International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 17, No. 2, 2024

Time Series Analysis for Predicting PM2.5 Concentration in Shanghai based on Machine Learning

Xuede Huang, Lingling Liu, Zhonglin Lin, Lingling Chen

MINNAN Univerity of science and technology, Quanzhou, China

Abstract: This study aims to use three different machine learning models -- ARIMA, LSTM and Random Forest -- to predict PM2.5 concentration in Shanghai. By analyzing and comparing the performance of these models in practical applications, we find that each model has its own unique advantages and limitations. The LSTM model performs best in processing complex multivariate time series data, showing high accuracy and excellent data processing capabilities, and its RMSE and MSE indicators are superior to other models. The random forest model excelled in feature importance analysis, providing valuable insights into understanding the drivers of changes in PM2.5 concentration.

Keywords: Time Series Analysis; PM2.5 Prediction; Machine Learning; Model Comparison.

1. Introduction

In the context of global environmental change, urban air quality monitoring has become an important topic. Especially in China's big cities, the high concentration of PM2.5 has seriously affected public health and social and economic development. PM2.5 refers to particulate matter with a diameter of 2.5 microns or less. Due to its small size and large area, it is easy to carry harmful substances, stay in the air for a long time, and has a strong ability to penetrate the respiratory system, posing a particularly serious threat to human health. Therefore, accurate prediction of PM2.5 concentration is of great significance for public health early warning system and environmental policy formulation. In recent years, machine learning technology has been widely applied in the field of environmental science, especially in the analysis of time series data. This study aims to apply and compare multiple machine learning models to predict the change of PM2.5 concentration in Shanghai. By learning from historical data, we hope to build a reliable prediction model that can provide scientific support and decision-making basis for environmental management and policy making.

2. Data Processing

2.1. Data Description

In this study, the data set used mainly includes three parts: PM2.5 concentration data, meteorological data and traffic flow data. The PM2.5 data came from multiple environmental monitoring stations in Shanghai, covering continuous measurement records over the past five years. The data is provided in hourly concentrations, which can reflect short-term trends and long-term environmental conditions. Meteorological data include temperature, humidity, wind speed and direction, all of which can affect changes in PM2.5 concentrations. Traffic flow data, obtained through the city's traffic monitoring system, is crucial to enhancing the model's prediction accuracy, given that vehicle emissions are an important source of PM2.5. All data underwent strict pre-processing before use, including filling in missing values, eliminating outliers and standardizing data, to ensure the accuracy of the analysis and the effectiveness of the model training.

2.2. Data Sources

The data used in this study mainly come from three sources. Both PM2.5 concentration data and meteorological data were collected from the official website of the Shanghai Environmental Monitoring Center, and these data were regularly uploaded by monitoring stations in the city to ensure real-time and accuracy of the data. The traffic flow data comes from the Shanghai Municipal Transportation Commission, which is collected through the city's road traffic monitoring system and reflects the vehicle flow in different time periods. Some of the data are also obtained through open data platforms such as the China Meteorological Data Network and the National Urban Traffic Operation Monitoring Platform, which provide open data access interfaces for researchers to obtain and use relevant data. All data sets are strictly reviewed and pre-processed after acquisition to meet research requirements and data analysis standards, ensuring the stability and scientificity of the research data base.

2.3. Data Collection and Preprocessing

For the data used in this study, the collection and preprocessing phases are critical to ensure the validity and reliability of the final analysis. The data collection phase involves taking data synchronously from multiple sources and integrating it into a unified data framework.

Collection phase: PM2.5 and meteorological data, automatically pulling data from the API of the Environmental Protection Bureau and the Meteorological Bureau on a regular basis and recording it every hour. Traffic flow data, obtained through the data interface of the urban Transportation Authority, usually in 15 minutes or 30 minutes as one data point.

The pre-processing phase begins by identifying missing values in all datasets. For a small range of missing values, the means of adjacent time points are used to fill in. For long periods of time, consider filling in gaps using linear interpolation or based on data from similar days. Use box plots and standard deviation methods to identify outliers. For data points marked as outliers, choose to delete or replace them with the mean of the surrounding data, depending on the situation. Due to the need of model training, the data is

normalized or standardized to eliminate the impact of different dimensions. Z-score normalization or Min-Max normalization methods are usually used. The PM2.5 concentration data, meteorological data and traffic data are aligned according to time stamps to form a single time series data set, which is convenient for subsequent machine learning processing.

Through these meticulous pre-processing steps, the data quality can be ensured and the accuracy and efficiency of subsequent machine learning model prediction can be improved.

Table 1. Sample data table

timestamp	PM2.5 (μg/m³)	Temperature (°C)	Humidity (%)	Wind speed (m/s)	Wind direction (degrees)	Traffic flow (car/hour)
2023-04-01 00:00	35	18.2	60	3.2	180	750
2023-04-01 01:00	38	18.0	62	3.0	175	732
2023-04-01 02:00	36	17.5	63	2.8	170	710
2023-04-01 03:00	34	17.0	65	2.5	165	690
2023-04-01 04:00	33	16.8	67	2.3	160	670

3. Application of Several Machine Learning Models

3.1. Autoregressive Integral Moving Average Model (ARIMA)

ARIMA model, or ARIMA, is a widely used time series prediction method, which is especially suitable for analyzing and predicting univariate time series data. When applying ARIMA model to predict PM2.5 concentration in Shanghai, the operation steps can be specifically divided into the following stages:

(1) Stationarity test of data: First, stationarity test of time series data is required, usually using unit root test (such as ADF test). If the data is not stationary, differential processing is needed. First-order difference is usually sufficient to bring the PM2.5 concentration data to a stationary state, but multi-order difference can also be performed if necessary.

(2) Determination of model parameters: Determining ARIMA model parameters (p, d, q) is a key step. Where p is the order of the autoregressive term, d is the difference number, and q is the order of the moving average term. These parameters can be estimated by looking at autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs. For example, the number of lags that are significantly non-zero in the PACF plot can be taken as the order of the AR term, and the number of lags that are significantly non-zero in the ACF plot can be taken as the order of the MA term.

(3) Model fitting: The ARIMA model is fitted using the selected parameters. In this step, the corresponding functions in statistical software can be utilized.

(4) Model diagnosis: After fitting is complete, a diagnostic check on the model is needed to ensure that the model does not violate any statistical assumptions. Checking whether the residuals are white noise can be done through the ACF plot of the residuals to ensure that there are no autocorrelations.

(5) Prediction and evaluation: Use a well-fitted model to make predictions of future values. In addition, it is very important to evaluate the accuracy of the prediction of the model, and the commonly used evaluation indicators include mean square error (MSE) and root mean square error (RMSE). The prediction performance directly affects the effectiveness of the model in practical application.

3.2. Long Short-term Memory Network Model (LSTM)

Long short-term memory network (LSTM) is a special type of recurrent neural network (RNN) that is well suited for processing and predicting time series data with long-term time dependence. Through its unique gating mechanism, LSTM solves the problem of gradient disappearance, which is often encountered in the processing of long series data by traditional RNNS, making it excellent in complex sequence prediction tasks. In the task of predicting PM2.5 concentration in Shanghai, which has multi-variable inputs and time series complexity, LSTM is able to effectively capture long-term dependencies and nonlinear patterns in environmental data.

(1) Data preprocessing: Before applying LSTM, data needs to be preprocessed first. This includes scaling (such as normalizing or standardizing) input variables to prevent gradient problems during model training. In addition, given that the input to the LSTM model requires a specific format, it is common to convert the data into the form of a sliding window (for example, using data from the last 24 hours to predict the PM2.5 concentration in the next hour).

(2) Model architecture: A typical LSTM model consists of one or more LSTM layers, followed by one or more dense connecting layers, and finally the output layer. For PM2.5 concentration prediction, the input layer of the model receives 3D data in the form of (sample number, time step, feature number). The LSTM layer can process these input data and extract the features of the time series. After the LSTM layer, a fully connected layer is usually added to integrate the learned features and make the final prediction.

(3) Training strategy: The training of the LSTM model includes selecting the appropriate loss function and optimizer. For regression problems such as PM2.5 concentration prediction, the commonly used loss function is mean square error (MSE). Optimizers such as Adam or RMSprop are often used to train deep learning models due to their adaptive learning rate properties. In addition, to prevent overfitting, techniques such as early stopping or Dropout can be employed.

(4) Performance evaluation: The performance of the model can be evaluated by its performance on the verification set. Common evaluation indexes include root mean square error (RMSE) and coefficient of determination (R^2). By comparing the performance under different model configurations, the best model can be selected for practical prediction.

Once the model is fully trained and verified, it can be deployed for real-time or regular PM2.5 concentration predictions. By continuously monitoring the difference between the model output and the actual concentration, the model parameters can be constantly adjusted to maintain the prediction accuracy. Through such a process, the LSTM

model provides a powerful tool for accurate prediction of PM2.5 concentration in Shanghai, helping relevant departments to better understand the changing trend of air quality and formulate corresponding countermeasures. The application of LSTM also demonstrates the potential and value of deep learning technology in the field of environmental science.

3.3. Random Forest Model

Random forest modeling is a powerful ensemble learning technique that improves the accuracy of predictions and the stability of the model by building multiple decision trees and aggregating their predictions. This method performs well when dealing with complex nonlinear problems and is well suited to applications in the field of environmental science, such as predicting PM2.5 concentration in Shanghai.

(1) Data preparation and feature selection: The success of the random forest model largely depends on the selection of input features. When preparing the data, not only the historical PM2.5 concentration, but also other environmental factors that may affect air quality should be considered, such as temperature, humidity, wind speed, wind direction and other pollutant indicators. All of these variables can be used as features for model inputs. Data usually requires a certain amount of pre-processing, including missing value processing, outlier detection, and data normalization, to ensure the effectiveness of model training.

(2) Model construction: When building a random forest model, it is first necessary to determine the number of trees (n_estimators), which is a key parameter affecting the model performance. Too few trees may lead to insufficient stability of the model, while too many trees may increase the complexity of the calculation but not necessarily improve the performance. In addition, other parameters need to be set, such as the maximum depth of the tree (max_depth), the minimum number of samples required to split a node (min_samples_split), etc., which will affect the fit degree and generalization ability of the model.

(3) Training strategy: The training of random forest includes the training of each decision tree using randomly sampled data and features. The advantage of this "bagging" technique is that it can reduce the variance of the model and increase the ability to generalize to new data. Typically, cross-validation is used to evaluate the performance of the model at different parameter Settings and to select the optimal parameter configuration.

(4) Performance evaluation: The performance of the model is evaluated by the prediction results on the test set. Commonly used evaluation indicators include mean square error (MSE), root mean square error (RMSE) and coefficient of determination (R^2). The accuracy and reliability of model predictions can be quantified by these indicators.

The main advantages of random forest models are their excellent accuracy, ability to handle large amounts of data, and natural handling of various types of variables. Once the model is trained, it can be used to make real-time or regular PM2.5 concentration predictions, providing a scientific basis for urban air quality management and policy making. In addition, random forests are also well explanatory, and the impact of each feature on the predicted results can be assessed by the feature importance score, which is very valuable for understanding the drivers of PM2.5 concentration changes. By applying the random forest model, the researchers can provide a tool that is both stable and reliable to help environmental scientists and policy makers better understand and predict changes in air quality, so that more effective countermeasures can be taken.

4. Results

4.1. Model Performance Evaluation

In order to comprehensively evaluate the performance of each model, root-mean-square error, mean-square error, coefficient of determination indexes were used. Root mean square error (RMSE) is a standard measure of the difference between the predicted value and the actual value, with lower values indicating more accurate predictions. Mean square error (MSE) is similar to RMSE, but it provides an average of the squared errors, giving greater weight to large errors. The coefficient of determination (R^2) measures how well the model explains changes in the data, with a value closer to 1 indicating the model's explanatory power.

Table 2. Evaluation results

Type of model	RMSE (μg/m³)	MSE (μg²/m⁶)	R^2
ARIMA	22.5	506.25	0.75
LSTM	15.3	234.09	0.85
Random Forest	18.7	349.69	0.80

4.2. Analysis of Results

Long Short Term Memory network (LSTM) models show high performance due to their ability to handle long term dependencies on time series data, especially when dealing with multivariate inputs related to weather and traffic data. The RMSE and MSE values of the LSTM model were 15.3 μg/m³ and 234.09 μg²/m⁶, respectively, while its R^2 value reached 0.85, reflecting its excellent forecasting accuracy and high data interpretation capability. Random Forest model, as a powerful ensemble learning algorithm, enhances the accuracy and stability of the model by constructing multiple decision trees and synthesizing their prediction results. In this study, the Stochastic Forest model performed slightly worse than LSTM, but was able to provide important insights into the impact of data features, which can be valuable for identifying and understanding key factors affecting PM2.5 concentrations. The model had an RMSE of 18.7 μg/m³, an MSE of 349.69 μg²/m⁶, and an R^2 value of 0.80. In contrast, the ARIMA model provides a better baseline forecast and is particularly suitable for processing time series data with clear trends and seasonality. While ARIMA is effective in univariate forecasting, it is less effective in multivariate forecasting. Its performance indicators are RMSE 22.5 μg/m³, MSE 506.25 μg²/m⁶, and R^2 0.75.

In a comprehensive evaluation of the performance of the three models -- ARIMA, LSTM and Random Forest -- in predicting PM2.5 concentration in Shanghai, we find that each model has its unique advantages and limitations. The LSTM model performs best on all evaluation indicators, especially in processing multivariate time series data containing complex dependencies. With its accurate prediction results and high coefficient of determination, LSTM provides a very effective tool for accurately monitoring and predicting air quality. The Random forest model, while slightly inferior to the LSTM in terms of accuracy, excelled in feature importance analysis, making it particularly valuable in explaining drivers of changes in PM2.5 concentration. And the ARIMA model, although it performs well in processing univariate time series data,

especially data with obvious seasonality and trend, is not as flexible and adaptable in multivariate environments as the other two models. Therefore, the selection of the right model needs to be determined according to the specific data characteristics and forecasting needs. Future research may explore the method of model integration to integrate the advantages of each model and further improve the accuracy and practicability of prediction.

References

[1] Pan Zhengtong. Predictive Analysis of China's cargo traffic volume based on Machine Learning time series [J]. China Storage and Transportation,2022,(01):85-86.]

[2] Jiang Yuan, Yang Bo, Zhao Donglai, et al. Mobile network traffic prediction technology based on time series analysis and machine learning [J]. Internet of Things Technology, 2019, 10(06):42-45.

[3] FU Zhiou, Zhou Yang, Chen Cheng, et al. [3] Fu Z, Zhou Y, Chen C, et al. Application of time series analysis and machine learning method in predicting the incidence trend of pulmonary tuberculosis [J]. China Health Statistics, 2019,37(02):190-195.