International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 6 Issue 5, July-August 2022 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Atmospheric Pollutant Concentration Prediction Based on KPCA-BP

Xin Lin, Bo Wang, Wenjing Ai

School of Information, Beijing Wuzi University, Beijing, China

ABSTRACT

$PM_{2.5}$ prediction research has important significance for improving human health and atmospheric environmental quality, etc. This paper uses a model combining nuclear principal component analysis method and neural network to study the prediction problem of meteorological pollutant concentration, and compares the experimental results with the prediction results of the original neural network and the principal component analysis neural network. Based on the O_3, CO, PM_{10}, SO_2, NO_2 concentrations and parallel meteorological conditions data of Beijing from 2016 to 2020, the $PM_{2.5}$ concentration was predicted. First, reduce the latitude of the data, and then use the KPCA-BP neural network algorithm for training. The results show that the average absolute error, root mean square error and expected variance score of the combined model are relatively good, the generalization ability is strong, and the extreme value prediction is the best, which is better than that of the single model.

KEYWORDS: *KPCA; prediction of atmospheric pollutants; BP neural network; PCA*

How to cite this paper: Xin Lin | Bo Wang | Wenjing Ai "Atmospheric Pollutant Concentration Prediction Based on KPCA-BP" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-6 | Issue-5, August 2022, pp.2008-2016, URL: www.ijtsrd.com/papers/ijtsrd51746.pdf

IJTSRD51746

1. INTRODUCTION

Beijing air pollution index has remained high in recent days, with $PM_{2.5}$ As the primary pollutant, due to its small particle size, it comes with a large number of toxic and harmful substances, and it is suspended in the air for a long time, which is a greater harm to human health. Looking at the governance of $PM_{2.5}$ in the past five years, the concentration of $PM_{2.5}$ is due to the proposal of relevant policies and the enhancement of people's awareness of environmental protection It has been significantly reduced, but the control of $PM_{2.5}$ and other related pollutants should continue to be strengthened[1]. According to the results of the national air quality forecast consultation published by the Ministry of Ecology and Environment of the People's Republic of China from November 2020 to October 2021[2]. In Beijing-Tianjin-Hebei and some surrounding areas, the air quality between April and October is mainly good to mild pollution, and its primary pollutant is $PM_{2.5}$. Air quality is relatively poor in autumn and winter, so being able to predict $PM_{2.5}$ concentration more accurately is an important issue.

Common prediction methods for $PM_{2.5}$ concentration include: artificial neural networks [3], wavelet-neural networks[4], multiple linear regression models[5], LSTM algorithm [6] et al. In these prediction methods, artificial neural network algorithms are most commonly used for complex nonlinear relationships, such as water resource prediction, traffic route prediction, etc. Due to the characteristics of the artificial neural network algorithm itself, it is a model of distributed parallel processing algorithm, and it is difficult to consider the influence of each factor in the multi-factor problem on the predicted value, so this paper adopts a dimensionality reduction method to analyze the multi-factor problem.

In the method of reducing the dimension, the concept of kernel function (kernel) is introduced in this paper because the dimensionality reduction effect of processing linear data in principal component analysis needs to be improved. To date, neural network algorithms based on nuclear principal component analysis have been used less in related studies on atmospheric pollutant concentration prediction. There are many influencing factors for the concentration of

pollutants in the atmosphere, and a hot issue in the current research is how to effectively extract the relevant information between the main factors, which is of great significance for the subsequent improvement of the accuracy of the $PM_{2.5}$ prediction model.

In this paper, when analyzing the influence of multiple factors on the concentration of $PM_{2.5}$ pollutants in the atmosphere, based on the prediction method of traditional neural network, a variety of dimensionality reduction methods are used to predict it. The experimental results of various prediction methods were analyzed and compared, and a relatively good method was proposed to improve the accuracy of predicting $PM_{2.5}$ concentration values. In this paper, based on the prediction method based on the traditional neural network, the TensorFlow-BP algorithm is used to PM the influencing factors of 5 factors (pollutant concentration) and 12 factors (pollutant concentration and meteorological conditions), respectively Concentration prediction; Principal component analysis and nuclear principal component analysis are mainly used to extract components and input them into tensorFlow-BP neural network model to predict $PM_{2.5}$ concentration. Finally, the prediction results of each model are analyzed and compared by using relevant indicators such as MAE and RMSE.

2. Data collection and evaluation indicators

2.1. data source

According to the big data information released by the National Meteorological Science Data Center, download the monitoring data of the Ecological Environment Monitoring Center Station in Beijing from 2016 to 2020, and the main pollution factors in the atmosphere are O_3 and CO, PM_{10}, $PM_{2.5}$, SO_2, NO_2, etc., and parallel to the daily wind speed, air temperature, surface temperature, sunshine hours, humidity, barometric pressure and cumulative precipitation. This article is selected from January 1, 2016 to December 2019 The $PM_{2.5}$ concentration value for the period on the 31st is used as a training sample set, from January 1, 2020 $PM_{2.5}$ concentration values for this period on December 31, 2012 were used as a test sample set.

2.2. Data processing

In the 2016-2019 data collected, some small amounts of missing data were populated with the mean of the data from the adjacent observatory. The data outliers were then processed using the boxplot and 3σ principles, and the z-score data for each feature in the training set and the test set were normalized, taking into account that the physical meaning and dimensions of the air pollutant concentration data and the meteorological condition data were not the same.

2.3. Evaluation indicators

This paper uses the mean absolute error (MAE), root mean square error (RMSE), and explained variance score as evaluation indicators to compare the degree of difference between the predicted values and the measured values of each model Do not show this in Equations (1)-(3).

$$MAE = \frac{\sum_{i=1}^{n} \left| X_{obs,i} - X_{\mathrm{mod}el,i} \right|}{n} \quad (1)$$

where: $X_{obs,\ i}$ represents the forecast, $X_{model,\ i}$ represents the measured data, and n represents the number of predictions. The smaller the value of the MAE, the better the fit between the predicted data and the real data, so the smaller the indicator, the better.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{obs,i} - X_{\mathrm{mod}el,i} \right)^2}{n}} \quad (2)$$

where: $X_{obs,\ i}$ represents the ith forecast, $X_{model,\ i}$ represents the measured data, and n represents the number of predictions. The smaller the value of RMSE, the smaller the error between the model's prediction data and the real data, so the smaller the indicator, the better.

$$\exp lained\_\mathrm{var} iance\left(X_{obs,i}, X_{\mathrm{mod}el,i}\right) = 1 - \frac{Var\left\{X_{obs,i} - X_{\mathrm{mod}el,i}\right\}}{Var\left\{X_{obs,i}\right\}} \quad (3)$$

where: $X_{obs,\ i}$ represents the ith predicted value, $X_{model,\ i}$ represents the true value, n represents the number of predictions, and it's The range of values is [0,1]. When the Application variance score is closer to 1, it shows that the independent variable can explain the variance change of the dependent variable. The better the support vector regression model is built, so the closer the value of the Deployed variance score is to 1, the better.

3. Experimental model

3.1. BP neural network model

BP Neural Network (BP Neural Network) belongs to the nonlinear dynamic information processing system of backpropagation algorithm, which is one of the most widely used models in meteorological forecasting applications [7], the algorithm does not need to clarify the functional relationship between input and output, and can make predictions about new data by adjusting parameters inside the network[8]。

This paper uses the BP neural network as a predictive model, and its structure mainly includes three layers, which are the input, output and implicit layers of the structure. First, the relevant data is input through the input layer, and then the data is passed to the implicit layer, and after the data is activated and enlarged, it is passed to the output layer and output by the output layer, the model for:

$$Y = g(W^{\mathrm{T}} X + b) \tag{4}$$

During operation, when the actual error is larger than the expected error, the backpropagation of the error will begin, and the corresponding values will be adjusted, that is, the weighting value and the threshold value The network is continued to be trained repeatedly so that the model parameters move in the direction of reduced loss values until the desired error is met, the mapping between the input and output is determined. ABP neural network implemented by the Keras library under the tensorflow framework in deep learning. Figure 3 depicts the computational flow of the relevant network data, and uses this as the basic computing node of the framework, responsible for maintaining and updating the node state.

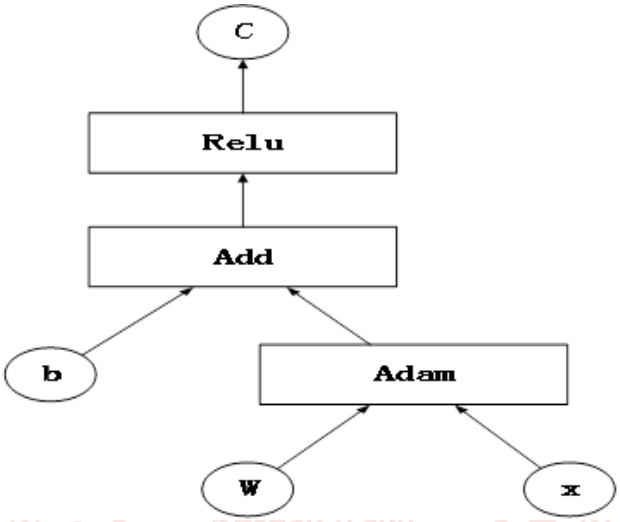

Figure 1 TensorFlow calculation graph

3.2. PCA-BP neural network model

Principal Component Analysis is a multivariate statistical method that transforms multiple variables that originally had a certain correlation into a few unrelated principal components through dimensionality reduction techniques[9]. The relevant process can be roughly divided into the following steps:

1. Standardization processing, the purpose of which is to eliminate variable dimensional relationships;
2. Establish a Pearson coefficient matrix;
3. Calculate the eigenvalues and eigenvectors corresponding to the Pearson coefficient matrix and sort them by size;
4. Calculate the matrix of cumulative contribution rate and principal component score coefficient.

First of all, in order to reduce the multiple correlations between various factors, the method of principal component analysis was used to select new indicators to predict the concentration of $PM_{2.5}$ more accurately; Then, the reduced complexity dataset is combined with the BP neural network to improve the running speed of the neural network algorithm, solve the nonlinear problem between multiple data, reduce the redundancy of the input data, and improve the accuracy of the prediction result.

3.3. KPCA-BP neural network model

The Nuclear Principal Component Analysis (KPCA) method is a nonlinear extension of the Principal Component Analysis (PCA) method, which first introduces a nonlinear mapping function Φ whose purpose is to better process the relevant data, mapping R^N sample vectors in the original space x_k to high-dimensional space F, that is, $R^N \to F, x_k \to \Phi(x_i)$ the relevant data can be converted from linearly indivisible to linearly separable, and the corresponding principal component analysis can be performed in high-dimensional space F.

This article makes the following assumptions: the centralized set of samples is denoted X, $\{x_1, x_2, \cdots, x_N\}$ which is R^N the set of samples in space, where the total number of samples is N, and the dimension of each sample is d, and then passes through in high-dimensional space The mapping in F can be obtained $\Phi(X)$, where

, $\sum_{i=1}^{N}\Phi(x_i)=0$. The corresponding eigenvectors in the high-dimensional space F are denoted $v_i, i=1,2,\cdots,d$, and the corresponding eigenvalues are denoted $\lambda_i,\ \ i=1,2,\cdots,d$, where the eigenvectors can be represented linearly by the collection of samples in the space, that is, there is a set of parameters $\alpha=(\alpha_1,\ \alpha_2,\cdots,\ \alpha_N)$ satisfied:

$$v_i=\sum_{i=1}^{N}\alpha_i\Phi(x_i)=\alpha\Phi(X) \quad (5)$$

The use of PCA in high-dimensional space F, availabl:

$$\Phi(X)\Phi(X)^{\mathrm{T}}v_i=\lambda_i v_i \quad (6)$$

After the operation, the equation 9 is obtained.

$$\Phi(X)^{\mathrm{T}}\Phi(X)\Phi(X)^{\mathrm{T}}\Phi(X)\alpha=\lambda_i\Phi(X)^{\mathrm{T}}\Phi(X)\alpha \quad (7)$$

In Equation (9), both sides of the equation contain it, $\Phi(X)^T\Phi(X)$ and it is replaced with a kernel matrix K. In general, kernel functions satisfy Mercer's theorem, and this article uses Gaussian kernel functions.

$$K(x,x_i)=\exp(-\frac{\|x-x_i\|^2}{2\sigma^2}) \quad (8)$$

where, σ is the argument to the kernel function.

At this point, any test sample vector mapped to a high-dimensional space $\mathrm{x}_{\mathrm{new}}$ has:

$$(\alpha^{\mathrm{k}}\Phi(\mathrm{x}_{\mathrm{new}}))=\sum\alpha_i^k(\Phi(\mathrm{x}_{\mathrm{i}})\cdot\Phi(\mathrm{x}))=\sum\alpha_i^k K(\mathrm{x}_{\mathrm{i}},\ \mathrm{x}) \quad (9)$$

After that, the PCA method of extracting principal components is used **to** calculate the projection of each data point on the corresponding characteristic vector to obtain the nuclear principal components. The KPCA method **is then used**, which is divided into four steps [10].

1. Extract the factors that affect the load, and use the generated variable matrix as the initial input matrix;
2. Select the corresponding kernel function to generate a kernel matrix by transforming and mapping;
3. The eigenvalues and eigenvectors corresponding to the computed kernel matrix;
4. Calculates the numerical value of the cumulative contribution rate to determine the number of input variables of the neural network.

4. Experimental results and analysis

4.1. PM2.5 based on BP neural network model predictions

Establish a BP neural network model under the TensorFlow framework according to Section 2.1, and enter the normalized data into the BP neural network for 5 contaminant factors and one for each 12 meteorological and pollutant factors[11] are trained. In the test set, January 1, 2020 to December 31, 2020 $PM_{2.5}$ concentration is predicted. After debugging, use ReLU to activate the function.

By O_3, CO, PM_{10}, SO_2, NO_2 predict the result of $PM_{2.5}$ concentration in the data of the influencing factors of various pollutants is recorded as ($PM_{2.5}$-5). By O_3, CO, PM_{10}, SO_2, NO_2 These five pollutants, as well as the average wind speed, average temperature, average surface temperature, sunshine hours, average relative humidity, average station pressure, 20-20 hours cumulative precipitation and other meteorological influencing factors data on PM $_{2.5}$ concentration prediction result model is denoted as ($PM_{2.5}$-12)[12].

The mean squared error (MSE) of the two models is smaller than that of the $PM_{2.5}$-12 model, indicating $PM_{2.5}$-12 The model is trained with higher accuracy.

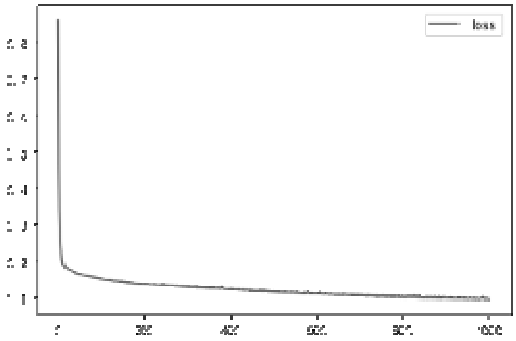

Figure 2 $PM_{2.5}$-5 Loss Function Diagram

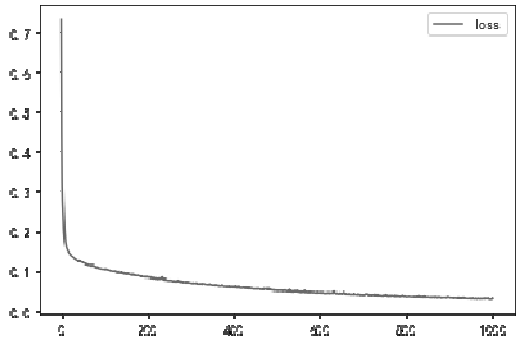

Figure 3 $PM_{2.5}$-12 Loss Function Diagram

Statistics were performed on three evaluation indicators of $PM_{2.5}$-5 model and $PM_{2.5}$-12 model. As can be seen from Table 1, the prediction accuracy of the BP neural network has been greatly improved after increasing the relevant meteorological factors.

Table 1 Comparative analysis of neural network prediction accuracy

	MAE	RMSE	Explained variance score
$PM_{2.5}$-5 models	0. 3963	0. 3595	0. 5636
$PM_{2.5}$-12 models	0. 3188	0. 2025	0. 7414

The prediction of the fitting curve results of the $PM_{2.5}$-5 model and the $PM_{2.5}$-12 model are shown in Figure 4 and Figure 5, respectively, and the fitting effect is good, which is confirmed the generalization and effectiveness of the model proposed in this article are discussed. Through comparative analysis, the $PM_{2.5}$-12 model predicts a relatively better prediction effect.

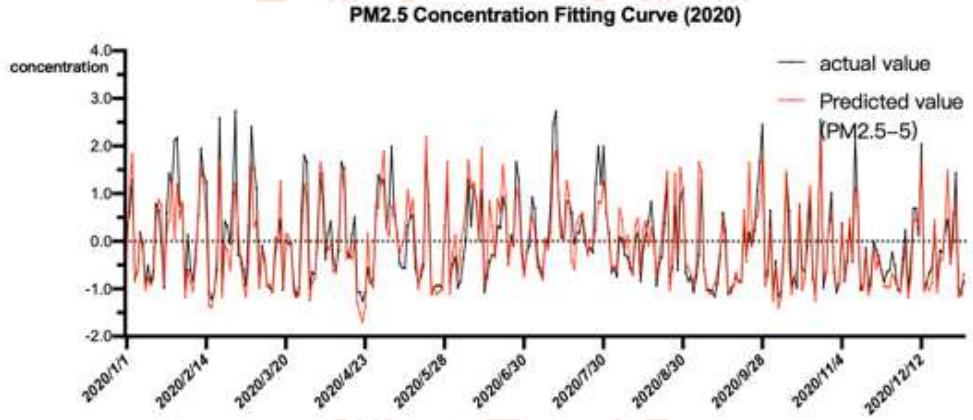

Figure 4 $PM_{2.5}$-5 Model Prediction Results

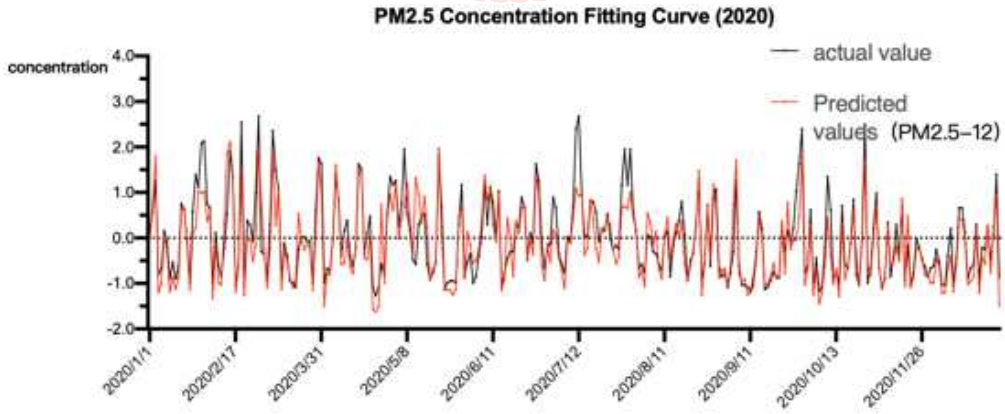

Figure 5 $PM_{2.5}$-12 Model Prediction Results

4.2. PM2 based on PCA-BP neural network model 5 predictions

4.2.1. The eigenvalue is determined with the principal component

In general, principal component analysis methods require that the various factors have a certain correlation with each other. The Pearson correlation coefficient $|r| > 0.35$ is usually required. Programmed using SPSS software, the correlation analysis is performed after standardizing the relevant 12 factors considered, the results of which are shown in Table 2, and two factors are excluded according to the above rules: sunshine hours, Cumulative rainfall and SO2 concentration.

Table 2 Pearson correlation coefficients

	Average wind speed	average temperature	Surface temperature	Hours of sunshine	Average humidity	Average air pressure	Cumulative precipitation	PM_{25}	PM_{10}	SO_2	CO	NO_2
Average wind speed	1	-0.064	-0.039	0.211	-0.444	0.029	-0.021	-0.26	-0.09	-0.005	-0.404	-0.37
Average temperatures	-0.064	1	0.984	0.026	0.414	-0.88	0.121	0.153	0.141	-0.374	-0.169	0.067
Surface temperatures	-0.039	0.984	1	0.098	0.352	-0.864	0.106	0.144	0.159	-0.359	-0.179	0.03
Hours of sunshine	0.211	0.026	0.098	1	-0.573	0.03	-0.071	-0.292	0.009	0.019	-0.038	-0.357
Average humidity	-0.444	0.414	0.352	-0.573	1	-0.4	0.087	0.477	0.05	-0.23	0.072	0.541
Average air pressure	0.029	-0.88	-0.864	0.03	-0.4	1	-0.097	-0.178	-0.17	0.316	0.189	-0.087
Cumulative precipitation	-0.021	0.121	0.106	-0.071	0.087	-0.097	1	0.115	0.077	-0.026	0.074	0.103
PM_{25}	-0.26	0.153	0.144	-0.292	0.477	-0.178	0.115	1	0.718	0.303	0.494	0.812
PM_{10}	-0.094	0.141	0.159	0.009	0.05	-0.171	0.077	0.718	1	0.432	0.583	0.551
SO_2	-0.005	-0.374	-0.359	0.019	-0.23	0.316	-0.026	0.303	0.432	1	0.517	0.394
CO	-0.404	-0.169	-0.179	-0.038	0.072	0.189	0.074	0.494	0.583	0.517	1	0.574
NO_2	-0.37	0.067	0.03	-0.357	0.541	-0.087	0.103	0.812	0.551	0.394	0.574	1

Next, principal component analysis is performed on the remaining 10 factors. The tangency quantity of KMO sampling amount is 0.721, and the Sig value is less than 0.05, so the obtained results are reference and scientific, see Table 3 below[13]。 The rotation process is performed using the Kaiser normalized maximum variance method, so that the factor load values deviate from 0 and 1, and some of the indicator content that has no obvious correlation is removed. The amount of data in this article is large, so it is set up multiple iterations. After 300 iterations, the results will converge. The characteristic values and contribution rates of each principal component are shown in

ingredients	Initial eigenvalue			Extract the sum of squares of the loads		
	total	Percentage of variance	Cumulative /%	total	Percentage of variance	Cumulative /%
1	3.559	35.592	35.592	3.559	35.592	35.592
2	3.227	32.274	67.867	3.227	32.274	67.867
3	1.299	12.986	80.852	1.299	12.986	80.852
4	0.726	7.257	88.109	0.726	7.257	88.109
5	0.451	4.515	92.624			
6	0.289	2.887	95.511			
7	0.168	1.682	97.192			
8	0.155	1.547	98.739			
9	0.113	1.134	99.873			
10	0.013	0.127	100.000			

components with a cumulative contribution rate of > 85% are extracted and determined as the final indicators, and the cumulative contribution rates of the four items are determined 88.109%, basically can reflect the cause sub-information, recorded as F1 ~ F4, for the original indicators of the load status see

ingredients	Initial eigenvalue			Extract the sum of squares of the loads		
	total	Percentage of variance	Cumulative /%	total	Percentage of variance	Cumulative /%
1	3.559	35.592	35.592	3.559	35.592	35.592
2	3.227	32.274	67.867	3.227	32.274	67.867
3	1.299	12.986	80.852	1.299	12.986	80.852

4	0.726	7.257	88.109	0.726	7.257	88.109
5	0.451	4.515	92.624			
6	0.289	2.887	95.511			
7	0.168	1.682	97.192			
8	0.155	1.547	98.739			
9	0.113	1.134	99.873			
10	0.013	0.127	100.000			

Table 。

Table1 KMO and Bartlett tests

Number of KMO sample tangents		0.721
Bartlett spherical degree test	Approximate chi-square	12732.239
	degree of freedom	45
	Significance	0.000

Table 4 Principal component characteristic values and contribution rates

ingredients	Initial eigenvalue			Extract the sum of squares of the loads		
	total	Percentage of variance	Cumulative /%	total	Percentage of variance	Cumulative /%
1	3.559	35.592	35.592	3.559	35.592	35.592
2	3.227	32.274	67.867	3.227	32.274	67.867
3	1.299	12.986	80.852	1.299	12.986	80.852
4	0.726	7.257	88.109	0.726	7.257	88.109
5	0.451	4.515	92.624			
6	0.289	2.887	95.511			
7	0.168	1.682	97.192			
8	0.155	1.547	98.739			
9	0.113	1.134	99.873			
10	0.013	0.127	100.000			

Table 5 Principal component load matrix

Index	1	2	3	4
$PM_{2.5}$	0.809	0.374	0.060	0.309
NO_2	0.768	0.465	-0.134	0.259
Average relative humidity (1%)	0.669	-0.200	-0.567	0.318
PM10	0.655	0.406	0.493	-0.062
The average temperature (0.1°C).	0.593	-0.752	0.172	-0.138
Average surface temperature (0.1°C).	0.568	-0.748	0.225	-0.162
SO2	0.122	0.738	0.367	-0.081
The average air pressure of this station (0.1hPa).	-0.581	0.705	-0.205	0.060
CO	0.468	0.684	0.027	-0.386
Average wind speed (0.1m/s).	-0.438	-0.180	0.675	0.504

4.2.2. Neural network construction and prediction results

After factor loading calculation, 4 principal components F1 to F4 are used as inputs. The model operation effect: the average absolute error is 0.2476, the root mean square error is 0.1150, and the interpretation expected score is 88.15%. Using the PCA-BP neural network model, the predicted values of PM2.5 concentration in 2020 are obtained and plotted the actual concentration of PM2.5 in 2020 and the model calculate a fitted curve for the predicted concentration, as shown in Figure 10. It can be seen from the fitting curve of Figure 10 that the experimental method of $PM_{2.5}$-PCA is used to predict, which is basically consistent with the actual $PM_{2.5}$ concentration value.

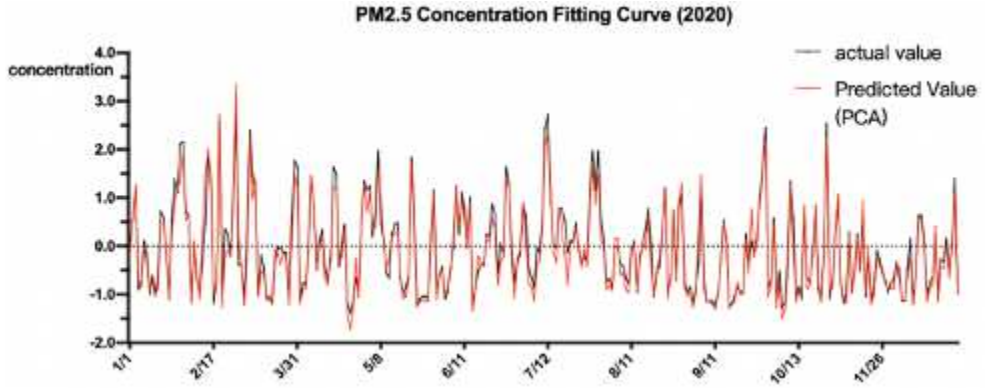

Figure 6 Fitted curve of PM2.5 concentration in 2020

Comparing the prediction effects of PM2.5-5, PM2.5-12, and PM 2.5-PCA, it can be seen that PM The 2.5-PCA model has the lowest RMSE, the highest interpretation expectation score, and the best fit curve effect. It can be obtained that after principal component analysis, multiple variables are reduced to 4 unrelated components, and neural network prediction is performed There will be a significant improvement in the model's predictive performance.

4.3. PM_2 based on KPCA-BP neural network model $_5$ predictions

The experimental environment is Python 3.8. The original data is first preprocessed, then the kernel function is introduced, the dimensionality reduction processing is carried out using the PCA method, and after debugging, the obtained data is used as training samples and test samples. In the KPCA processed data, the data from 2016 to 2019 are used as training data[14], 2020 as test data. After many trainings, the results are output. Table 6 shows the parameters corresponding to the model.

Table 6 KPCA model parameter values

Training parameters	Ways and values
Forecast duration	2021. 1. 1-2020. 12. 31
Output layer activation function	ReLU
The number of neurons in the output layer	128
The number of neurons in the hidden layer	64
Loss function	Mean variance (MSE).
Optimize iterative algorithms	Adam
Epoch	1000
batch_size	128

To compare the prediction accuracy of PCA-BP neural network and KPCA-BP neural network, see Table.

Table 7 Model Fitting Effects

	MAE	RMSE	Explained variance score
PM2.5-PCA Neural Network	0. 2476	0. 1150	0. 8815
PM2.5-KPCA Neural network	0.2358	0. 0978	0. 8921

To make the results clearer, a small amount of data is randomly selected. The experiment takes the forecast data for the 22nd of each month in 2020, and the prediction results of the above two models are shown in Figure 12.

The dot data marker points represent the true values, the triangular data marker points represent the predicted values of the KPCA-BP neural network, and the cross data marker points represent the predictions of the PCA-BP neural network Value. As can be seen from the figure, the distance from the triangular data marker point to the dot data marker point is closer, which illustrates the prediction of the KPCA-BP neural network model The effect is better than other predictive models.

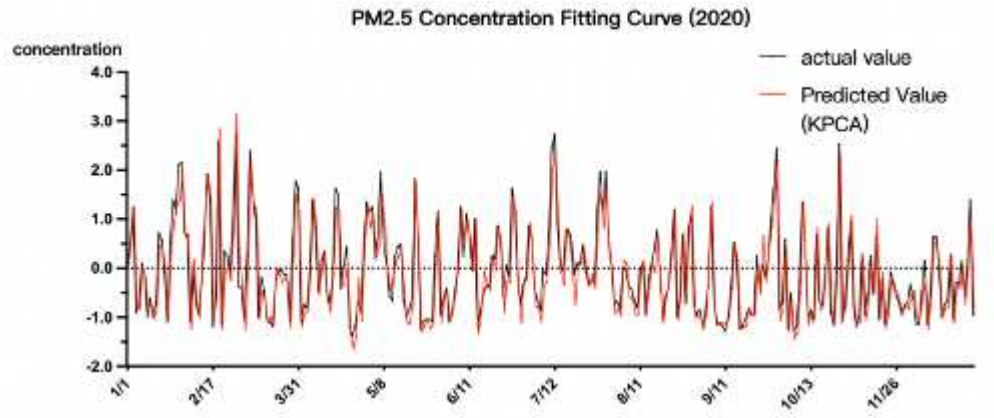

Figure 7 2020 $PM_{2.5}$ concentration fitting curve

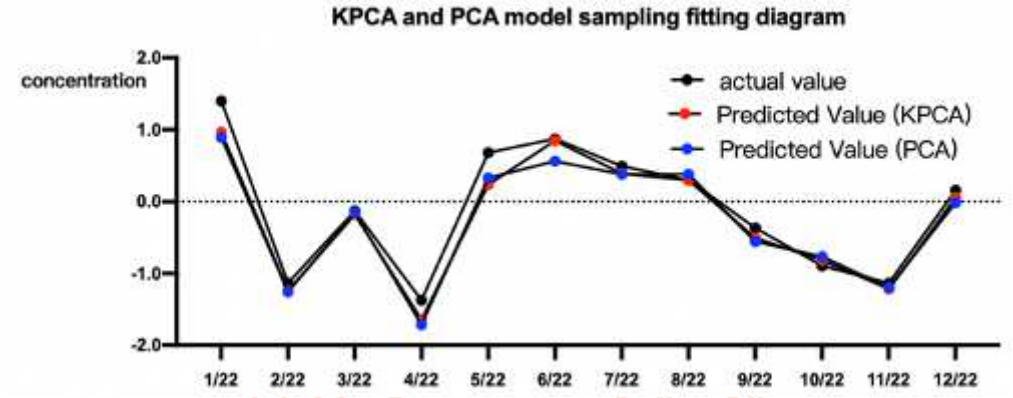

Figure 8 Comparison of KPCA and PCA model predictions

5. Conclusion

In this paper, three neural network models were studied by predicting PM2.5 concentrations in 2020 Accuracy in predicting the annual concentration of meteorological pollutants. A comprehensive evaluation of the comparative analysis of the above models:

1. The K PCA-BP neural network model predicts better results than other models. Compared with the BP neural network and the PCA-BP neural network prediction method, the three evaluation indicators are lower and the stability is better.
2. Based on the prediction results, both combination models have better prediction results, indicating that extracting components can effectively learn the impact factor.
3. Based on the predictions, nuclear principal component analysis is superior to traditional principal component analysis in predicting pollutant concentrations. When the factor increases, the latitude of the data can be lowered more reasonably to improve the accuracy of the prediction results.

With the development of science and technology, the instantaneous acquisition processing power of big data will become more and more mature. The prediction method presented in this paper will have a promising use and is suitable for prediction of the concentration of the other 5 pollutants. Further analysis based on the forecast results will provide support for the decision-making of the prevention and control department.

Bibliography

[1] [1] China Environment News. Transcript of the regular press conference of the Ministry of Ecology and Environment in February [EB/OL](2020-02-26) [2021-11-01]. https://www.cenes.com.cn/news/202102/t20210226_970791.html?from=singlemessage&isappinstalled=0

[2] Ministry of Ecology and Environment of the People's Republic of China. Air Quality Forecast. 京政办發 [A/OL](2021-08-31)[2021-11-01]. https://www.mee.gov.cn/hjzl/dqhj/qgkqzlzk/index_1. shtml

[3] Nitin Tiwari, Neelima Satyam. Coupling effect of pond ash and polypropylene fiber on strength and durability of expansive soil subgrades: an integrated experimental and machine learning approach[J]. Journal of Rock Mechanics and Geotechnical Engineering, 2021, 13(05): 1101-1112.

[4] Xu Yixin, Ren J, Feng Lei, Liang Yinglu, Liu Yiming. Optimization of PM_(2. 5) concentration prediction model based on wavelet analysis[J]. Environmental Monitoring

Management and Technology, 2021, 33(02): 24-28+34.

[5] Ilaboya I R. Performance of Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) as Predictive Tool for Rainfall Modelling. 2019.

[6] Zhao YM. LSTM algorithm based on spatio-temporal correlation and application to PM_(2. 5) concentration prediction [J]. Computer Applications and Software, 2021, 38(06): 249-255+323.

[7] ZJ, Hu HP, Bai YP, Li Q. Research on urban precipitation prediction based on BP, PCA-BP and PLS algorithms[J]. Journal of North Central University (Natural Science Edition), 2016, 37(02): 181-186.

[8] Jiang Shi. Application of PSO-BP neural network in the safety inventory prediction of a coal machine enterprise [J]. Coal Technology, 2017, 36(10): 305-307.

[9] Xie, Zhonghua. Matlab statistical analysis and applications: 40 case studies [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2010.

[10] Ren Jaguar, Gong K, Ma FJ, Gu QB, Wu Qianqian. Prediction of heavy metals and PAHs in contaminated site soil based on BP neural network [J]. Environmental Science Research, 2021, 34(09): 2237-2247.

[11] Liu, Lanfang, Tan, Binglin, Zhang, Ke, Wu, Jinyu. Risk assessment of urban haze disasters in Hunan Province based on principal component analysis [J]. Disaster Science, 2021, 36(01): 76-81.

[12] Guo S. K., Jane T., Dong Y. L. Radar one-dimensional distance image recognition based on PSO-KPCA-LVQ neural network [J]. Electro-Optics and Control, 2019, 26(06): 22-26.

[13] Zhou Q. Y., Zhang S. R., Lai X. T., Wang J. B. Research on fault diagnosis algorithm based on PCA/KPCA initial dictionary optimization[J]. Journal of Sensing Technology, 2020, 33(11): 1599-1607.

[14] Li Xinning, Wu Hu, Yang Xianhai. Multi-view recognition of fruit packing boxes based on features clustering angle[J]. High Technology Letters, 2021, 27(02): 200-209.