Application of Statistical Models in Air Quality Monitoring

Ziyi Han[1,a,*]

*[1]Mathematics and Statistics International Foundation Program, The University of Warwick,
Coventry, CV4 7AL, The United Kingdom
a. u5673263@live.warwick.ac.uk
corresponding author

Abstract: Currently, statistical models are vital instruments for examining and forecasting diverse environmental variables, including air quality analysis. With the development of big data, foundational statistical techniques—including regression and time series analysis—are adapting to manage larger datasets and complex environmental dynamics. This paper examines the recent applications of statistical models used to assess and forecast air quality in Chinese cities based on existing literature and data research. By focusing on various modeling approaches, such as Poisson regression, grey correlation, and neural networks, it discusses the advantages and drawbacks of each model. Furthermore, it examines the enhancement of classical analytical methodologies inside a big data framework to augment the accuracy of air quality analysis and forecasting. The result shows that although datasets have been more extensive, to increase the accuracy of pollution forecasting, diverse data and more excellent computational capabilities are still needed. Advances in machine learning and optimisation algorithms show promise for overcoming these challenges in the future.

Keywords: Big data, Statistical models, Air quality, Regression analysis

1. Introduction

The effective use of statistical models is critical for analysing atmospheric pollution and providing solutions in environmental science. With the growth of urbanisation and industrialisation, ecological complexity and the severity of environmental pollution have increased sharply, raising the demand for precision in statistical analysis models. Traditional models, such as multiple linear regression and time series models, have provided a foundation for analysing air pollution data and have been effective to some extent in examining pollutant trends and their impacts on human health[1-2]. The advent of big data has introduced both opportunities and difficulties to environmental science. Sources such as satellite imagery, meteorological stations, and IoT sensors now generate vast amounts of data, continually increasing data diversity. To achieve a more comprehensive and accurate analysis, it is clear that improvements in computational capabilities and structural frameworks are required[3]. Improved computational capacity and innovative algorithm optimisation enable researchers to integrate diverse datasets, thus liberating models, and allowing them to play a more significant role[4]. This paper synthesises recent research on statistical models in environmental science, examining the integration of big data and its impact on traditional methods, such as regression and time series analysis[5-6].

2. Air quality monitoring

With air quality issues getting more and more serious, especially in urban areas, it is vital to explore air quality monitoring methods and apply them into practice to better monitor air quality, predict possible air pollution and prepare solutions to improve air quality. In recent years, air pollution, especially haze, has posed a threat to residents' physical health and mental wellness—some severe pollution can even cause ischemic heart disease, stroke and other serious diseases[1-2]. Moreover, poor air quality could adversely impact sustained economic growth. Thus, it is essential to accurately excavate and detect the information on air quality and monitor the conditions of air pollution to facilitate the provision of pollution prevention and find measures. Not only will this enable environmental agencies to take action in time, but it will also mitigate the risks of health diseases[4].

3. Statistical models

3.1. Time series forecasting models

1) ARIMA (Auto-Regressive Integrated Moving Average)

The ARIMA model, which was introduced by Box and Jenkins, is a combined model of three statistical models—Autoregressive (AR), Integrated (I) and Moving Average (MA), and is used to predict a value in a response Time Series. This model is presented as ARIMA(p, d, q), where p, d, q represent the autoregressive order respectively[7].

When using ARIMA to make predictions, after performing the stabilization process on the sequence of unstable time, a regression model will be established to determine the lag value based on the current value and lag value of the random error difference. Suppose that γ represents the original sequence while Y illustrates the sequence of difference, the prediction for Y can be presented in:

$$\widehat{Y} = c + \emptyset_1 \gamma_{t-1} + \cdots + \emptyset_p \gamma_{t-p} + \cdots - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \quad (1)$$

where c represents constant, $\emptyset_1 \gamma_{t-1}, \ldots, \emptyset_p \gamma_{t-p}$ represent AR, $\theta_1 e_{t-1}, \ldots, \theta_q e_{t-q}$ represent MA[4].

2) Prophet model

The Prophet Model is a forecasting model based on Bayesian time series that handles missing data, outliers, and irregular sampling, manages abrupt shifts and can model multiple seasonal patterns simultaneously.

The core formula is: $y_t = g_t + s_t + h_t + \varepsilon_t$ (Additive Form) or $y_t = g_t \cdot s_t \cdot h_t + \varepsilon_t$ (Multiplicative Form).

where: g_t is the trend component, s_t is the seasonality component, h_t is the external variables component and ε_t is the error term[2].

The additive form and multiplicative form are equivalent, due to the ability to transform between addition and multiplication through simple mathematical operations, providing flexibility to match the data's behaviour.

3) LSTM (Long Short-Term Memory)

LSTM is a recurrent neural network (RNN) architecture that is widely used in predicting trends in pollution levels, weather and so on. LSTM can be employed in air quality forecasting to capture temporal dependencies in pollutants such as PM2.5. Compared with RNN, this model applies three "gates" to transform information. The core four formulas are as follows:

Forget Gate:

$$f_t = \sigma * (W_f[h_{t-1}, x_t] + b_f) \quad (2)$$

where f_t represents the forget gate activation vector, W_f represents the forget gate weigh matrix, $[h_{t-1}, x_t]$ represents the concatenation of previous hidden state and current input vector, b_f represents the forget gate bias vector, and σ means the sigmoid activation function $(\sigma(z) = \frac{1}{1+e^{-z}})$

Input Gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$\widetilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (3)$$

where i_t represents the input gate activation vector, $\widetilde{C}_t$ represents the candidate cell state values; W_i and W_C represent the weight matrices for input gate and cell state update, respectively; b_i and b_C are bias vectors.

Cell State Update:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \quad (4)$$

where $\odot$ represents element-wise multiplication.

Output Gate:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \odot \tanh(C_t)$$
$$\sigma = \frac{1}{1+e^{-x}}$$
$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (5)$$

where o_t represents the output gate activation vector and h_t represents the current hidden state (also the output of the LSTM at time t)[4,8].

4) Same-period prediction models

The model can be used in prediction that is in a short period of time. The core formula is as follows:

$$Y_t = a_0 + a_1 \times Y_{t-1} + a_2 \times Y_{t-2} + a_3 \times Y_{t-3} \quad (6)$$

where $a_0, a_1, a_2 \ldots a_n$ represent regression coefficients, Y_{t-n} represents the pollutant concentration of n days before the day[9].

3.2. Linear regression models

1) Linear regression

Linear Regression is used to find the relationship between two or more than two variables. For instance, one can utilize it to investigate and measure the correlation between air quality indicators (a dependent variable) and influencing factors such as temperature, wind speed, humidity, or emission levels (independent variables). Some core formulas are as follows:

$$E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K] = \phi(x_1, x_2, \ldots, x_K) \quad (7)$$

where Y represents the response variable, $X_1, X_2, \ldots, X_K$ represents the explanatory variables.

$$\phi\left(x_1, x_2, \ldots, x_K\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K \quad (8)$$

which is linear in parameters β_j[10].

2) Multiple linear regression (multivariate)

Multiple Linear Regression extends linear regression to multiple predictors. For n mutually independent values(x, y):

$$Y = X\beta + \varepsilon$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \tag{9}$$

This can also be represented in a matrix form)in which, the vector Y is the dependent variable, represented as $Y = (y_1, \ldots, y_n)^T$; the matrix $X \in R^P$ is the independent variable, represented as $X = (1, x_1, \ldots, x_{p-1}), \beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$, β_0 represents the intercept, $\beta_i =$ the coefficient of the i-th Predictor Variable, error terms $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ with each error term $\varepsilon_i \sim N(0, \sigma^2)$[3].

3) Generalized Linear Models (GLM)

The Generalized Linear Models can be used when the dependent variable Y follows a discrete distribution. It extends linear regression to non-normal distributions (e.g., Poisson, Binomial). Core formulas are as follows:

Supposed that $E(Y) = u = (E(Y_1), \ldots, E(Y_n))^T = (u_1, \ldots, u_n)^T$,

$$g(u) = X\beta \tag{10}$$

in which, $X = (1, x_1, \ldots, x_{p-1}), x_i$ represents the i -th factor that influence Y , $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})', \beta_0$ represents the intercept, $\beta_i =$ the coefficient of the i-th Predictor Variable[3].

4) Poisson regression

$$Y_i = E(Y_i) + \varepsilon_i, i = 1, \ldots, n \tag{11}$$

Supposed that $E(Y) = u = (E(Y_1), \ldots, E(Y_n))^T = (u_1, \ldots, u_n)^T$, and there exists a linear relation between Y and X, denoted as $\mu = (x_i, \beta)$.

$$(i)\mu_i = (x_i, \beta) = x_i^T \beta$$

$$(ii)\mu_i = (x_i, \beta) = \exp(x_i^T, \beta)$$

$$(iii)\mu_i = (x_i, \beta) = \log_e^{(x_i^T, \beta)}$$

$$Y_i = \exp(x_i^T, \beta),\ i = 1, \ldots, n$$

$$\mu_i = \exp(x_i^T, \beta)$$

$$E(Y_i) = \mu_i \tag{12}$$

in which, $\mu_i \geq 0$[3].

5) Robust regression with mean shift penalization

$$g(\mu) = X\beta + \gamma$$

$$\mu = E(Y) = (\mu, \ldots, \mu_n)^T$$

$$\gamma = (\gamma, \ldots, \gamma_n)^T \tag{13}$$

in which, γ_i represents the i-th mean shift parameter of the observed value[3].

3.3. Machine learning models and Gray Relational Analysis (GRA)

There are many effective machine learning models that can be used in air quality investigation, such as SVR (Support Vector Regression), XGBoost (Extreme Gradient Boosting), ELM (Extreme Learning Machine), Cluster Regression Models and BP Neural Networks[2,4,6,9].

GRA emphasizes measuring the relationships among sequences and can be employed for feature selection or assessing variable significance[1,6].

4. Applications and advances in statistical models

4.1. Regression analysis in air quality research

Linear regression models are frequently applied to explore connections between environmental conditions and indicators of air quality, including PM2.5 levels and AQI. For example, studies have applied regression methods to investigate how socio-economic and weather-related variables impact air quality across Chinese cities, illustrating the utility of classical regression analysis in detecting pollution patterns[2,11]. The study above found that including covariates (e.g., temperature or pollution shock variables) improved the accuracy of predictive in forecasting AQI using models like ARIMA[2]. Recently, Poisson regression models have been formulated to assess the impacts of pollutant concentrations and climatic variables on urban air quality, demonstrating the versatility of regression techniques in managing both skewed and discrete data[3].

4.2. Time series models for air quality forecasting

Time series analysis is crucial in forecasting air pollution patterns, particularly for short-term predictions. In recent years, China’s rapid development has led to significant environmental challenges, especially in air pollution, which has profoundly impacted residents' health conditions and living standards. In light of this context, Zhang et al. initiated a research program employing mathematical models and conventional time series methodologies, including ARIMA and seasonal decomposition, to address air pollution. Their findings indicated that the selected time series models were effective in predicting daily variations in the AQI, as demonstrated in studies concerning Beijing's air quality[2]. Motivated by the rising severity of air pollution caused by the rapidly growing industrialization in China, another study conducted by Li et al. aimed to address the challenge by using some new mathematical models due to the unsatisfied results of existing predictive systems like the WRF-FMAQ. With the growth of available data, however, advanced techniques like Long Short-Term Memory (LSTM) neural networks have emerged, enhancing accuracy by identifying nonlinear patterns within time series data. Comparisons between ARIMA and LSTM models reveal that, although neural networks are more computationally intensive, they offer superior precision for handling complex datasets[12].

4.3. Grey models and hybrid approaches

With the rapid economic growth in China, air pollution became a serious environmental issue, which also affected public health. However, the traditional models failed to effectively address noise and complex patterns in AQI data, prompting the great need for hybrid models. That is when Grey models have been applied to study pollution in cities where data availability is limited, utilising interval-based relationships to estimate pollutant levels[1]. Hybrid models that integrate grey correlation with neural networks or other machine learning techniques are becoming increasingly popular for their flexibility in managing both small and extensive datasets[4]. These hybrid methods combine the simplicity of traditional models with the adaptability of machine learning, resulting in more dependable predictive outcomes[5].

5. Challenges and improvements related to big data

Traditional models, such as regression analysis, were not originally designed for the extensive datasets produced by modern environmental sensors. As datasets grow, computational constraints become a barrier, especially for methods like Poisson regression, which may become computationally intensive with high-dimensional data. Machine learning methods optimised for big data, such as neural networks, present potential solutions but require considerable computational resources[12].

Extensive datasets often comprise heterogeneous information, integrating both structured and unstructured data from sources like satellites and sensors. Integrating these diverse data types into statistical models presents challenges, as each type demands specific preprocessing methods[9,13]. To address this, researchers are increasingly employing data fusion and transformation techniques to harmonise the data, a step essential for improving model accuracy.

A further challenge when utilising big data in statistical modeling is ensuring that models remain generalisable across different regions and time periods. Models specifically designed for certain locations, such as time series models for cities like Beijing or Chongqing, may be less effective elsewhere due to unique environmental factors and sources of pollution[5-6]. Therefore, big data analytics should integrate adaptive methods that allow these models to generalise effectively across diverse datasets.

6. Conclusion

The utilization of big data has significantly enhanced the efficacy of statistical models in environmental science, particularly in forecasting air quality trends. Conventional methods, such as regression and time series analysis, have evolved through the integration of hybrid methodologies and machine learning, leading to enhanced accuracy and predictive efficacy. Nevertheless, the adoption of big data introduces numerous challenges, such as the complexity of diverse data types, substantial computational requirements, and difficulties in achieving model generalisation across various regions and environmental contexts. To propel statistical modeling forward in ecological science, it is essential to develop models that are not only computationally optimised but also adept at integrating multiple data types. By addressing these issues with innovative data processing and adaptable modeling techniques, researchers and policymakers can harness big data more effectively, facilitating the creation of targeted, data-driven strategies for environmental management across different regions. The essay briefly demonstrates the potential of statistical models in air quality monitoring but might fall short in addressing critical challenges. Thus, future research will focus on computational solutions for big data, explore robust data integration techniques, improve model generalization strategies, and address data quality and validation comprehensively.

References

[1] Zhang, W., Yang, W.S., Bai, Q., et al. (2022). Study the Correlation Between Motor Vehicle Ownership and Urban Air Quality in Xi'an Based on Descriptive Statistics and the Grey Correlation Model. Transport Research, 8(3), 111-119.

[2] Zhang, N. (2023). Analysis and Prediction of China's Air Quality Index Based on Statistical Models. Shanghai University of Finance and Economics.

[3] Lu, G. (2021). Study on Factors Affecting Air Quality in Chinese Cities Using Robust Poisson Regression. Environmental Studies Journal, 3, 56-67.

[4] Wang, F. (2022). Air Quality Prediction Using ELM and Multi-objective Grey Wolf Optimization Algorithm. Environmental Forecast Journal, 12, 88-96.

[5] *Zou, J. (2022). Prediction of Chongqing Air Quality Index Using Combined Weight Models. Southwest University, Master's Thesis.*

[6] *Zhao, C. (2023). Application of Mathematical Modeling in Roadside Air Quality Analysis. Chinese Science and Technology Information, 9, 60-63.*

[7] *Araghinejad, S. and Araghinejad, S., 2014. Time Series Modeling (pp. 85-137). Springer Netherlands.*

[8] *Siami-Namini, S., Tavakoli, N. and Namin, A.S., 2018, December. A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE international conference on machine learning and applications (ICMLA) (pp. 1394-1401). Ieee.*

[9] *Liu, M. (2014). Establishment and Application of Winter Environmental Air Quality Prediction Model in Shenyang. Environmental Monitoring in China, 30(4), 10-16.*

[10] *Seber, G.A. and Lee, A.J., 2012. Linear regression analysis. John Wiley & Sons.*

[11] *Lin, B. (2010). Multiple Linear Regression Analysis and Its Applications. China Science and Technology Information, 9, 60-63.*

[12] *Li, G., Qiu, Z., Miao, J., et al. (2023). LSTM-based Air Quality Prediction Model. Journal of Southwest Minzu University, 49(1), 67-75.*

[13] *Ye, S.Q., Huang, S.Y., Chen, D.H., et al. (2017). Application of Statistical Models in Urban Air Quality Forecasting. Guangdong Environmental Monitoring Center, 510308.*