.207.

Karimian H, Li Q, Wu C, et al. (2019). Evaluation of different machine learning approaches to forecasting PM2. 5 mass concentrations. Aerosol and Air Quality Research, 19(6): 1400-1410, DOI: 10.4209/aaqr.2018.12.0450

Le V D, Bui T C, Cha S K. (2020). Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. 2020 IEEE international conference on big data and smart computing (BigComp). IEEE, 55-62, DOI: 10.1109/BigComp48618.2020.00-99.

Li X, Peng L, Hu Y, et al. (2016). Deep learning architecture for air quality predictions. Environmental Science and Pollution Research, 23(22): 22408-22417, DOI:10.1007/s11356-016-7812-9.

Nadeem I., Ilyas A.M., Uduman P.S. Analyzing and forecasting ambient air quality of Chennai city in India. (2020) GEOGRAPHY, ENVIRMENT, SUSTAINABILITY.13(3), 13-21, https:// doi.org/10.24057/2071-9388-2019-97.

Pan B. (2018). Application of XGBoost algorithm in hourly PM2. 5 concentration prediction, IOP conference series: earth and environmental science. IOP publishing, 2018, 113(1): 012127, DOI: 10.1088/1755-1315/113/1/012127.

R. Stern, P. Builtjes, M. Schaap, R. Timmermans, R. Vautard, A. Hodzic, M. Memmesheimer, H. Feldmann, E. Renner, R. Wolke, et al., (2008). A model inter-comparison study focussing on episodes with elevated pm10 concentrations, Atmospheric Environment, 42(19), 4567–4588, DOI: 10.1016/j.atmosenv.2008.01.068.

Schachter E N, Moshier E, Habre R, et al. (2016). Outdoor air pollution and health effects in urban children with moderate to severe asthma. Air Quality, Atmosphere & Health, 9(3): 251-263. DOI:10.1007/s11869-015-0335-6.

Senarathna M., Priyankara S., Jayaratne R., Weerasooriya R., Morawska L., Bowatte G. (2022). Measuring Traffic Related Air Pollution Using Smart Sensors In Sri Lanka: Before And During A New Traffic Plan. GEOGRAPHY, ENVIRONMENT, SUSTAINABILITY. 15(3):27-36. https://doi.org/10.24057/2071-9388-2022-011

Stojov V, Koteli N, Lameski P, et al. (2018). Application of machine learning and time-series analysis for air pollution prediction. Proceedings of the CIIT.

Tao Q, Liu F, Li Y, et al. (2019). Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU. IEEE access, 7: 76690-76698, DOI: 10.1109/ACCESS.2019.2921578

Tong W, Li L, Zhou X, et al. (2019). Deep learning PM2.5 concentrations with bidirectional LSTM RNN. Air Quality, Atmosphere & Health, 12(4): 411-423, DOI: 10.1007/s11869-018-0647-4.

Wang K, Yin H, Chen Y. (2019). The effect of environmental regulation on air quality: A study of new ambient air quality standards in China. Journal of Cleaner Production, 215: 268-279, DOI:10.1016/j.jclepro.2019.01.061

Wang, S., & Hao, J. (2012). Air quality management in China: Issues, challenges, and options. Journal of Environmental Sciences, 24(1): 2-13, DOI:10.1016/S1001-0742(11)60724-9.

Zaib S, Lu J, Bilal M. (2022). Spatio-Temporal Characteristics of Air Quality Index (AQI) over Northwest China. Atmosphere, 13(3): 375, DOI:10.3390/atmos13030375.

APPENDICES

Table 5. AQI model confusion matrix results

	Good	Light Polluted	Moderate Polluted	Heavy polluted
F1 score	0.8858	0.8247	0	0
AUC	0.8894	0.8783	0.9560	0.9398

Table 6. AQCI model confusion matrix results

	Good	Light Polluted	Moderate Polluted	Heavy polluted
F1 score	0.7644	0.7688	1	1
AUC	0.8665	0.8508	1	1