Journal of Computer and Communications, 2025, 13(2), 142-154
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Hybrid Air Quality Prediction Method Based on VAR and Random Forest

Minghao Yi, Fuming Lin*

College of Mathematical and Statistics, Sichuan University of Science and Engineering, Zigong, China
Email: *2889455243@qq.com

How to cite this paper: Yi, M.H. and Lin, F. (2025) A Hybrid Air Quality Prediction Method Based on VAR and Random Forest. *Journal of Computer and Communications*, **13**, 142-154.
https://doi.org/10.4236/jcc.2025.132009

Received: January 23, 2025
Accepted: February 21, 2025
Published: February 24, 2025

Abstract

To improve the efficiency of air quality analysis and the accuracy of predictions, this paper proposes a composite method based on Vector Autoregressive (VAR) and Random Forest (RF) models. In the theoretical section, the model introduction and estimation algorithms are provided. In the empirical analysis section, global air quality data from 2022 to 2024 are used, and the proposed method is applied. Specifically, principal component analysis (PCA) is first conducted, and then VAR and Random Forest methods are used for prediction on the reduced-dimensional data. The results show that the RMSE of the hybrid model is 45.27, significantly lower than the 49.11 of the VAR model alone, verifying its superiority. The stability and predictive performance of the model are effectively enhanced.

Keywords

Var Model, Principal Component Analysis, Random Forest Model

1. Introduction

With the rapid development of the global economy and the acceleration of urbanization, environmental issues have become increasingly prominent. Among these, air pollution has emerged as a major global challenge. Air pollution not only poses serious threats to human health but also exerts profound impacts on global climate change. According to a report by the World Health Organization (WHO), approximately 90% of the global population lives in environments with substandard air quality. Health issues caused by air pollution, such as respiratory and cardiovascular diseases, have become urgent public health problems that governments worldwide need to address [1].

In recent years, with the rapid advancement of data technologies, the use of big data and statistical methods for comprehensive analysis and prediction of air

quality has become an essential tool for environmental protection and policy-making [2]. However, existing methods still face significant limitations:

- **Linear Assumptions of VAR Models:** While the VAR model captures linear dynamic dependencies in multivariate time series [3], its linear structure struggles to describe complex nonlinear interactions (e.g., the combined effects of PM2.5 with temperature and humidity). For instance, [4] demonstrated that VAR models exhibit weak explanatory power for nonlinear responses when analyzing lagged meteorological impacts on air quality, leading to constrained prediction accuracy.
- **Time Series Modeling Limitations of Random Forest:** Although Random Forest excels in handling high-dimensional nonlinear data [5], its independent tree structure lacks inherent capability to model temporal dependencies. As noted by [6], directly applying Random Forest to air quality prediction may ignore autocorrelation and intertemporal dependencies of pollutant concentrations, thereby reducing long-term forecasting stability.
- **Information Loss Risks in PCA:** PCA effectively reduces dimensionality [7], but [8] found that minor components in complex environmental data may contain critical seasonal patterns of specific pollutants. Excessive dimensionality reduction could diminish sensitivity to localized features.

To address these limitations, this study proposes a hybrid framework integrating PCA, VAR, and Random Forest. By combining the nonlinear modeling capacity of Random Forest with the temporal dependency analysis of VAR, and optimizing data structure through PCA, our approach mitigates the shortcomings of individual methods.

First, Principal Component Analysis (PCA), a dimensionality reduction technique, has been widely applied in environmental studies. Pearson (1901) [9] and Hotelling (1933) were pioneers in proposing and developing PCA. Its primary advantage lies in effectively reducing data dimensions while retaining most of the in-formation, thus simplifying data analysis [10]. Liu *et al.* (2019) utilized PCA to extract key components of air pollutants and identified PM2.5 as one of the critical factors affecting air quality.

Second, the Vector Autoregressive (VAR) model has been extensively used in time series analysis to study dynamic relationships among variables [11] [12]. Sims (1980) initially introduced the VAR model to address lag effects and interactions in multivariate time series. Liu *et al.* applied the VAR model to investigate the lag effects of PM2.5 concentrations, revealing a significant temporal dependence in air pollution. Kou *et al.* further employed the VAR model to analyze the relationship between air quality and meteorological conditions, highlighting the lag effects of factors, such as temperature and humidity on air quality. Qiu *et al.* utilized the VAR model to study the dynamic response mechanism between environmental regulation and agricultural carbon emissions, providing insights into the model's application.

Lastly, the Random Forest model, an ensemble algorithm based on decision

trees, has garnered attention in environmental studies due to its superior performance in handling high-dimensional and noisy data. Breiman first proposed the Random Forest algorithm, and numerous studies have demonstrated its predictive performance surpasses traditional linear regression models in environmental forecasting.

In summary, building on previous research, this study integrates PCA, the VAR model, and the Random Forest model [13]-[15] to develop a new meta-model aimed at providing a more accurate and robust tool for air quality analysis and prediction.

The main contributions of this study are twofold. First, based on existing air quality analysis methods, this study proposes a novel analytical framework combining PCA, the VAR model, and the Random Forest model. By integrating statistical analysis with machine learning models, the proposed framework not only effectively reduces data dimensions but also enhances the accuracy of air quality predictions, enriching the theoretical foundation for air quality forecasting. Second, this study explores the interactions between various environmental variables and air quality, deepening the understanding of the causes of air pollution. This can provide valuable decision-making references for governments and assist in formulating more effective environmental policies. Moreover, the integration of data technologies and statistical methods in this study offers a research pathway for future environmental protection efforts and promotes the application of big data technologies in environmental management.

2. Theoretical Framework

Since Principal Component Analysis (PCA) is a well-known method, this section primarily introduces the Vector Autoregressive (VAR) model, the Random Forest model, and the proposed framework in this study.

2.1. Vector Autoregressive (VAR) Model

The VAR model is used to describe the interdependencies among multiple time series. It predicts the current and future values of time series by considering the lagged values of multiple variables, making it suitable for handling multivariate time series data.

The basic equation of the VAR model is as follows:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \epsilon_t \tag{1}$$

where:

- c: a constant vector (intercepts);
- $A_1, A_2, \cdots, A_p$: lag coefficient matrices, representing the effects of each lag period;
- ϵ_t: the error term, assumed to be white noise.

To predict the value at a future time $t+h$, the above equation is recursively applied:

$$y_{t+h} = c + A_1 y_{t+h-1} + A_2 y_{t+h-2} + \cdots + A_p y_{t+h-p} \tag{2}$$

2.2. Random Forest Model

Random Forest is an ensemble learning method that improves model accuracy and robustness by constructing multiple decision trees and combining their prediction results.

For regression problems, the prediction of Random Forest is expressed as:

$$\hat{y} = \sum_{b=1}^{B} T_b(x) \tag{3}$$

where:

- B : the number of decision trees in the Random Forest;
- $T_b(x)$: the prediction result of the b-th decision tree for the input features x;
- $\hat{y}$: the final prediction, which is the average of the predictions from all trees.

2.3. Proposed Framework

The basic idea of the proposed method is as follows: First, the original dataset is divided into a training set and a testing set. A Random Forest model is trained using the training set and is then used to predict the target values for the time steps in the testing set. The training data is combined with the predictions obtained from the Random Forest model to form a new dataset, which is then used to fit and forecast using the VAR model.

2.3.1. Theoretical Derivation of RF-VAR Synergy

The synergy stems from two complementary mechanisms:

- **Nonlinear Pattern Capture**: RF approximates complex mappings through ensemble trees:

$$\hat{y}_{RF}^{k,t} = \frac{1}{B}\sum_{b=1}^{B} T_b^{(k)}(X_t) \quad \text{(Tree Ensemble)} \tag{4}$$

where $T_b^{(k)}$ denotes the b-th tree for variable k, effectively modeling interactions between lagged terms and external factors.

- **Linear Temporal Dependency**: VAR enforces Granger causality constraints through its coefficient matrices:

$$\frac{\partial y_{k,t}}{\partial y_{j,t-i}} = A_{k,j,i} \quad \text{(Linear Propagation)} \tag{5}$$

The hybrid model combines these via additive decomposition:

$$y_t = \underbrace{\sum_{i=1}^{p} A_i y_{t-i}}_{\text{Linear Dynamics}} + \underbrace{\beta \cdot RF(X_t)}_{\text{Nonlinear Correction}} + \epsilon_t \tag{6}$$

2.3.2. Implementation Framework

First, use the Random Forest model to predict each variable:

$$\hat{y}_{RF}^{k,t} = RF_k\left(X_t\right) \tag{7}$$

where:

- $\hat{y}_{RF}^{k,t}$: the prediction of the k-th variable at time t by the Random Forest model;
- RF_k : the Random Forest model for predicting the k-th variable;
- X_t : the feature vector for the Random Forest model, typically including lagged values of all variables.

Feature Engineering Details:

$$X_t = \left[y_{1,t-1}, y_{1,t-2}, \cdots, y_{1,t-p}, \cdots, y_{K,t-p}\right] \in \mathbb{R}^{K\times p} \tag{8}$$

where K is the number of variables, and p is the maximum lag order (determined via AIC/BIC).

Each RF model RF_k is trained to minimize:

$$\min_{\theta_k} \sum_{t=p+1}^{N_{\text{train}}} \left(y_{k,t} - RF_k\left(X_t\right)\right)^2 \tag{9}$$

where θ_k represents tree split parameters.

RF predictions may introduce future information, causing endogeneity bias. To address endogeneity issues:

- **Strict Sample Partitioning**:

  - Training set $Y_{\text{train}} = \left\{y_1, y_2, \cdots, y_T\right\}$, test set $Y_{\text{test}} = \left\{y_{T+1}, \cdots, y_{T+h}\right\}$.

  - Train RF solely on Y_{train}. For prediction, input $X_{T+1} = \left[y_T, y_{T-1}, \cdots, y_{T-p+1}\right]$ to generate $\hat{y}_{T+1}^{RF}$.

- **Rolling Prediction (Walk-Forward Validation)**: For each test time step $t \in \left\{T+1, T+2, \cdots\right\}$:

  (a) Dynamically update the training window.

  (b) Retrain RF and VAR models to ensure predictions rely only on historical data.

Next, incorporate the Random Forest predictions into the VAR model:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \beta \hat{y}_t^{RF} + \epsilon_t \tag{10}$$

where:

- $y_t = \left[y_{1,t}, y_{2,t}, \cdots, y_{K,t}\right]^{\mathrm{T}}$: the K-dimensional time series vector at time t;
- c : a K-dimensional constant vector (intercepts);
- $A_1, A_2, \cdots, A_p$: $K \times K$ coefficient matrices for different lags;
- $\hat{y}_t^{RF} = \left[\hat{y}_{1,t}^{RF}, \hat{y}_{2,t}^{RF}, \cdots, \hat{y}_{K,t}^{RF}\right]^{\mathrm{T}}$: the prediction vector of the Random Forest model at time t;
- β : a $K \times K$ coefficient matrix measuring the impact of the Random Forest predictions on actual values;
- ϵ_t : the error term, typically assumed to follow a multivariate normal distribution with zero mean and covariance matrix Σ.

The comprehensive model equation becomes:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \beta \cdot RF\left(X_t\right) + \epsilon_t \tag{11}$$

For each variable k, the equation can be expressed as:

$$y_{k,t} = c_k + \sum_{i=1}^{p}\sum_{j=1}^{K} A_{k,j,i} y_{j,t-i} + \sum_{m=1}^{K} \beta_{k,m} \hat{y}_{m,t}^{RF} + \epsilon_{k,t}, \quad \forall k = 1,2,\cdots,K \tag{12}$$

where:

- c_k : the intercept for the k -th variable;
- $A_{k,j,i}$: the coefficient of the j -th variable at the i -th lag for the k -th variable;
- $\beta_{k,m}$: the coefficient measuring the impact of the Random Forest prediction of the m -th variable on the k -th variable;
- $\epsilon_{k,t}$: the error term for the k -th variable at time t .

3. Algorithm

The algorithm is described as follows:

1) Data Preprocessing & Dimensionality Reduction

- **Input:** Raw dataset Y ($N \times K$ matrix, N : timesteps, K : variables)
- **Process:**

* Standardize Y .

* Apply PCA to extract top M principal components, obtaining reduced data Y_{pca} ($N \times M$).

- **Output:** Reduced-dimension dataset Y_{pca}

2) Data Splitting

- **Input:** Y_{pca}
- **Process:**

* Split Y_{pca} chronologically into training (70%), testing (15%), and forecast (15%) sets: Y_{train} , Y_{test} , Y_{forecast} .

- **Output:** Y_{train} , Y_{test} , Y_{forecast}

3) Random Forest Training & Prediction

- **Input:** Y_{train}
- **Process:**

* Train independent RF models RF_k for each variable k :

$$\min_{\theta_k} \sum_{t=p+1}^{N_{\text{train}}} \left(y_{k,t} - \text{RF}_k \left(Y_{\text{train}}[t] \right) \right)^2$$

with parameters: $B = 100$, random seed = 42.

* Predict testing set Y_{test} using lagged test features X_{test} .

- **Output:** Predicted test set $\hat{Y}_{\text{test}}$

4) Data Merging & VAR Fitting

- **Input:** Y_{train} , $\hat{Y}_{\text{test}}$
- **Process:**

* Merge datasets chronologically: $Y_{\text{new}} = \left[Y_{\text{train}}; \hat{Y}_{\text{test}} \right]$.

* Fit extended VAR model with RF predictions as exogenous variables:

$$y_t = c + \sum_{i=1}^{p_{\text{var}}} A_i y_{t-i} + \beta \hat{y}_t^{\text{RF}} + \epsilon_t$$

where $p_{var} = m$ (selected via AIC).

- **Output:** Fitted VAR coefficients A_i, β.

5) Forecasting & Evaluation

- **Input:** Fitted VAR model, $Y_{forecast}$.
- **Process:**

* Generate forecasts $\hat{Y}_{forecast}$.

* Compute RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{N_{forecast}} \sum_{t=1}^{N_{forecast}} \left(Y_{forecast}[t] - \hat{Y}_{forecast}[t] \right)^2}$$

- **Output:** Forecast results $\hat{Y}_{forecast}$, RMSE value.

The algorithm flow chart is as follows Figure 1:

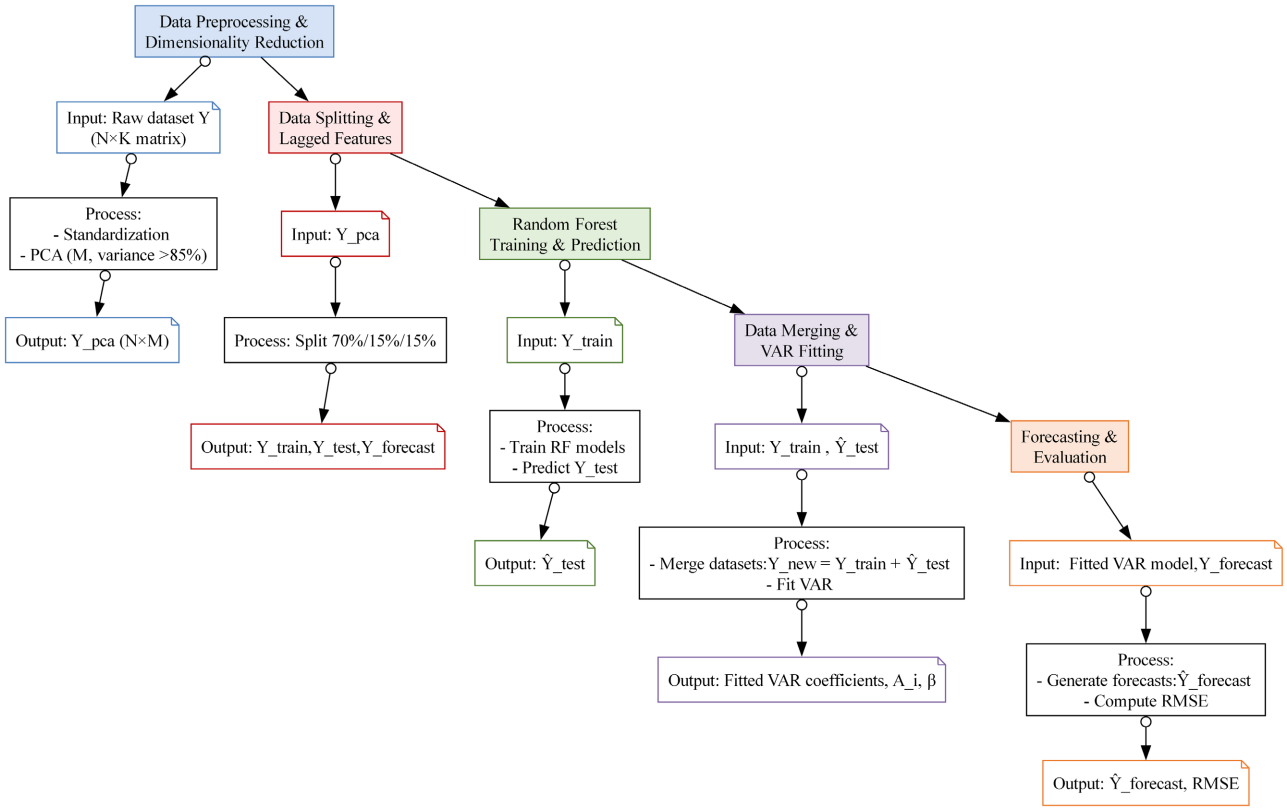

Figure 1. The algorithm flow chart.

4. Empirical Analysis

4.1. Data and Statistical Description

The data used in this study is sourced from the Kaggle dataset, which includes air quality data (AQI dataset) for various continents from 2022 to 2024, as well as country/region mappings, ISO codes, and continent/region data (Country/Region Mapping - ISO, Continent, Region). The two datasets were integrated and matched according to continents, resulting in a new dataset that includes variables such as year, month, and AQI values. By combining the air quality data with the world

region classification data, visualization analysis was conducted in Python, and the results are as follows:

This Figure 2 shows the frequency distribution of global Air Quality Index (AQI) statuses in the dataset. The X-axis represents the AQI statuses, including "Good," "Moderate," "Unhealthy for Sensitive Groups," "Unhealthy," "Very Unhealthy," and "Hazardous." The Y-axis represents the frequency of each status globally. The values above each bar indicate the number of occurrences of each AQI status, showing how many countries fall under each category.

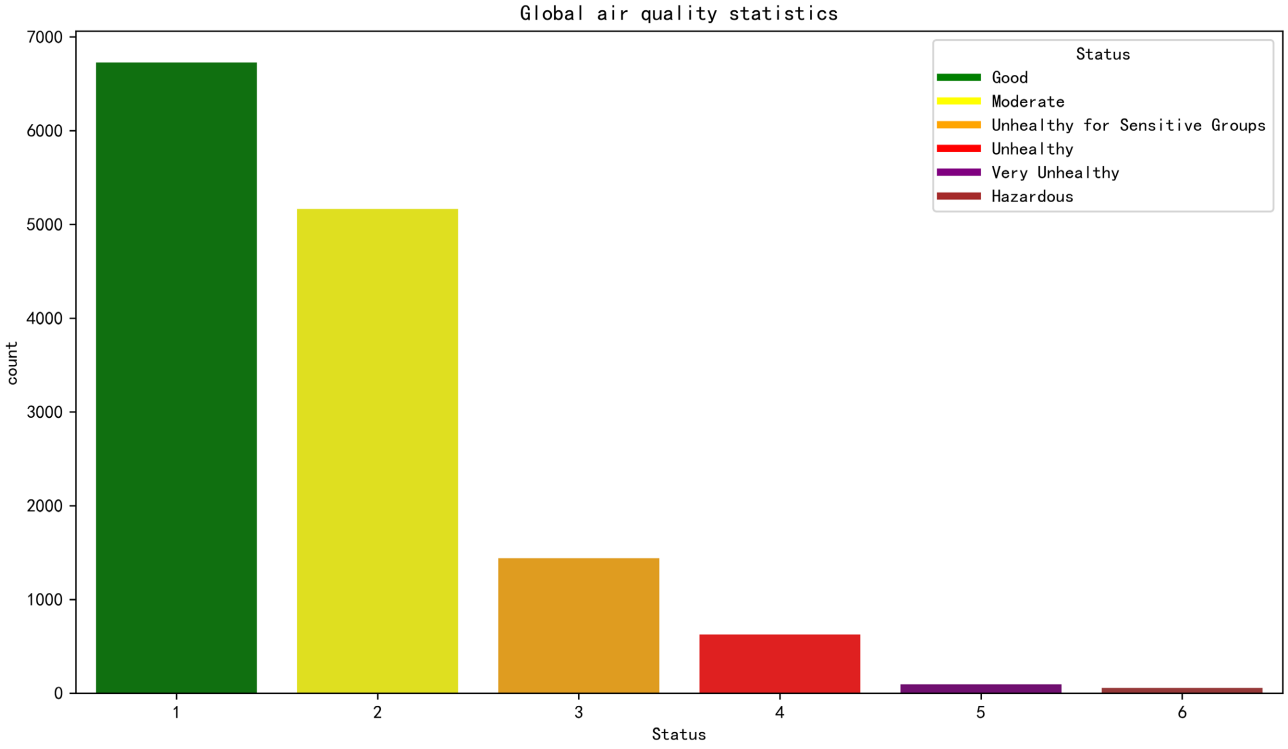

Figure 2. Global air quality statistics.

This Figure 3 compares the distribution of air quality statuses across different continents, displaying the air quality classification for each region. The X-axis represents the regions, such as Asia, Europe, Africa, etc., while the Y-axis shows the frequency of each air quality status within the region. The values above each bar indicate the specific count of each air quality status in the region, illustrating the distribution of air quality in that area.

This Figure 4 illustrates the distribution of the Air Quality Index (AQI) across different continents, visualized using a boxplot. The X-axis represents regions, such as Asia, Europe, Africa, etc., while the Y-axis shows the AQI values, where higher values indicate more severe air pollution. The median line in the boxplot represents the median AQI for each region (shown as the values on each box). The box represents the inter-quartile range of AQI values in each region, reflecting the central tendency and range of variation in air quality. The points marked as "outliers" represent abnormal values, which are AQI readings that are far from

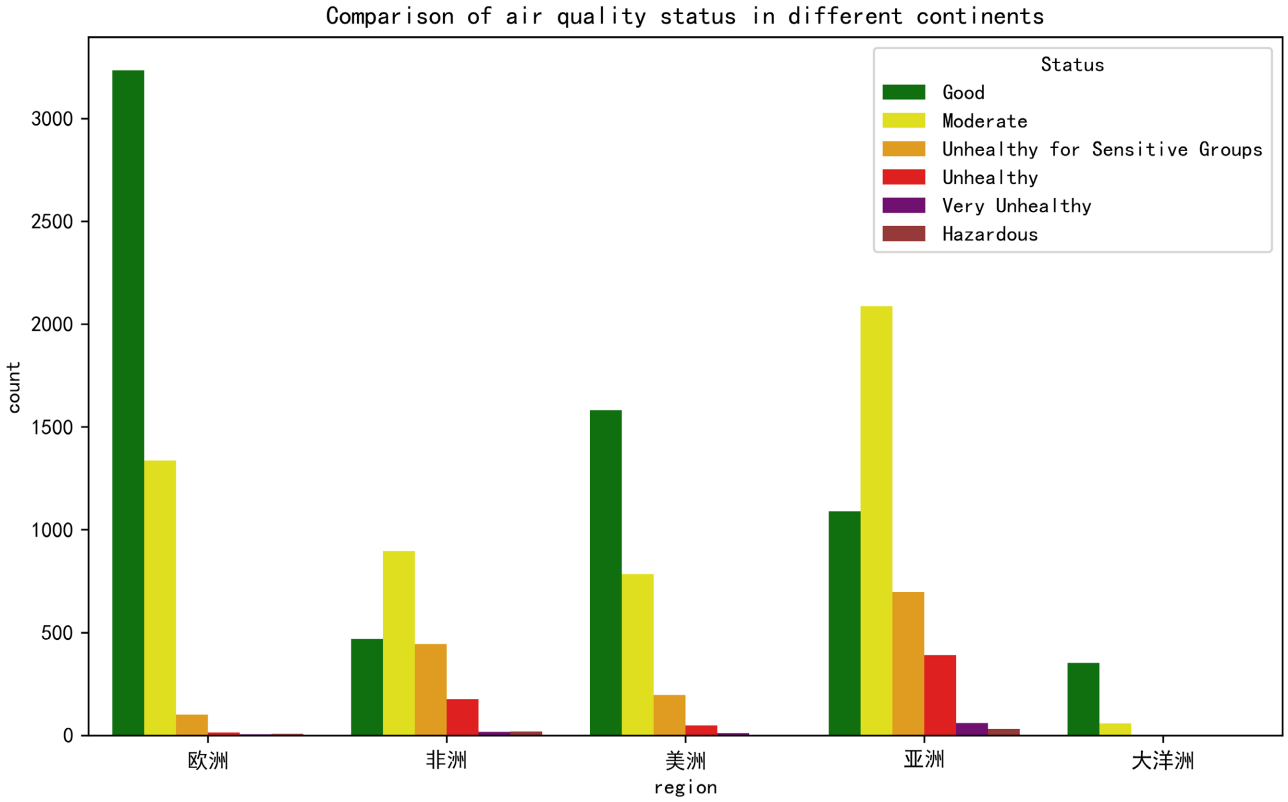

Figure 3. Comparison of air quality status in different continents.

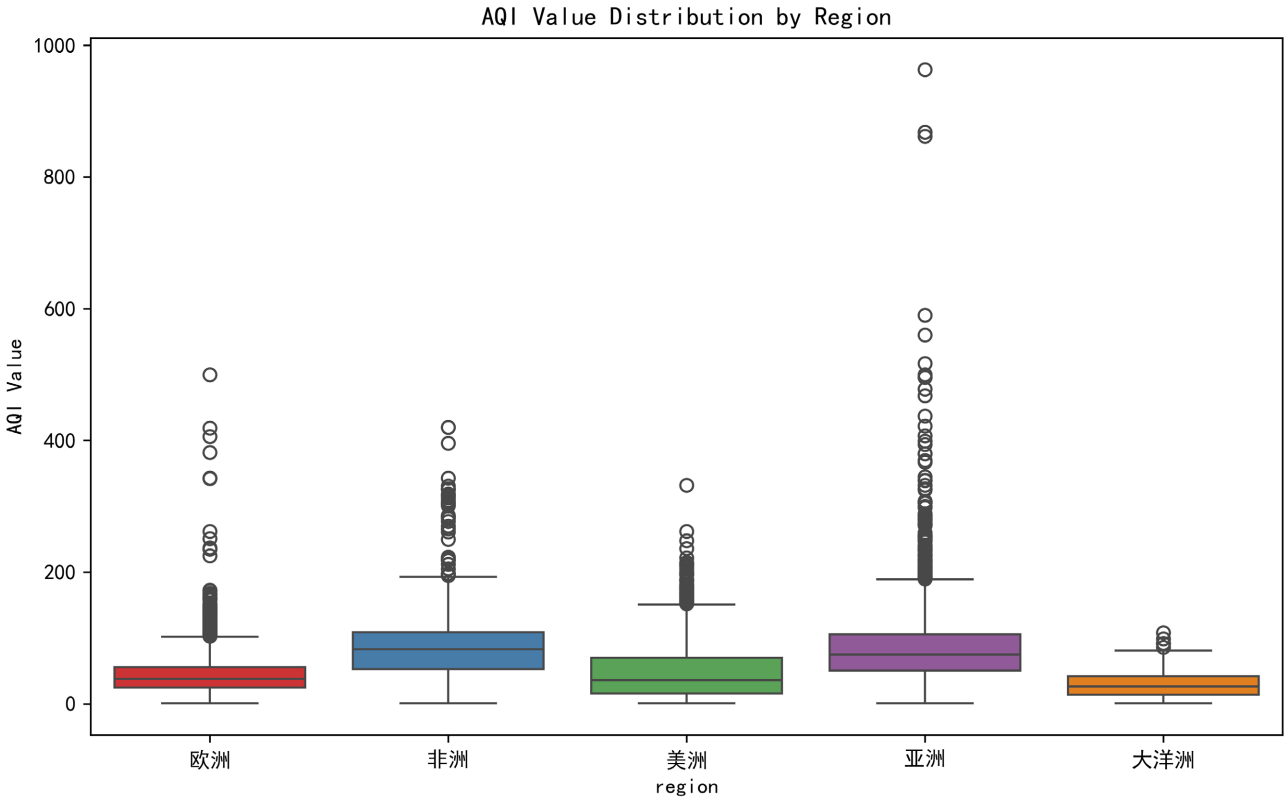

Figure 4. AQI value distribution by region.

the majority of the data points.

4.2. Algorithm 1

Algorithm 1 PCA and VAR Model Combined for Prediction

1: Dimensionality reduction using Principal Component Analysis (PCA);
2: Data splitting: set the ratio of training set, testing set, and prediction set as 70%, 15%, and 15%, respectively;
3: Fit the VAR model;
4: VAR model prediction and RMSE calculation;

4.3. Algorithm 2

Algorithm 2 Prediction Algorithm Combining PCA, Random Forest and VAR Model

1: Dimensionality reduction using Principal Component Analysis (PCA);
2: Data splitting: set the ratio of training set, testing set, and prediction set as 70%, 15%, and 15%, respectively;
3: Random Forest model training and prediction;
   1. Initialize the Random Forest regressor, setting the number of trees and random seed;
   2. Train the model: use the training set as features and the testing set's time steps as the target time steps, then use the model to predict the testing set and obtain the new testing set;
4: Integrate the new dataset;
   1. Restore the new testing set data to the original feature space;
   2. Set the new dataset as the combined training set, new testing set, and prediction set;
5: Perform dimensionality reduction on the new dataset;
6: Use the new dataset to fit the VAR model;
7: VAR model prediction and RMSE calculation;

Through the calculations of the two algorithms, the RMSE values of the VAR model without Random Forest training and the VAR model with Random Forest training are shown in Table 1:

Table 1. Model comparison.

Model	RMSE Value
VAR RMSE	49.1113
NEW Model RMSE	45.2686

In this method, the Random Forest regression model is used, with dimensionality-reduced features for training. By constructing multiple decision trees and employing ensemble learning, the model effectively identifies and models nonlinear patterns in the data. Additionally, the robustness of the Random Forest model enhances the overall model's performance when dealing with data noise and outliers. The linear modeling capability of the VAR model ensures that the synergistic effects in multivariate time series are effectively utilized. By combining the strengths of both, it is evident that the RMSE value obtained by the proposed

method for predicting the forecast set length is lower than that obtained by the VAR model alone, indicating superior performance compared to the single VAR model.

The prediction results of the two models are shown in Figure 5:

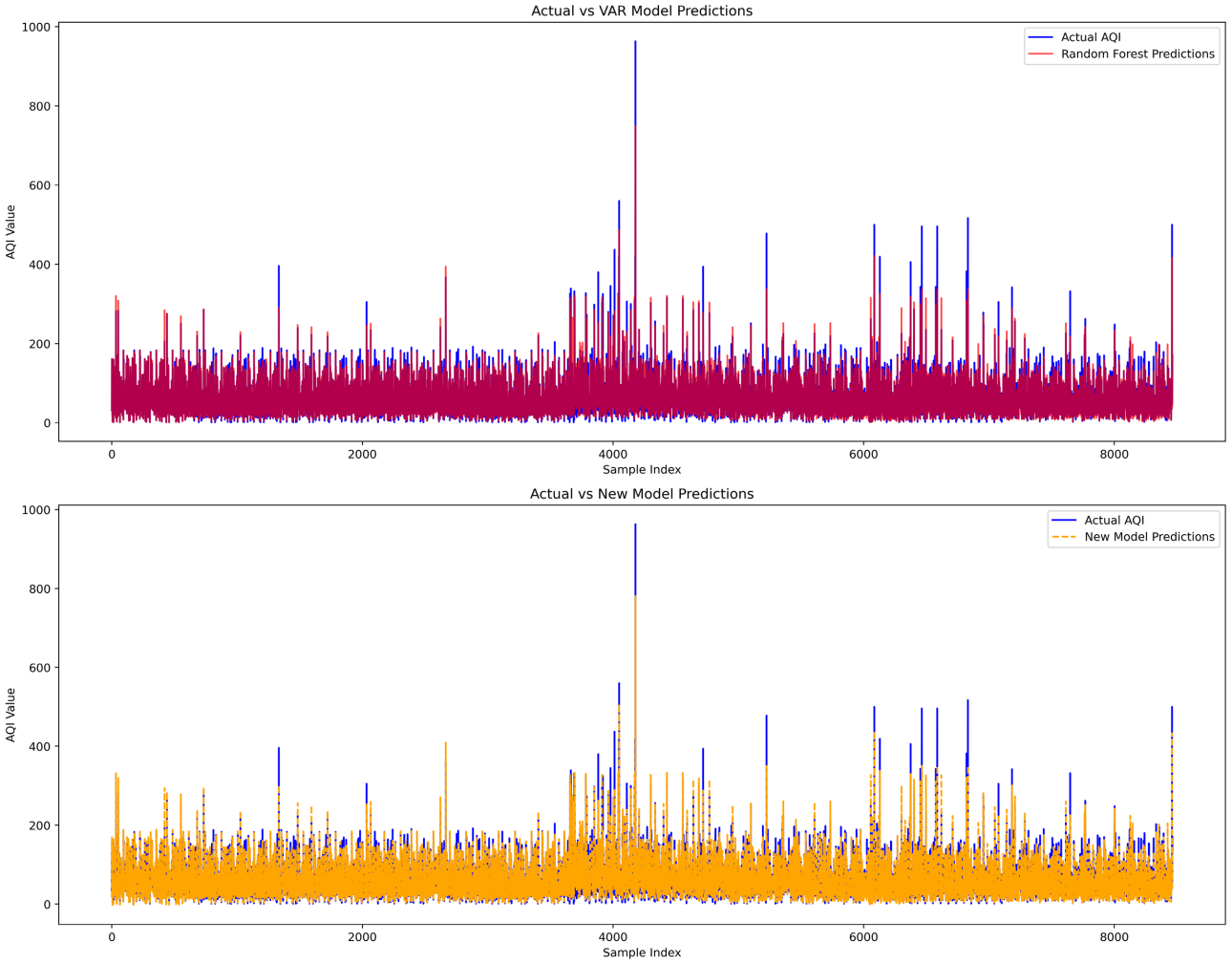

Figure 5. Prediction results.

5. Conclusion

In this study, a new model for air quality prediction was developed by combining PCA, the VAR model, and the Random Forest regression model. By applying dimensionality reduction, the model effectively simplified the input variables while retaining key information. During the subsequent modeling process, the Random Forest model successfully captured nonlinear relationships and complex patterns in the data, while the VAR model excelled at handling linear dependencies in multivariate time series. The model, which combines the strengths of both, significantly improved the accuracy and stability of air quality prediction. Experimental results show that the combination of the Random Forest model and the VAR model has a clear advantage in handling high-dimensional data, especially in

terms of robustness when dealing with data noise and outliers. The proposed meta-model framework can be further extended and optimized, offering great potential for applications in big data environments. Future research can explore the integration of more machine learning algorithms or time series models to further enhance the accuracy of air quality prediction. Moreover, to improve the model's generalization ability, more effective methods can be explored to handle air quality data from different regions and various climate conditions. These improvements will provide stronger data support for environmental management and policy-making and promote the further application of big data technologies in environmental science.

Acknowledgements

This paper was completed under the careful guidance of my supervisor, Associate Fuming Lin. In the determination of the research scheme, theoretical analysis, data processing, and the writing and finalization of the article, Lin gave me attentive teaching and selfless help. On the occasion of completing this paper, I would like to express our deep gratitude to Lin. This little-by-little achievement is the condensation of our blood and sweat. No matter where I go, your teachings will always be engraved in my heart.

Funding

This work is partly supported by the Graduate Textbook Construction Project of Sichuan University of Science and Engineering (Grant No. KA202011) and the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (2024QYY02).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (2021) Global Air Quality Guidelines. WHO Publications.

[2] Liu, J., He, C., Si, Y., Li, B., Wu, Q., Ni, J., *et al.* (2024) Toward Better and Healthier Air Quality: Global $PM_{2.5}$ and O_3 Pollution Status and Risk Assessment Based on the New WHO Air Quality Guidelines for 2021. *Global Challenges*, **8**, Article ID: 2300258. https://doi.org/10.1002/gch2.202300258

[3] Sims, C.A. (1980) Macroeconomics and Reality. *Econometrica*, **48**, 1-48. https://doi.org/10.2307/1912017

[4] Kou, L., Liao, J., Li, X., *et al.* (2022) Climate Change Prediction in Canada Based on VAR Model. *Computer and Modernization*, **10**, 13-18.

[5] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[6] Qiu, W. and Lu, D. (2019) Analysis of Factors Affecting Agricultural Carbon Emission Based on VAR Model and Its Dynamic Response Mechanism. *Hubei Agricultural*

Sciences, **58**, 271-276.

[7] Hotelling, H. (1933) Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**, 417-441. https://doi.org/10.1037/h0071325

[8] Liu, H. and Zhang, H. (2019) Atmospheric Environmental Quality Evaluation of City Based on Principal Component Analysis. *China Resource Comprehensive Utilization*, **37**, 141-143. (In Chinese)

[9] Pearson, K. (1901) LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space*. *The London*, *Edinburgh*, *and Dublin Philosophical Magazine and Journal of Science*, **2**, 559-572. https://doi.org/10.1080/14786440109462720

[10] Richards, L.E. and Jolliffe, I.T. (1988) Book Review: Principal Component Analysis. *Journal of Marketing Research*, **25**, 410-410. https://doi.org/10.1177/002224378802500410

[11] Hou, P.S., Fadzil, L.M., Manickam, S. and Al-Shareeda, M.A. (2023) Vector Autoregression Model-Based Forecasting of Reference Evapotranspiration in Malaysia. *Sustainability*, **15**, Article No. 3675. https://doi.org/10.3390/su15043675

[12] Nachouki, M., Mohamed, E.A., Mehdi, R. and Abou Naaj, M. (2023) Student Course Grade Prediction Using the Random Forest Algorithm: Analysis of Predictors' Importance. *Trends in Neuroscience and Education*, **33**, Article ID: 100214. https://doi.org/10.1016/j.tine.2023.100214

[13] Lin, J. and He, J. (2022) Parallel Random Forest Prediction Algorithm Based on PCA Stratified Sampling in the Big Data Environment. China Management *Informationization*, **25**, 172-176. (In Chinese)

[14] Huang, S. and Zhang, Z. (2023) Study on the Dynamic Relationship between Energy Consumption and Environmental Pollution in Chongqing City: Empirical Analysis Based on VAR Model. *China-Arab States Science and Technology Forum* (*Chinese and English*), **2023**, 23-27. (In Chinese)

[15] Wei, X. (2023) Treatment and Application of Outlier in VAR Model. *Science and Technology and Economy*, **36**, 101-105. (In Chinese) https://doi.org/10.14059/j.cnki.cn32-1276n.2023.06.021