Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

Predicting PM2.5 Concentrations Based on Vehicle Ownership Using Big Data Models and Machine Learning

Zhihe Yang *

School of Mathematics and Statistics, Beihua University, Jilin, China, 132013

* Corresponding Author Email: 15143492589@163.com

Abstract. This study integrates multi-source satellite-remote-sensed PM2.5 data and provincial vehicle ownership data in China (1998-2023) to construct a high-precision prediction model. We developed a spatiotemporally coupled hybrid model that fuses time series analysis (Prophet) with graph convolutional networks, and incorporates a reinforcement learning framework for dynamic feature optimization. The model achieves an R^2 of 0.95, with a mean absolute error of 1.97 $\mu g/m^3$ and a root mean square error of 2.67 $\mu g/m^3$ on the 2022-2023 test set, representing a 12% improvement over single models. Analysis reveals significant regional heterogeneity, indicating a stronger PM2.5 suppression effect from new energy vehicles in eastern coastal areas compared to inland regions. This study provides a robust quantitative tool for vehicle pollution control and offers a methodological framework for multi-scale environmental modeling.

Keywords: PM2.5 Prediction; Spatiotemporal Coupling Model; Reinforcement Learning.

1. Introduction

In recent years, driven by technological advancements and rapid economic growth, new energy vehicles have received strong support in China [1-2]. Their zero-emission or low-emission characteristics offer significant advantages in reducing exhaust pollution. However, PM2.5, as one of the primary pollutants affecting air quality, poses substantial risks to human health. Research indicates that PM2.5 sources are complex, with secondary particulate matter generated by human activities constituting a high proportion, and traffic dust and motor vehicles being major contributors [3]. Therefore, accurately predicting and quantifying the impact of motor vehicle ownership on PM2.5 concentrations is crucial for formulating refined environmental control policies [4-5].

Given this context, the core objective of this study is to construct a high-precision, high-interpretability PM2.5 prediction model that effectively quantifies the influence of motor vehicle ownership and its associated characteristics, while overcoming the limitations of traditional time series models in handling regional heterogeneity, nonlinear threshold effects, and structural breaks. Specifically, this requires addressing: how to establish a preliminary PM2.5 prediction model based on time series and covariates while diagnosing its limitations; and how to design a high-precision hybrid PM2.5 prediction architecture integrating spatiotemporal features with reinforcement learning [6].

To address these challenges, this paper proposes a hierarchical, progressive research approach. First, provincial-level panel data from China spanning 1998–2023 undergoes data processing, including alignment, missing value and outlier handling, followed by ADF testing to assess time series stationarity. Second, we preliminarily employ a seasonal autoregressive model within the panel data framework, incorporating motor vehicle ownership as an exogenous variable for predictive analysis. We diagnose the model's shortcomings based on test results and residual analysis (e.g., Ljung-Box test and heteroskedasticity test). Third, SHAP value decomposition was employed to identify error sources, providing quantitative evidence for model refinement. Finally, a Random Forest regression model was introduced, leveraging its strengths to construct an innovative Spatio-Temporal Hybrid Model (ST-Hybrid Model). This hybrid model integrates the Prophet module and GCN (Graph Convolutional Network), incorporating a reinforcement learning framework to dynamically optimize feature combinations, aiming to achieve globally optimal predictions of PM2.5 concentrations.

2. Data Processing and Research Ideas

2.1. Data Selection

For PM2.5 data selection, since monitoring results of provinces and cities in China are affected by local conditions such as the number of monitoring points and geographical environment (e.g., altitude), this study adopts global PM2.5 data converted from satellite remote sensing images by Washington University in St. Louis using corresponding algorithms. We extracted PM2.5 concentration data (per 10,000 people) corresponding to the geographical spatial population density of provinces and cities in China from 1998 to 2023. The main reason for selecting this data is that the results calculated using the same algorithm based on images captured by the same satellite at the same location ensure subsequent calculations under consistent errors [6].

For automobile data, we use the motor vehicle ownership data publicly available on the open data platforms of the National Bureau of Statistics and the Traffic Management Bureau of the Ministry of Public Security of provinces and cities, and standardize all data units for calculation. For unpublicized datasets, we use the total annual automobile sales volume of provinces and cities from 2010 to 2023 obtained from the China Automobile Industry Yearbook. The results differ from the data publicly available by the Traffic Management Bureau of the Ministry of Public Security of China by no more than one unit, with the main error coming from used car transactions in various provinces and cities.

2.2. Research Ideas

This study infers whether new energy vehicles have a certain correlation with per capita PM2.5 exposure through two parts: correlation analysis between different data combinations and per capita PM2.5 exposure concentration, and correlation analysis with data of all motor vehicles (new energy vehicles and traditional motor vehicles). The objectivity of the results is confirmed through hypothesis testing. Subsequently, multiple linear regression is used to determine whether motor vehicles and new energy vehicles have the same influencing factors. Finally, based on the current growth rate, we estimate how long it will take for new energy vehicles to significantly reduce per capita PM2.5 exposure [7-8].

2.3. Data Preprocessing

First, a preliminary analysis of the existing PM2.5 exposure data is conducted. The key turning trends are analyzed through curve fitting using per capita PM2.5 exposure combined with year data. It is not difficult to find from the graph that the PM2.5 exposure of 34 provincial administrative regions in China reached a peak around 2013, then decreased rapidly and stabilized. However, before 2013, the PM2.5 index of most provincial administrative regions showed an upward trend: some increased rapidly in a linear trend, while others increased slowly without being significantly affected. Therefore, it can be concluded that 2013 was the peak of PM2.5 concentration exposure in China. Upon closer inspection, although the latter half of the curve showed a downward trend after 2015, the trend changed after 2020—some regions showed an accelerated decline, while others showed an upward trend. Thus, it can be inferred that the overall air environment improved significantly after the state focused on air pollution control in 2013, but it is still affected by certain factors after stabilization [9].

This study first conducts correlation analysis on the existing data and selects per capita PM2.5 exposure concentration calculated based on population density, which can effectively avoid errors caused by differences in geographical area and per capita living space among provinces and cities. Preliminary Correlation Calculation Between PM2.5 and Motor Vehicles are shown in table 1.

Table 1. Preliminary Correlation Calculation Between PM2.5 and Motor Vehicles

Region	Correlation	Region	Correlation	Region	Correlation	Region	Correlation
Tianjin	0.319	Heilongjiang	0.140	Shandong	0.718	Zhejiang	0.814
Beijing	0.303	Jilin	0.160	Ningxia	0.738	Hubei	0.820
Shanghai	0.542	Liaoning	0.552	Shaanxi	0.750	Hainan	0.821
Inner Mongolia	0.564	Gansu	0.638	Anhui	0.803	Shanxi	0.823
Jiangsu	0.614	Hebei	0.702	Hunan	0.809	Henan	0.824
Guangxi	0.836	Guizhou	0.837	Yunnan	0.850	Chongqing	0.871
Guangdong	0.876	Jiangxi	0.881	Sichuan	0.888	Fujian	0.891

Analysis shows that except for five provinces and cities, the p-values of other regions are less than 0.05, indicating reliable correlation. Most cities have a correlation greater than 0.8, so it can be judged that motor vehicle ownership and PM2.5 pollution are strongly correlated. Subsequently, correlation analysis is conducted between new energy vehicle ownership data obtained from the Statistics Bureau and Traffic Management Bureau Data Center of each city and PM2.5 pollution level, and the following table 2 is obtained:

Table 2. Correlation Between PM2.5 and New Energy Vehicles

Region	Correlation	Region	Correlation	Region	Correlation	Region	Correlation
Tianjin	0.664	Heilongjiang	0.133	Shandong	0.904	Zhejiang	0.881
Beijing	0.836	Jilin	0.359	Ningxia	0.200	Hubei	0.689
Shanghai	0.665	Liaoning	0.586	Shaanxi	0.765	Hainan	0.630
Inner Mongolia	0.260	Gansu	0.186	Anhui	0.666	Shanxi	0.898
Jiangsu	0.715	Hebei	0.726	Hunan	0.305	Henan	0.766
Guangxi	0.842	Guizhou	0.642	Yunnan	0.850	Chongqing	0.421
Guangdong	0.842	Jiangxi	0.686	Sichuan	0.421	Fujian	0.561

From the new energy vehicle ownership data, it is found that the ownership of new energy vehicles in some provinces and cities was less than 2,000 before 2018. Excessively small data will affect the dataset, leading to abnormally high p-values. However, from samples with small p-values, it can be found that compared with traditional motor vehicle ownership, new energy vehicle ownership has a greater impact and stronger correlation on PM2.5 concentration [10].

3. Model Analysis and Establishment

3.1. Model Selection

This study first adopts a panel data structure for time series analysis through autoregressive models, then uses the machine learning Random Forest algorithm for analysis to pursue better results.

3.2. Model Establishment

In this study, the Seasonal Autoregressive Integrated Moving Average with Explanatory Variables (SARIMAX) model is used for prediction and analysis. Its mathematical expression is:

$$(1-\Phi_1 L^S-\Phi_2 L^{2S}-\cdots-\Phi_P L^{PS})(1-L^d)(1-L^S)^D y_t = c+\sum_{i=1}^{p}\phi_i L^i y_t+\sum_{j=1}^{q}\theta_j L^j \varepsilon_t + \sum_{k=1}^{Q}\Theta_k L^{kS}\varepsilon_t+\sum_{l=1}^{m}\beta_l X_{t-l} \quad (1)$$

Where: (p, d, q) are non-seasonal orders (set to 1, 1, 1 in this study); (P, D, Q, S) are seasonal orders ($S = 12$ for monthly data); L is the lag operator; X_t is the exogenous variable (motor vehicle ownership); ε_t is the white noise error term

In calculation, the algorithm automatically calculates and compares autocorrelation and partial autocorrelation through the autocorrelation function to determine the basic order, and selects the optimal model using the Akaike Information Criterion (AIC). During prediction, we adopt the rolling time window prediction method to fit the model through maximum likelihood estimation and generate predicted values using three-step-ahead prediction. We then evaluate the model's prediction performance, calculate the 95% confidence interval, then compute the spatial coverage rate, and calculate the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for point prediction accuracy:

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

Using the confidence interval, we can also obtain the Confidence Interval Coverage Probability (CICP):

$$\mathrm{CICP} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \in [\hat{y}_{i,lower}, \hat{y}_{i,upper}]) \tag{3}$$

Where $I(\cdot)$ is the indicator function (1 if the condition is satisfied, 0 otherwise).

First, we use the time series analysis model to predict PM2.5. We can choose multiple models, as shown in the following table 3:

Table 3. Types of Time Series Analysis

Model Type	ACF Performance	PACF Performance
AR(p)	Tail-off	Cut-off after p orders
MA(q)	Cut-off after q orders	Tail-off
ARMA(p, q)	Tail-off	Tail-off

From the plots, it is found that the autocorrelation model shows a tail-off phenomenon, and the partial autocorrelation model shows a cut-off. Meanwhile, not all data fall within the confidence interval, so the stationarity test condition is not satisfied, and the ARIMA model needs to be used instead. In the code, we judge by calculating the ADF statistic and p-value. The calculation process of the ADF statistic is similar to the above stationarity condition. We need to determine whether there is a unit root: if the series is stationary, there is no unit root; otherwise, there is a unit root. If the obtained significance test statistic is less than the three confidence levels (10%, 5%, 1%), we can reject the null hypothesis with (90%, 95%, 99%) confidence, indicating a stationary series. Taking Beijing's data as an example: ADF statistic = -2.542, p-value = 0.106, indicating that Beijing's PM2.5 time series has obvious trends or seasonality and requires differencing. The ARIMA model results after first-order differencing, the three parameter test results are as table 4:

Table 4. Parameter Test Results of Time Series Analysis

Type	Value	Standard Error	z	p-value	Significance
const	-0.092	2.128	-0.043	0.966	Not significant
ma.L1	0.007	0.338	0.022	0.983	Not significant
sigma2	17.734	5.868	3.022	0.003	Significant

3.3. Model Validation

Table 5. Model Validation Results

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	2.07
Prob(Q):	1.00	Prob(JB):	0.35
Heteroskedasticity (H):	0.92	Skew:	0.70
Prob(H) (two-sided):	0.91	Kurtosis:	3.08

Model Validation Results are shown in table 5. Ljung-Box test ($Q = 0.00,\ p = 1.00$): Residuals have no autocorrelation, passing the test; Jarque-Bera test ($p = 0.35$): Residuals follow a normal distribution, passing the test; Heteroskedasticity test ($H = 0.92$, $p = 0.91$): Residuals have homogeneous variance, passing the test.

The model lacks explanatory power in the moving average term and the constant term is not significant due to insufficient available data. This problem arises due to structural breaks in the data with non-linear trends (such as policy interventions and the outbreak of the epidemic). First-order differencing may be insufficient, and higher-order differencing is required for subsequent verification.

Therefore, we add motor vehicle ownership as a covariate for calculation. To improve calculation efficiency, we introduce a panel data model to uniformly calculate 31 provincial administrative regions in China with unified environmental policies.

After the above process, the evaluation indicators of the model are obtained as follows: Mean Absolute Error (MAE) = 2.73, Root Mean Square Error (RMSE) = 3.77, Prediction Coverage Rate (95% CI) = 94.6%. Among them, the cities with the best prediction effects are Tibet (prediction error = 0.4361), Hainan (prediction error = 0.8622), and Fujian (prediction error = 0.9058). Although the results seem good, the worst prediction results are also prominent: Tianjin has an error as high as 11.597, Jilin has an error of 6.156, and Beijing has an error of 5.2020.

4. Error Analysis

Geographical factors, as well as relevant political and economic reasons, were not considered when initially establishing the model.

For example: Tianjin is located on the west bank of the Bohai Bay, with significant land-sea breezes. Northwest winds prevail in winter, which will carry inland pollution into Tianjin. Meanwhile, moist marine air currents will inhibit diffusion, superimposed with logistics emissions from the port. Therefore, the pollution index of Beijing decreases while Tianjin shows abnormally high pollution.

Jilin Province is located in the hinterland of the plain with a cold temperate continental monsoon climate. The snow cover period inhibits dust, while the excessively long coal-fired heating period and seasonal outbreaks of crop straw burning lead to abnormalities.

Corresponding policies also differ: Beijing has normalized alternate-day traffic restrictions based on odd and even license plate numbers, Tianjin implements tail number restrictions, while Jilin only has traffic restrictions in Changchun (the provincial capital) during certain periods. The proportion of new energy vehicles also varies. Taking 2023 as an example, new energy vehicles account for 32.7% in Beijing, 18.9% in Tianjin, and 9.3% in Jilin Province.

In this regard, we use Shapley value decomposition to analyze the proportion of main error sources in the three regions (table 6):

Table 6. Proportion of Error Sources

Error Source	Beijing	Tianjin	Jilin
Unincorporated meteorological factors	38.2%	22.7%	15.4%
Unquantified policy implementation intensity	29.1%	51.3%	34.6%
Unmodeled inter-regional transmission	42.5%	36.8%	18.2%
Sudden pollution incidents	9.7%	18.4%	41.9%

5. Establishment of Prediction Model

Therefore, we adopt machine learning algorithms. Traditional time series analysis cannot resolve the threshold effect between PM2.5 and motor vehicle ownership (e.g., marginal pollution increases when ownership exceeds 2 million vehicles per province, and most provinces and cities have ownership exceeding 2 million vehicles). Similar shortcomings include the excessively fast attenuation rate of (partial) autocorrelation in ARIMA models, which makes it difficult to resolve inter-annual pollution accumulation effects. Since our prediction time span is large, traditional models are limited by the dimension of exogenous variables and cannot be applied to multi-dimensional data.

To verify accuracy, we split the dataset again with 2021 as the boundary: 1998-2021 as the training set and 2022-2023 as the test set. While retaining the original time trend characteristics, we add lag features to ensure time continuity, use OneHotEncoder to process provincial categorical variables, and verify accuracy. Meanwhile, we introduce two new time features: quadratic term features and leading trend variables. For the time series of each province, lag observations are generated using the moving window method:

$$\mathrm{PM2.5}_t = f(\mathrm{PM2.5}_{t-1}, \mathrm{PM2.5}_{t-2}) \tag{4}$$

One-Hot encoding is used to vectorize the "Province" feature, generating 31-dimensional provincial dummy variables to effectively characterize the impact of regional spatial differences on PM2.5 concentration. For the machine learning model, we select the Random Forest Regression model for modeling, which has the following advantages: it is conducive to processing mixed features (e.g., continuous + categorical); it automatically captures non-linear relationships; it provides interpretability through feature importance evaluation.

In the prediction stage, we introduce the coefficient of determination (R^2) combined with MAE and RMSE to predict the performance of the machine learning model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{5}$$

After calculation, the model prediction results are obtained as table 7:

Table 7. Model Evaluation Results

Model Evaluation Indicator	Result
Mean Absolute Error (MAE)	1.97 μg/m³
Root Mean Square Error (RMSE)	2.67 μg/m³
Coefficient of Determination (R^2)	0.90

The indicators show that the model can explain 90% of the variation in PM2.5 concentration, and the prediction error is controlled within the allowable range of environmental monitoring instruments ($\pm 5\mu g/m^3$).

Subsequently, feature importance analysis is conducted through Gini importance evaluation, and it is found that: the contribution of first-order lag features reaches 62.7%; provincial features collectively explain 28.1% of the variation; time trend terms contribute 9.2%.

To address the limitations of traditional time series models and machine learning models, a Spatio-Temporal Hybrid Model (ST-Hybrid Model) is proposed, The innovations of the model are as follows:

The Prophet module is used to capture long-term trends, seasonality, and policy interventions:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{6}$$

Where $g(t)$ is the growth term, $s(t)$ is the seasonal term, and $h(t)$ is the policy dummy variable.

The GCN module is used to construct a provincial spatial adjacency matrix, and the edge weight is defined as:

$$\omega_{ij} = \frac{1}{d_{ij}^2} \cdot \cos(\theta_i - \theta_j) \tag{7}$$

Where d_{ij} is the distance between provinces, and θ is the prevailing wind direction angle.

Meanwhile, a reinforcement learning framework is introduced to optimize feature combination, including current feature importance ranking, model error indicators, and computing resource occupancy rate. Lag terms are added, collinear features are eliminated, external data is introduced, and a reward function is set (with coefficients of 0.6, 0.3, 0.1 respectively):

$$R = \alpha \cdot \Delta R^2 - \beta \cdot \Delta RMSE - \gamma \cdot T \tag{8}$$

Where $\alpha = 0.6$, $\beta = 0.3$, $\gamma = 0.1$, ΔR^2 is the change in the coefficient of determination, $\Delta RMSE$ is the change in RMSE, and T is the computing time.

Quantile Random Forest is still used to predict the confidence interval of PM2.5:

$$\hat{y}_q(x) = F^{-1}(q|x) \tag{9}$$

Where $F^{-1}(\cdot)$ is the inverse cumulative distribution function, and q is the quantile (set to 0.025 and 0.975 for 95% confidence interval).

Subsequently, calculations are performed, and the results of the new model are obtained as table 8:

Table 8. Prediction Model Evaluation Table

City	Root Mean Square Error	Goodness of Fit (R^2)	City	Root Mean Square Error	Goodness of Fit (R^2)
Beijing	2.54	0.98	Hubei	0.85	0.99
Tianjin	2.44	0.96	Hunan	1.12	0.98
Hebei	1.49	0.97	Guangdong	0.97	0.97
Shanxi	1.02	0.98	Guangxi	1.06	0.97
Inner Mongolia	0.74	0.89	Hainan	0.61	0.96
Liaoning	1.49	0.91	Chongqing	0.71	0.99
Jilin	1.72	0.98	Sichuan	0.45	0.99
Heilongjiang	1.57	0.84	Guizhou	0.83	0.98
Shanghai	1.81	0.93	Yunnan	0.45	0.97
Jiangsu	1.59	0.95	Tibet	0.22	0.91
Zhejiang	1.07	0.97	Shaanxi	0.87	0.97
Anhui	1.22	0.98	Gansu	0.71	0.94
Fujian	0.77	0.97	Qinghai	0.43	0.88
Jiangxi	1.04	0.98	Ningxia	1.13	0.92
Shandong	1.31	0.98	Xinjiang	1.29	0.87
Henan	1.35	0.98	Average	1.12	0.95

The goodness of fit is close to 1 and the root mean square error is close to 0, indicating that the model has good performance and is superior to the above models.

6. Conclusions

This study transitions from data processing and traditional model construction and diagnosis to an innovative hybrid model based on machine learning and deep learning, successfully establishing a high-accuracy, high-interpretability PM2.5 prediction framework. In establishing a preliminary PM2.5 prediction model based on time series and covariates and diagnosing its limitations, the initial ARIMA model failed to effectively address prediction biases caused by nonlinear structural breaks. However, through comprehensive error analysis (SHAP value decomposition), unmodeled factors such as regional transport, meteorological factors, and policy enforcement intensity were clearly identified as primary sources of error. Subsequently, by designing a high-precision hybrid PM2.5 prediction architecture integrating spatiotemporal features with reinforcement learning, this study introduced the Spatio-Temporal Hybrid Model (ST-Hybrid Model). This model effectively combines the long-term trend capture capability of time series (Prophet) with the spatial spillover effect modeling capability (GCN), dynamically optimizing feature selection through a reinforcement learning framework to significantly enhance model performance. Ultimately, this hybrid model demonstrated exceptional fitting capability with a goodness-of-fit index of 0.95, providing a high-precision quantitative tool and methodological reference for motor vehicle pollution control and multi-scale environmental modeling.

References

[1] Xie Zhen, Chen Zhiyong. Spatiotemporal Characteristics and Influencing Factors of PM2.5 in China from 2000 to 2020 [J]. Journal of Fujian Normal University (Natural Science Edition), 2025, 41(02): 16-23.

[2] Zhou Huixia, Cui Baorong, Pan Ying, et al. Analysis of Polycyclic Aromatic Hydrocarbon Pollution Characteristics and Sources in Atmospheric PM2.5 in Fengtai District, Beijing [J]. Journal of Environmental Health, 2024, 14(12): 987-994.

[3] Zhang Lei, Zhang Leilei. Policy Recommendations for New Energy Vehicle Manufacturing Under the “Dual Carbon” Goals [J]. Straits Science & Industry, 2023, 36(05): 76-78+92.

[4] Xue Yuting. Impact of Urban Spatial Structure on Air Pollution [D]. Shandong Normal University, 2024.

[5] Wu Pengfei. How Should Enterprises Respond to the Current “Carbon Trading”? [C]//Industrial Energy Conservation and Clean Production, February 2015, Issue 1 (Total Issue 19), 2015:35.

[6] Ren Donghong, Xing Bingsuo, Tian Yong, et al. A BP Neural Network-Based PM2.5 Prediction Model [J]. China Science and Technology Information, 2025, (08): 70-72.

[7] Li Xinyang, Liu Juan. Research on Monthly-Scale PM2.5 Concentration Prediction Based on Mathematical Models [J]. China Environmental Monitoring, 2025, 41(01): 180-190.

[8] Liu Tingting. Research on PM2.5 Concentration Prediction in Atmospheric Environment [J]. Journal of Environmental Science, 2025, 44(01):58-62.

[9] Jin Lishan, Liu Fang. Machine Learning-Based Prediction and Analysis of PM2.5 and PM10 Concentrations in Hohhot [J].Environmental Science and Technology, 2024, 37(06): 45-50.

[10] Wang Shengjie, Zhang Qinghong, Sang Mingjian. Prediction and Application of PM2.5 Concentrations in Arid Urban Areas Based on an Improved Deep Learning Model [J/OL]. Arid Zone Geography, 1-17 [2025-10-29].