Air Quality Prediction based on Neural Network

Tianyu Wang *

School of Business Administration, Chengdu Jincheng college, Chengdu, 610000, China

* Corresponding Author Email: wangtianyu1@cdjcc.edu.cn

Abstract. Air pollution, driven by various human activities, has become a critical global issue with significant impacts on human health and the environment. Accurate forecasting and monitoring of air quality are essential to mitigate these effects. This paper introduces a novel deep learning model, AirPhyNet, which integrates atmospheric physics principles into its design, aiming to improve the precision of air quality predictions. The model is evaluated against several baseline models, demonstrating its robustness and effectiveness in forecasting air quality dynamics. The study emphasizes the importance of incorporating physical insights into deep learning models to enhance prediction accuracy and interpretability. The model's performance is compared systematically with existing baseline models, showcasing its superior accuracy and generalisation capabilities. The paper also discusses the use of machine learning algorithms for air quality prediction, highlighting the need for accurate results to inform public health measures and environmental strategies. The study utilizes real-world air quality datasets from Beijing and Shenzhen, China, and employs evaluation metrics to objectively compare the models. The results indicate that AirPhyNet outperforms state-of-the-art deep learning models and classical approaches, providing a promising tool for air quality forecasting in various environmental settings. The paper concludes by suggesting future research directions, including refining the model architecture and exploring its applications in real-world environmental management.

Keywords: AirPhyNet; Air quality forecasting; Deep learning.

1. Introduction

The escalation of air pollution, stemming from diverse sources such as residential activities, vehicular emissions, and industrial operations, has evolved into a pressing global concern. This urgent challenge mandates meticulous forecasting and monitoring of air pollution levels, given their profound ramifications on human health and environmental sustainability. The intricate interplay of atmospheric pollutants poses formidable hurdles for conventional methods to accurately predict air quality, particularly amidst the backdrop of rapid urbanization and industrial expansion. Urban particulate air pollution, containing ultra-fine particles, is associated with increased respiratory disease exacerbations and cardiovascular deaths in older individuals [1].

Amidst these challenges, deep learning presents a beacon of hope for enhancing the precision of air quality forecasts. However, the current state-of-the-art is marred by models that often lack interpretability and consistency, primarily stemming from their disregard for the fundamental principles of atmospheric science [2]. By infusing insights gleaned from atmospheric physics into deep learning models, the reliability and comprehensibility of air quality forecasts can be augmented.

This study endeavours to bridge the divide between data-driven methodologies and the wealth of knowledge housed within the field of atmospheric science, with the ultimate aim of refining air quality forecasting. Leveraging insights garnered from prior research endeavours, a novel approach is proposed that amalgamates the strengths of deep learning techniques with the foundational principles of atmospheric physics [2]. This integration holds the promise of furnishing more accurate and interpretable predictions of air pollutant concentrations.

At the core of the research agenda lies the exploration and evaluation of a pioneering deep learning framework dubbed AirPhyNet, meticulously tailored for air quality prediction [3]. Methodologically, the approach entails a comprehensive review of literature, meticulous design of the AirPhyNet architecture, rigorous data collection and preprocessing, exhaustive model training and testing, and systematic comparison with existing baseline models [4]. By embedding principles from atmospheric

physics into the very fabric of the model structure, AirPhyNet aspires to attain unprecedented levels of accuracy and interpretability in air quality forecasting.

Through these concerted endeavours, the research endeavours to furnish invaluable insights that can inform the development of more effective strategies for environmental conservation and public health preservation. Machine learning algorithms can help predict air quality, but accurate results are still needed for public health measures [5]. Machine learning can accurately predict air quality levels in cities and areas by using trained artificial neural networks and autoregression methods [2]. Deep Learning algorithms outperform Machine Learning algorithms in predicting air pollution, with SimpleRNN reducing computational complexity by 98.90% and improving accuracy by 7.5% [6].This paper presents an air quality prediction model using machine learning methods, with an accuracy of over 90% in predicting pollutants' concentrations [7]. Air pollution contributes to climate change and human health issues, necessitating public awareness and multidisciplinary solutions [8]. Air pollution exposure significantly increases the risk of respiratory diseases, emphasizing the need for improved policy initiatives on air quality in both high- and low-income countries [9]. The updated World Health Organization Global Air Quality Guidelines set interim target levels for six key pollutants, aiming to improve air quality and reduce health risks [10].

2. Methods

2.1. Data Source

In this study, researchers utilised real-world air quality datasets collected from two urban centres in China, namely Beijing and Shenzhen. PM2.5 concentrations in Beijing between 14:00:00 on 1 January 2017 and 23:00:00 on 6 January 2017 were selected as the dataset, and the dataset contained a total of 130 sample points.

2.2. Indicator Explanation

In this study, in order to objectively compare the predictive effectiveness of each model, the coefficients of determination (R^2) Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE) are used as the evaluation indexes of the models. The related calculation formulas are shown in Equations (1)-(6).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{5}$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{(\hat{y}_i - |y_i|)/2}\right| \tag{6}$$

In equations (1)-(6), the n is the number of samples in the data set, and $\hat{y}_i$ is the model predicted value, and y_i is the true value.

2.3. Method Description

In this study, the ARIMA model is known as the Autoregressive Differential Moving Average Model, which is commonly used for modelling and forecasting of time series data. ARIMA model

combines Autoregressive (AR), Differential (I) and Moving Average (MA) models, the expression of the ARIMA model is shown in Equation (7).

$$Y_t = c + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (7)$$

In the above equation, c is the constant term, and r_i is the autocorrelation coefficient and ε_t is the error. After obtaining the data, a missing value test was carried out on the data and it was found that there was a missing value in the data, in order to facilitate the subsequent modelling, the missing value was treated with a mean substitution. In this study, the first 70% of the dataset was used as a training set to build the model, and the second 30% of the data was used as a test set to evaluate the model's generalisation ability.

3. Results and Discussion

3.1. Stationary Test

The initial time series plot is shown in Figure 1. In order to objectively determine the smoothness of the time series, the time series was subjected to the ADF smoothness test, which has a p-value of 0.03932<0.05, proving that the time series is a non-stationary time series.

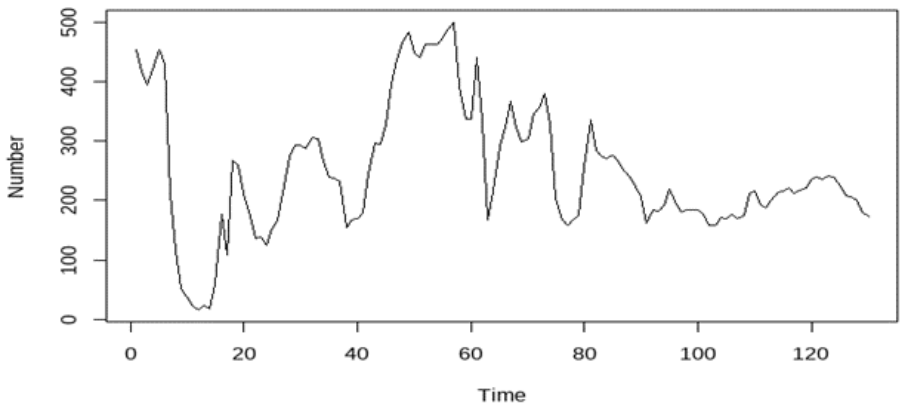

Fig. 1 Initial time series plot

The white noise test is performed on the smooth time series and the white noise test p-value is 2.2e-16, which is less than 0.05, indicating that the smooth time series is non-white noise and can be subjected to subsequent ARIMA modelling (Table 1).

Table 1. White noise test results

Box-Pierce test statistic	(number of) degrees of freedom	P-value
104.52	1	2.2e-16

3.2. Model Construction

In order to determine the p-value and q-value of the ARIMA model, plots of the autocorrelation function and partial autocorrelation function are shown in Figure 2 and 3.

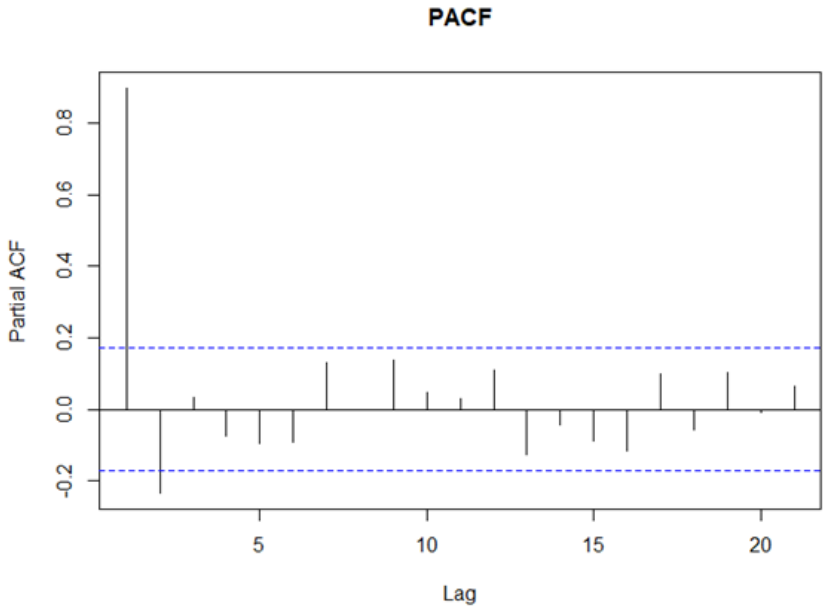

Fig. 2 Partial autocorrelation function plot

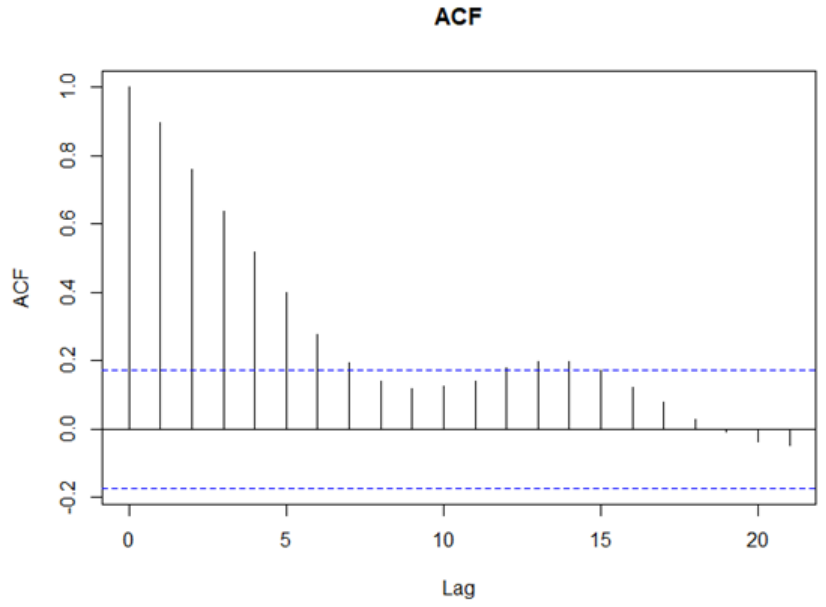

Fig. 3 Autocorrelation function plot

From Figure 2 and 3, the autocorrelation plot is truncated after order 15 and the partial autocorrelation plot is truncated after order 1, so it results in an ARIMA model with q=1 and p=15, so an ARIMA (1,0,15) model is established. ARIMA (1,0,15) model coefficient estimates are shown in Table 2.

In Table 2, intercept = 252.5847, which represents the predicted value of the model when the other independent variables are zero; ar1 = 0.4673, which means that there is a positive correlation between the current observation and the previous observation. ma1 = 0.7174, which means that there is a positive correlation between the current observation and the error at the previous time point. If ma1 is positive, it means that the current value is positively affected by the past error; if it is negative, it means that it is negatively affected.

Table 2. ARIMA model coefficient estimation

Name	Value
intercept	252.5847
ar1	0.4673
ma1	0.7174
ma2	0.4650
ma3	0.4292
ma4	0.4564
ma5	0.5252
ma6	0.4416
ma7	0.2903
ma8	0.1783
ma9	0.0155
ma10	0.0643
ma11	-0.2215
ma12	-0.0914
ma13	-0.0004
ma14	0.0066
ma15	-0.0302

3.3. Model Prediction Results

After the model building was completed and passed the residual white noise test, the prediction was made in the test set using the established ARIMA (2,1,4) model and the prediction results are shown in Figure 4.

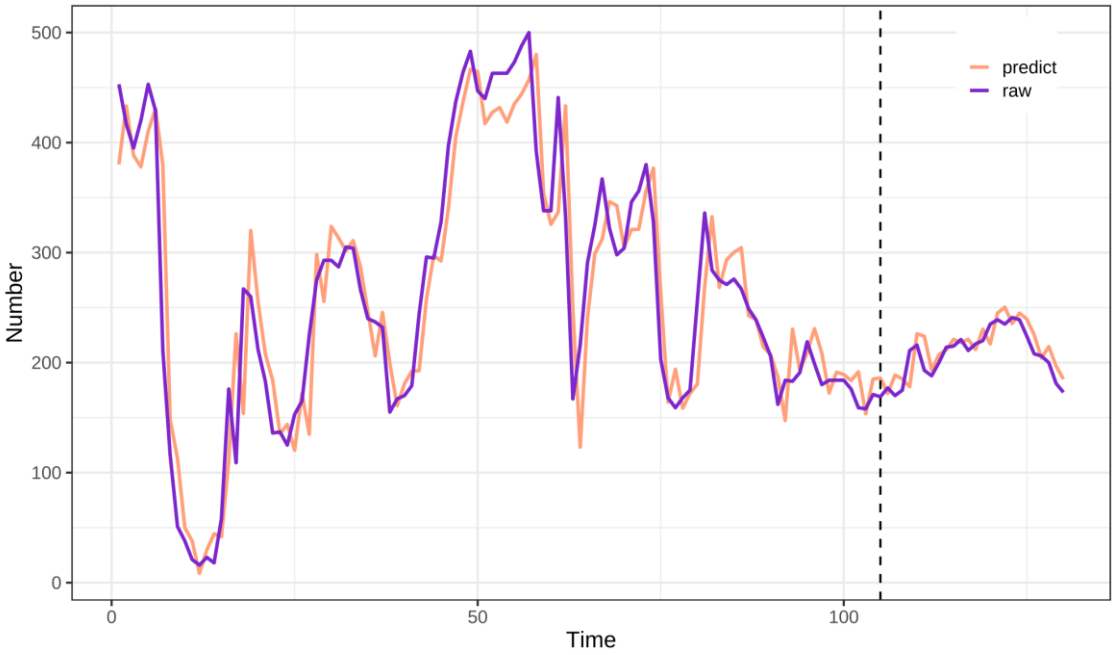

Fig. 4 ARIMA model prediction results

The evaluation metrics of the ARIMA model in the test set are shown in Table 3.As can be seen from Table 3, the ARIMA model predicted R2 = 0.61 in the test set, indicating that the model fits better; where the error evaluation indexes MSE = 197.01, RMSE = 14.04 and MAE = 11.67, indicating that the model predicts better; and the accuracy evaluation indexes MAPE = 0.06 and SMAPE = 0.06, which are close to 0, indicating that the model's medical precision is high.

Table 3. Evaluation metrics for the ARIMA model test set

Evaluation indicators	Value
R^2	0.61
MSE	197.01
RMSE	14.04
MAE	11.67
MAPE	0.06
SMAPE	0.06

After building the ARIMA (1,0,15) model, the residual QQ plots are shown in Figure 5. The residuals are tested for white noise, and the white noise test p-value is 0.8201, which is greater than 0.05. The combination of the QQ plot and the white noise test p-value indicates that the residuals are white noise.

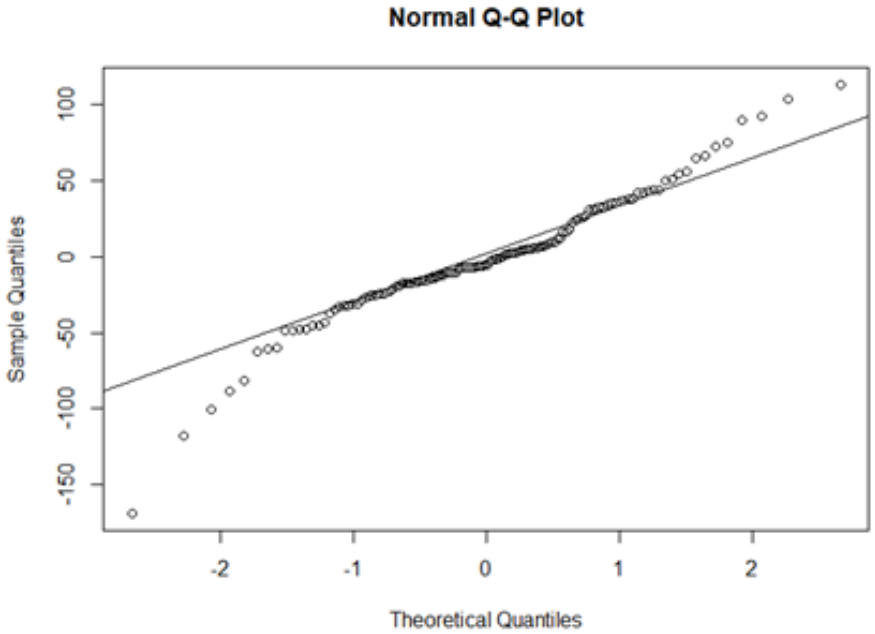

Fig. 5 Residual QQ diagram

4. Conclusion

With the unveiling of AirPhyNet, a significant leap forward has been achieved in air quality prediction. This deep learning framework is distinguished by its integration of atmospheric physics principles within its structure. The innovative fusion of these elements meets the urgent demand for models that are both accurate and interpretable, with the capacity to grasp the intricate dynamics of air pollution. AirPhyNet's excellence in forecasting air quality has been validated through extensive evaluation and analysis, demonstrating its superiority over contemporary deep learning models and conventional approaches.

The assimilation of physics-informed insights into the model's design not only elevates the precision of forecasts but also imparts a more profound comprehension of the factors driving changes in air quality. Such comprehension is vital for decision-makers and environmental stewards, enabling them to devise and execute strategies that effectively counteract air pollution and its detrimental impacts on public health and the ecosystem.

The resilience of AirPhyNet against abrupt fluctuations, coupled with its generalizability across varied datasets, renders it an adaptable instrument for predicting air quality in a multitude of environmental contexts. The model's applicability transcends scholarly pursuits, providing tangible advantages for urban development, public health endeavors, and the preservation of nature.

Prospective research may concentrate on enhancing the AirPhyNet framework to manage more intricate datasets and incorporate additional variables that affect air quality. This enhancement could entail the integration of real-time data from diverse sources, such as vehicular traffic and meteorological conditions, to amplify the model's predictive prowess. Moreover, deploying AirPhyNet in various global locales could yield insights into its versatility and efficacy across different atmospheric and pollution scenarios.

Subsequent investigative paths might delve into refining the architectural design of the model, embedding further domain-specific knowledge, and probing its utility in practical environmental governance contexts.

References

[1] Seaton A, et al. Particulate air pollution and acute health effects. The Lancet, 1995, 345: 176-178.

[2] Mazinani A, Davoli L, Pau D, Ferrari G. Air Quality Estimation with Embedded AI-Based Prediction Algorithms. 2023 International Conference on Information Technology Research and Innovation (ICITRI), 2023, 87-92.

[3] Mao W, Wang W, Jiao L, Zhao L, Liu A. Modelling air quality prediction using a deep learning approach: Method optimisation and evaluation. Sustainable Cities and Society, 2020, 102567.

[4] Kumari N, et al. Prediction of Air Quality in Industrial Area. 2020 International Conference on Recent Trends on Electronics, Information, Communication&Technology (RTEICT), 2020, 193-198.

[5] Madan T, Sagar S, Virmani D. Air Quality Prediction using Machine Learning Algorithms –A Review. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICANN), 2020, 140-145.

[6] Khan S, Yadav M. FORECASTING AIR QUALITY USING MACHINE LEARNING. International Journal of Engineering and Applied Science Trends, 2020, 5, 420-426.

[7] Shi C, Wang Y, Wan Y, Wu S. Air Quality Prediction Based on Machine Learning. 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), 2022, 1-5.

[8] Manisalidis I, et al. Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health, 2020, 8.

[9] Bălă G, et al. Air pollution exposure-the visible risk factor for respiratory diseases. Environmental Science and Pollution Research International, 2021, 28.

[10] Goshua A, Akdis C, Nadeau K. World Health Organization global air quality guideline recommendations: Executive summary. Allergy, 2021, 77: 1955-1960.