Advances in Environmental Protection 环境保护前沿, 2020, 10(5), 745-754
Published Online October 2020 in Hans. http://www.hanspub.org/journal/aep
https://doi.org/10.12677/aep.2020.105090

Hans汉斯

近三年四川省空气污染物浓度与天气因素联合分析

王　楚，韩　琳*

成都信息工程大学大气科学学院，四川 成都
Email: scream1023@163.com, *hanlin@cuit.edu.cn

收稿日期：2020年10月1日；录用日期：2020年10月22日；发布日期：2020年10月29日

摘　要

为了深入研究四川主要空气污染物浓度($PM_{2.5}$, PM_{10})与天气因素和其他污染物(CO, NO_2, O_3, SO_2)的相互影响，本文选用了2016~2018四川14个地区空气质量站点监测日数据集以及对应站点的中国气象台站地面气象资料日值数据集(V3.0)，运用相关分析的方法计算了主要空气污染物与气象要素(温度，气压，降水，相对湿度，风速)和其他污染物浓度的相关系数，并运用多元线性回归与BP神经网络的统计方法分别建立$PM_{2.5}$浓度预测模型，进而对比了两种模型的预报效果。结果表明：1) 主要污染物浓度与天气因素和其他污染物浓度的相关性在不同地区有着较大差异。总体来看，主要空气污染物浓度与温度、气压、降水有显著相关关系；2) 运用多元线性回归和BP神经网络两种方法分别建立的$PM_{2.5}$浓度预测模型显示，在相同数据条件下，BP神经网络预测的预报效果相较于多元线性回归更具优越性。

关键词

空气污染物，气象要素，相关分析，多元线性回归，BP神经网络

Joint Analysis of Air Pollutant Concentration and Weather Factors in Sichuan Province in the Past Three Years

Chu Wang, Lin Han*

School of Atmospheric Sciences, Chengdu University of Information Technology, Chengdu Sichuan
Email: scream1023@163.com, *hanlin@cuit.edu.cn

*通讯作者。

文章引用: 王楚, 韩琳. 近三年四川省空气污染物浓度与天气因素联合分析[J]. 环境保护前沿, 2020, 10(5): 745-754.
DOI: 10.12677/aep.2020.105090

Received: Oct. 1st, 2020; accepted: Oct. 22nd, 2020; published: Oct. 29th, 2020

Abstract

In order to further study the interaction between the main air pollutant concentrations in Sichuan ($PM_{2.5}$, PM_{10}) and weather factors and other pollutants (CO, NO_2, O_3, SO_2), this paper selects the air quality monitoring daily data sets of 14 regional air quality stations in Sichuan from 2016 to 2018 and the daily value data sets (V3.0) of surface meteorological data of China Meteorological Observatory of corresponding sites, using relevant analysis methods to find the main air pollutants correlation coefficients with meteorological elements (temperature, barometric pressure, precipitation, relative humidity, wind speed) and other pollutant concentrations, and using multiple linear regression and BP neural network methods to establish the $PM_{2.5}$ concentration prediction model, and then compare the prediction effects of the two models. The results show that 1) in the past three years, the correlation between the concentration of main pollutants and weather factors and the concentration of other pollutants varies greatly in different regions. Overall, the concentration of major air pollutants is significantly correlated with temperature, air pressure, and precipitation. 2) The $PM_{2.5}$ concentration prediction models established by multiple linear regression and BP neural network methods show that under the same data conditions, the prediction effect of BP neural network prediction is superior to multiple linear regression.

Keywords

Air Pollutants, Meteorological Elements, Correlation Analysis, Multiple Linear Regression, BP Neural Network

1. 引言

随着四川实体工业经济的迅速发展，城市化进程进一步扩大。污染物排放量维持在较高排放水平。四川盆地大致位于亚洲的中南部，中国的西南部，盆地的东部为长江三峡，南部为云贵高原，西部为青藏高原，北部为大巴山，形成四周高中间低的信封盆状的地形[1]。特殊的地形条件形成了暖湿风少、云雾多、相对湿度大、平均年静风率46%以上等气象条件[2]，对大气污染物时空分布具有重要的影响。四川盆地属中亚热带湿润季风气候，四川省的空气污染在冬季最严重，而在夏季最轻。风速和相对湿度对颗粒物浓度的影响显著，并显示出明显的相关性。迄今为止，对于大气污染物与气象条件相关分析的研究，大多数运用相关分析[3] [4]；或探讨重度污染日当天环流形势，分析污染物的输送与扩散条件[5] [6] [7]；或利用常规台站资料结合中尺度气象模式 WRF 模拟结果，分析研究地区气象特征湍流特征，以此讨论污染的扩散和水平输送[8] [9] [10] [11]。本文通过收集了四川省 14 个主要地区 2016~2018 年空气污染物浓度和其对应站点气象要素数据，系统地分析了主要空气污染物($PM_{2.5}$ 和 PM_{10})和 AQI 指数的年际变化、季节特征、天气特征以及与其它空气污染物、地面常规气象要素之间的系统关系。并根据主要空气污染物与各要素之间的相关关系建立多元线性回归预测模型与 BP 网络预测模型，对比了两个模型的预报效果为污染物浓度的预报提供思路。

2. 资料与方法

2.1. 资料选取

资料选用，2016~2018 四川 14 个地区空气质量站点监测日数据集以及对应的中国气象台站地面气象资料日值数据集(V3.0)。因日数据存在缺测值与异常值，所以对各项数据进行月平均处理。

2.2. 研究方法

2.2.1. 相关分析

相关分析方法是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。两个数据之间的相关性程度通过相关系数 R 来表示。相关系数的计算公式为：

$$R=\frac{\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\Sigma\left(X_i-\bar{X}\right)^2\Sigma\left(Y_i-\bar{Y}\right)^2}} \tag{1}$$

相关系数 R 的值大于−1 小于 1。当 R 大于 0 小于 1，两数据呈正相关时；当 R 大于−1 小于 0，两数据呈负相关。R 的绝对值越接近 1，两数据之间的相关性越强，r 的绝对值越接近 0，两数据之间的相关性越弱。

2.2.2. 多元线性回归

多元线性回归是分析有两个或多个因素自变量的线性相关关系的方法。一种数据常常是与多个因素相联系的，由多个自变量的最优组合共同来预测因变量的值，比利用单个自变量进行预测更有效，更符合实际。因此多元线性回归普遍用于某些数据的预测上。实际操作时，利用 MATLAB 的 regress 函数，求出各项系数，从而建立回归方程。

2.2.3. BP 神经网络

BP 神经网络，即误差逆向传播算法训练的多层前馈神经网络，是 1986 年由 Rumelhart 和 McClelland 为首的外国科学家共同提出的概念，在当今世界得到广泛运用。

人工神经网络无需事先设定输入输出之间映射关系的数学方程，仅通过自身的训练和学习，在给定输入值时就能得到最接近期望的输出值的结果。BP 神经网络是一种按误差反向传播训练的多层前馈网络，它的核心方法为梯度下降法，使网络的实际输出值和期望输出值的误差均方差为最小[9]。

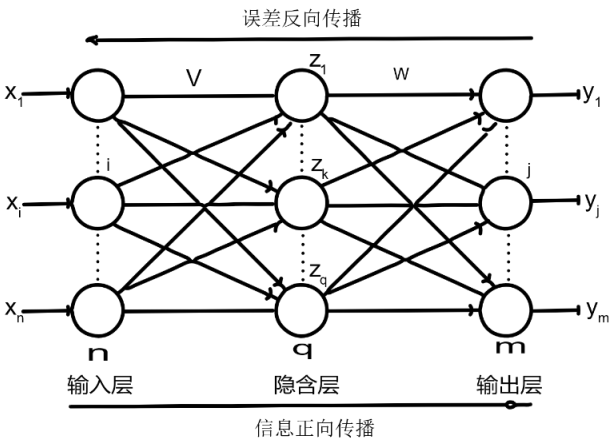

Figure 1. BP neural network structure diagram
图 1. BP 神经网络结构图

如图 1 神经网络结构图，BP 神经网络的基本结构分为输出层，隐含层和输出层，各层由传输函数连接，但是各层中的数据单元之间无联系。网络运行前，设置训练目标，当网络运行时，输入数据通过隐含层作用域输出节点，经过非线性变换，产生输出数据，当输出值误差大于期望时，误差则经过反向传递作用到各层传输节点(图中 V 和 W)，调整其连接权值和阈值，并再次输出，直到达到设定的目标精度。实际训练时，利用 MATLAB 建立模型，设定好网络参数。利用 mse 函数求出其均方误差。

3. 四川三年月平均气象数据与 $PM_{2.5}$，PM_{10} 浓度相关分析

由图 2 可以看出，成都 $PM_{2.5}$ 和 PM_{10} 月平均浓度与月平均温度变化有显著的负相关关系，且相关性最高，其相关系数分别为为−0.792 ($p < 0.001$)和 0.793 ($p < 0.001$)，随温度升高，$PM_{2.5}$ 和 PM_{10} 浓度降低；成都月平均气压与月平均 $PM_{2.5}$ 和 PM_{10} 浓度具有显著正相关性质，相关系数分别为 0.711 ($p < 0.001$)和 0.725 ($p < 0.001$)，随气压升高，$PM_{2.5}$ 和 PM_{10} 浓度升高；成都月平均降水量与 $PM_{2.5}$ 和 PM_{10} 浓度具有显著负相关性质，相关系数分别为−0.604 ($p < 0.001$)和−0.574 ($p = 0.03$)，随降水量增加，$PM_{2.5}$ 和 PM_{10} 浓度降低；成都月平均相对湿度与月平均 $PM_{2.5}$ 和 PM_{10} 浓度之间均无显著相关性($R = -0.25$, $p = 0.142$ 和 $R = -0.149$, $p = 0.488$)；成都月平均风速与月平均 $PM_{2.5}$ 和 PM_{10} 浓度之间存在较为显著的负相关性，相关系数分别为−0.454 ($p = 0.05$)和−0.498 ($p = 0.013$)，平均风速较大的月份 $PM_{2.5}$ 和 PM_{10} 浓度值较小。

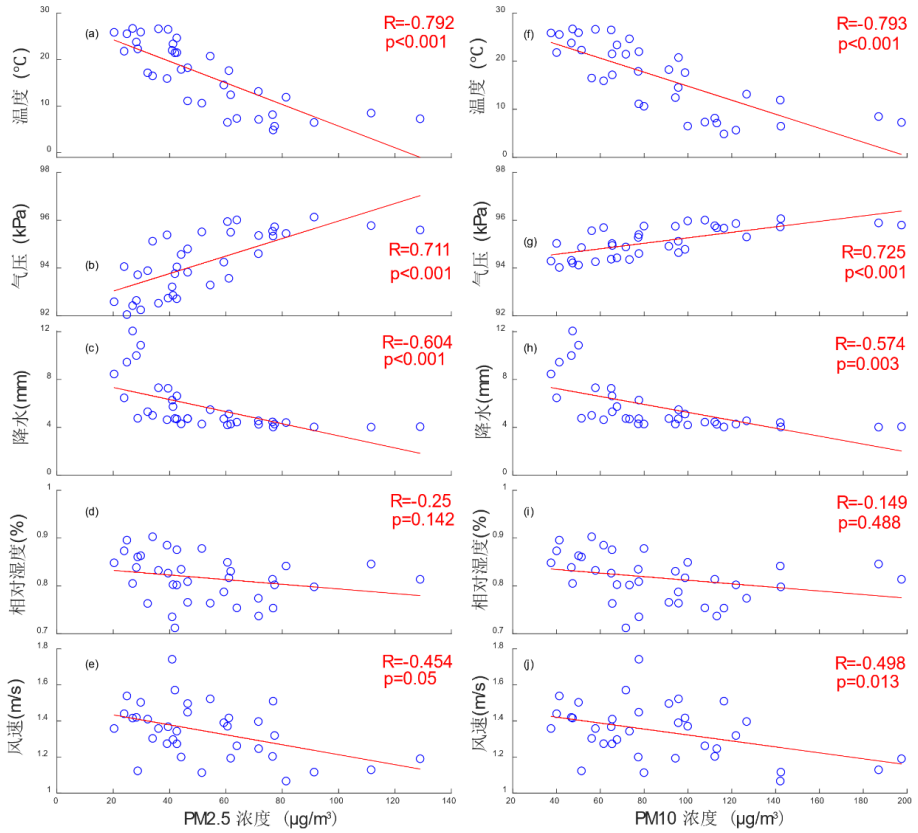

Figure 2. Scatter plots of major pollutants and various meteorological elements (a) Temperature and $PM_{2.5}$, (b) Air pressure and $PM_{2.5}$, (c) Precipitation and $PM_{2.5}$, (d) Relative humidity and $PM_{2.5}$, (e) Wind speed and $PM_{2.5}$, (f) The relationship between average temperature and PM_{10}, (g) Atmospheric pressure and PM_{10}, (h) Precipitation and PM_{10}, (i) Relative humidity and PM_{10}, (j) Wind speed and PM_{10} concentration

图 2. 主要污染物与各类气象要素散点图(a) 气温与 $PM_{2.5}$，(b) 气压与 $PM_{2.5}$，(c) 降水量与 $PM_{2.5}$，(d) 相对湿度与 $PM_{2.5}$，(e) 风速与 $PM_{2.5}$，(f) 平均气温与 PM_{10} 的关系，(g) 气压与 PM_{10}，(h) 降水量与 PM_{10}，(i) 相对湿度与 PM_{10}，(j) 风速与 PM_{10} 浓度

Table 1. Correlation coefficients between the $PM_{2.5}$ concentration values of 12 cities in Sichuan and the average monthly meteorological elements
表 1. 四川 12 个城市 $PM_{2.5}$ 浓度值与各个月平均气象要素的相关系数

地区	温度	气压	降水	相对湿度	风速
成都	−0.792	0.711	−0.604	−0.250	−0.454
广元	−0.872	0.794	−0.555	−0.608	−0.264
遂宁	−0.802	0.715	−0.582	−0.050	−0.323
内江	−0.832	0.781	−0.645	−0.036	−0.406
乐山	−0.817	0.764	−0.654	0.013	−0.641
眉山	−0.770	0.514	−0.689	−0.563	0.703
达州	−0.848	0.809	−0.550	0.271	−0.688
雅安	−0.826	0.784	−0.672	0.010	−0.382
巴中	−0.841	0.783	−0.696	0.195	−0.712
阿坝州	−0.542	−0.208	−0.590	−0.555	0.366
甘孜州	−0.543	−0.623	−0.445	−0.714	−0.026
凉山州	−0.755	0.455	−0.601	−0.403	0.122

Table 2. Correlation coefficients between PM_{10} concentration values and monthly average meteorological elements in 12 cities in Sichuan
表 2. 四川 12 个城市 PM_{10} 浓度值与各个月平均气象要素的相关系数

地区	温度	气压	降水	相对湿度	风速
成都	−0.793	0.725	−0.574	−0.149	−0.498
广元	−0.818	0.723	−0.557	−0.660	−0.212
遂宁	−0.736	0.646	−0.522	−0.225	−0.136
内江	−0.796	0.734	−0.612	−0.130	−0.310
乐山	−0.814	0.748	−0.651	−0.030	−0.617
眉山	−0.771	−0.550	−0.694	−0.617	0.734
达州	−0.848	0.797	−0.537	0.233	−0.675
雅安	−0.842	0.798	−0.702	−0.016	−0.423
巴中	−0.841	0.761	−0.713	0.067	−0.653
阿坝州	−0.629	−0.339	−0.652	−0.697	0.498
甘孜州	−0.437	−0.601	−0.463	−0.708	0.191
凉山州	−0.759	0.367	−0.647	−0.562	0.283

由表 1 和表 2 可以看出，温度，气压和降水与两项主要污染物($PM_{2.5}$, PM_{10})浓度存在显著相关关系。在 13 个主要地区，温度都与两项主要污染物浓度之间存在负相关关系。值得一提的是，在甘孜州和阿坝州，月平均气压与 PM_{10} 浓度之间存在负相关关系，甘孜州月平均气压与 $PM_{2.5}$ 浓度存在负相关关系，甘孜州平均海拔为 4100 米，阿坝州平均海拔为 3000 米，判断污染物浓度可能受地形影响。

4. 线性回归与 BP 神经网络

4.1. 线性回归

1. 数据处理

由成都月平均气象要素和 $PM_{2.5}$ 浓度相关分析结论，所以将一月日平均温度、日平均气压、日平均降水量、日平均风速、其他污染物一月日平均浓度以及前一日 $PM_{2.5}$ 浓度作为自变量，因变量为当日 $PM_{2.5}$ 日平均浓度值。

2. $PM_{2.5}$ 浓度预测模型训练过程

1) 将 2016~2018 年 $PM_{2.5}$ 一月份日平均浓度数据和各气象要素日平均数据作为训练集。并将各项数据归一化至−1 到 1 之间。

2) 运用 k 折交叉检验将训练集随机划分成 10 个子集(即 k = 10)，其中 9 个作为训练集 1 个作为测试集。

3) 利用 MATLAB 建立多元线性回归模型，利用 regress 函数求出各项系数，函数模型建立后，输出项为一月份前 10 天日平均 $PM_{2.5}$ 浓度。

4) 因为 k 折交叉检验子集为随机划分，所以再将模型运行 1000 遍，选取训练集中均方误差最小的模型。

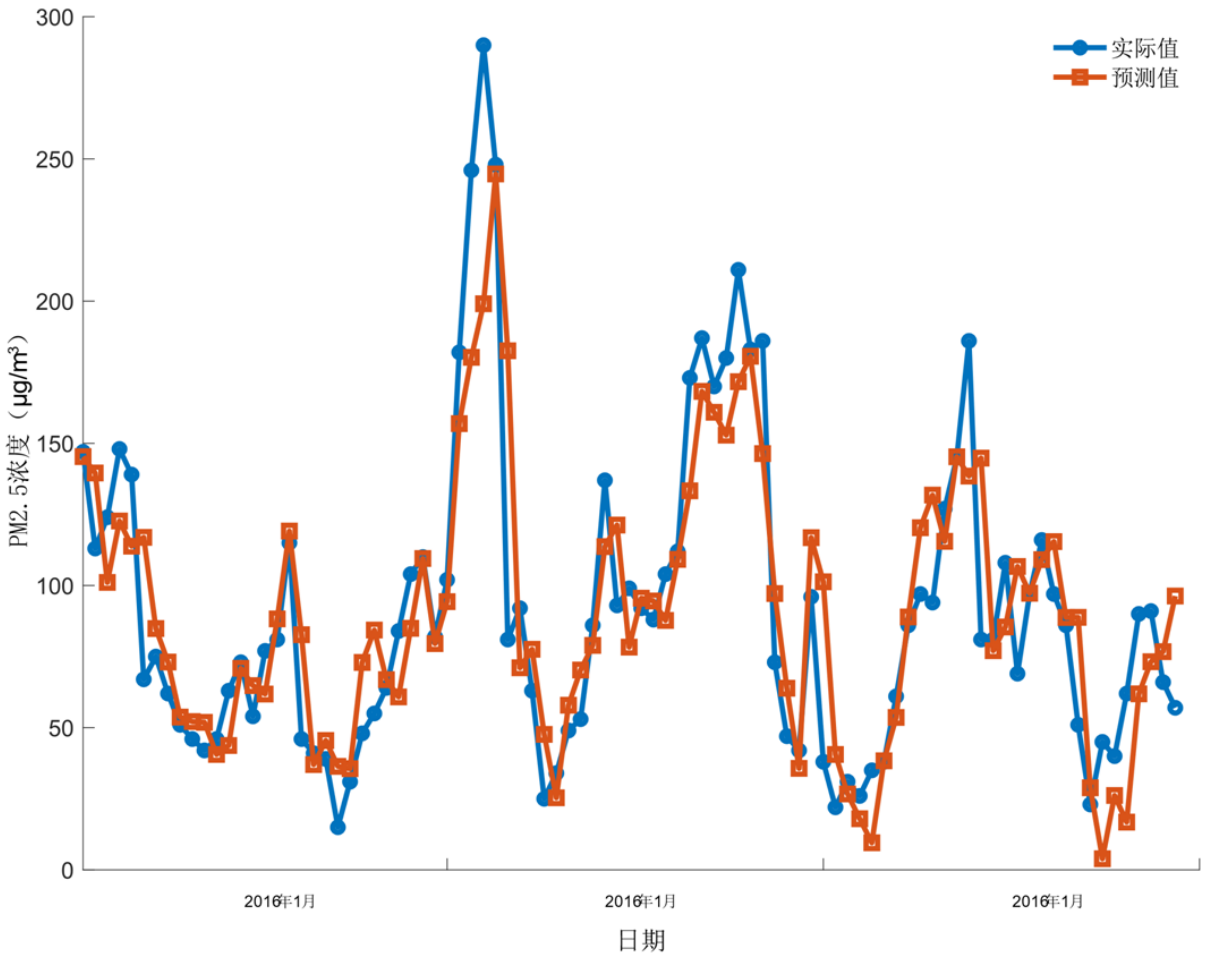

Figure 3. Effect map of linear regression training
图 3. 训练效果图

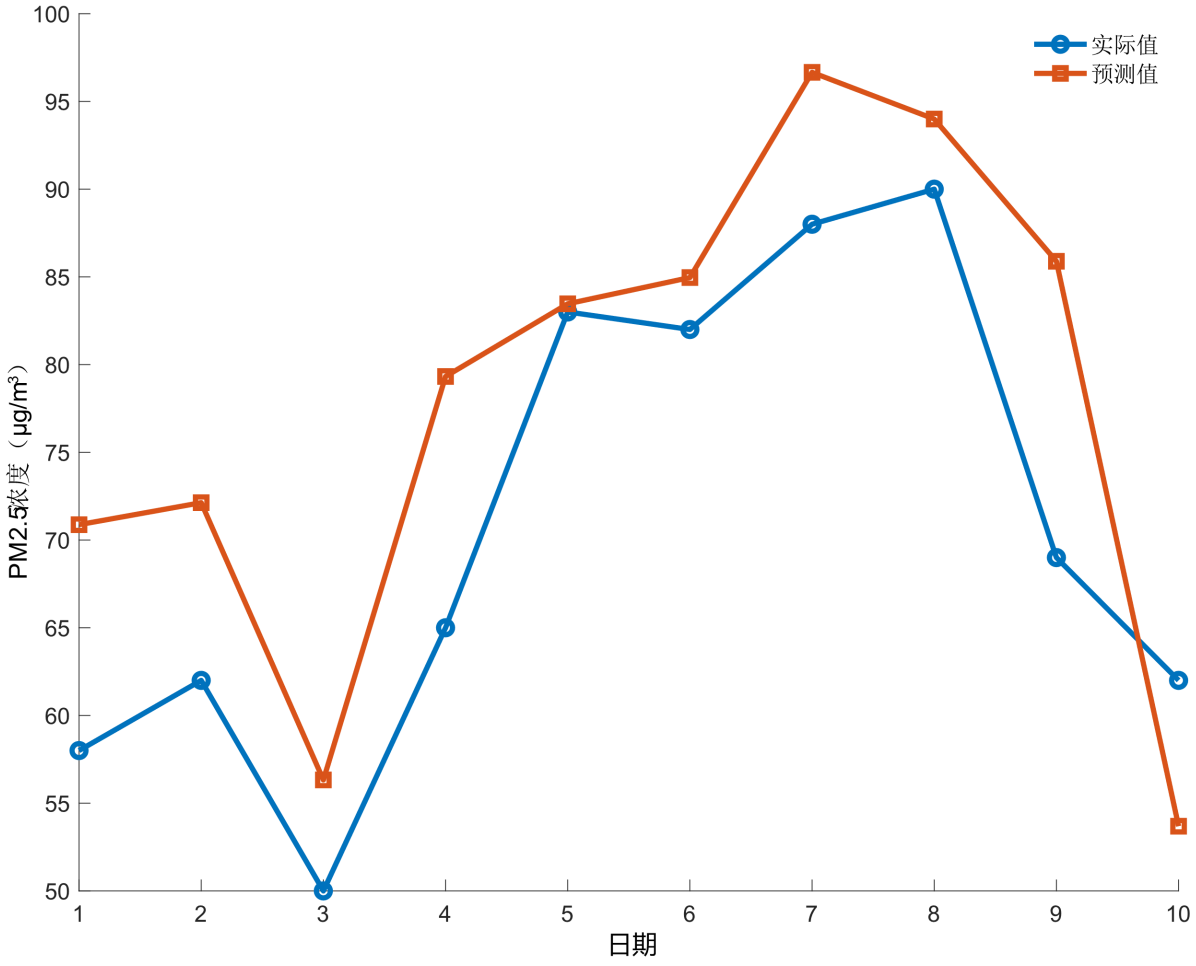

Figure 4. Effect chart of linear regression test
图 4. 线性回归测试效果图

Table 3. Relative set error model and absolute multiple error prediction
表 3. 多元线性预测模型测试集绝对误差和相对误差

时间	实际值	预测值	绝对预报误差	相对预报误差
1 月 1 日	58.00	70.87	−12.87	−22.20%
1 月 2 日	62.00	72.13	−10.13	−16.34%
1 月 3 日	50.00	56.33	−6.33	−12.65%
1 月 4 日	65.00	79.32	−14.32	−22.03%
1 月 5 日	83.00	83.47	−0.47	−0.56%
1 月 6 日	82.00	84.96	−2.96	−3.61%
1 月 7 日	88.00	96.66	−8.66	−9.84%
1 月 8 日	90.00	94.00	−4.00	−4.44%
1 月 9 日	69.00	85.89	−16.89	−24.48%
1 月 10 日	62.00	53.69	8.31	13.41%

由图 3 线性回归训练效果可以看到，均方误差为 759.54，$R^2 = 0.794$ ($p < 0.001$)。月平均 $PM_{2.5}$ 浓度低值，预测曲线可以达到，但 $PM_{2.5}$ 浓度高值，预测曲线无法符合。由图 4 与表 3 可以看到均方误差为 96.78，$R^2 = 0.744$ ($p < 0.001$)。除了一月 10 日预测值小于实际值，其余预测值均大于实际值，并且随着预报时次延长，预报效果减弱，最大相对误差达到 22.48%。

4.2. BP 神经网络

1. 数据处理

自变量和因变量与多元线性回归模型一样，做相同处理。

2. $PM_{2.5}$ 浓度预测模型训练过程

1) 并将各项数据归一化至−1 到 1 之间。

2) 运用 k 折交叉检验将训练集随机划分成 10 个子集(即 k = 10)，其中 9 个作为训练集一个作为测试集。

3) 为确定最佳隐含层节点数，运用经验公式：$m = \sqrt{n+l} + \alpha$ (其中，m 为隐含层节点数，n 为输入层节点数，l 为输出层节点数，α 为 1~10 之间的常数)，因为本次输入层节点数 n = 9，输出层节点数 l = 1，所以本次 BP 神经网络模型的最佳隐含层节点数在 5 到 15 之间。设置循环，选出最佳隐含层节点数。

4) 设置此次 BP 网络预测模型参数：a) 最大训练次数为 1000 次，b) 为了防止训练集过拟合导致测试集预报效果差，设置训练目标精度为 0.001；c) 为防止过大学习率，临近最佳点产生动荡，导致无法收敛，设置学习率为 0.01；d) 设置输入层传输函数为 logsig，输出层传输函数为 purelin。

5) 因为 k 折交叉检验子集为随机划分，所以再将模型运行 1000 遍，选取训练集中均方误差最小的模型。

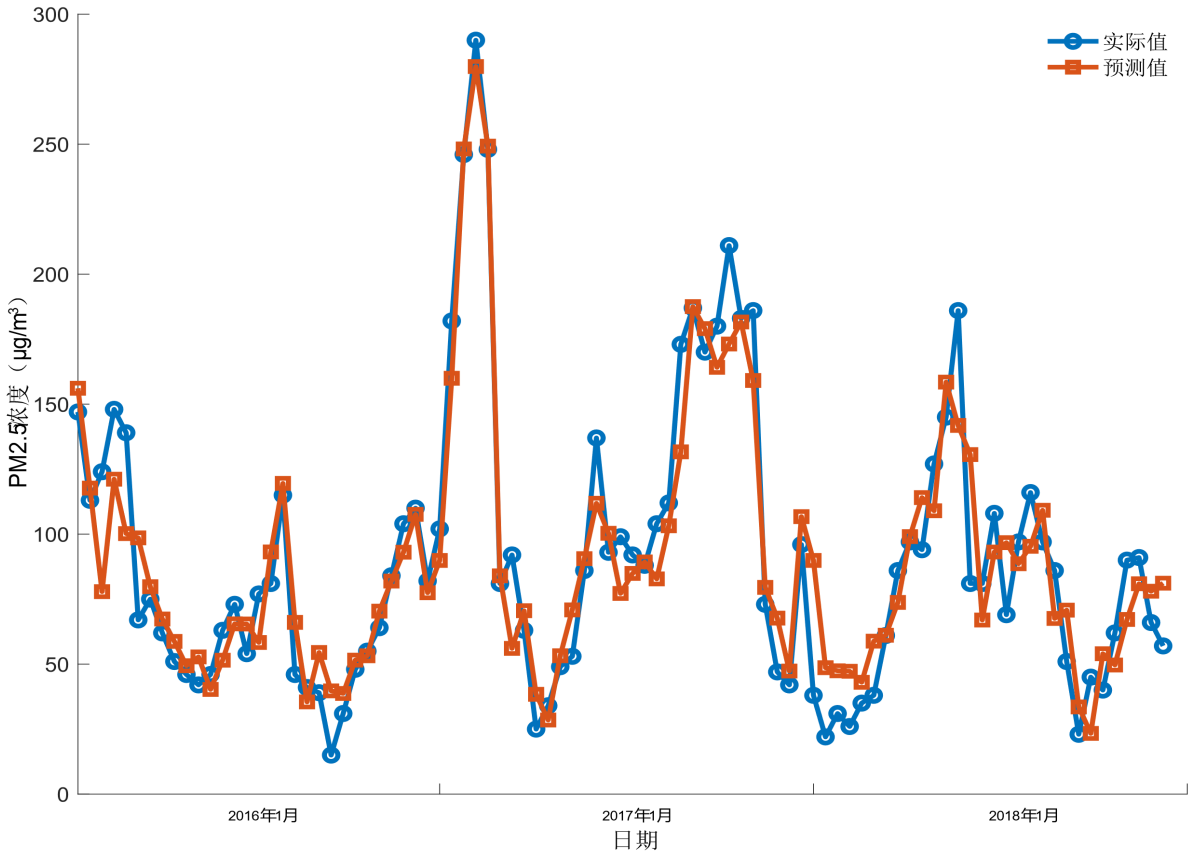

Figure 5. Training effect chart of BP neural network
图 5. BP 网络训练效果图

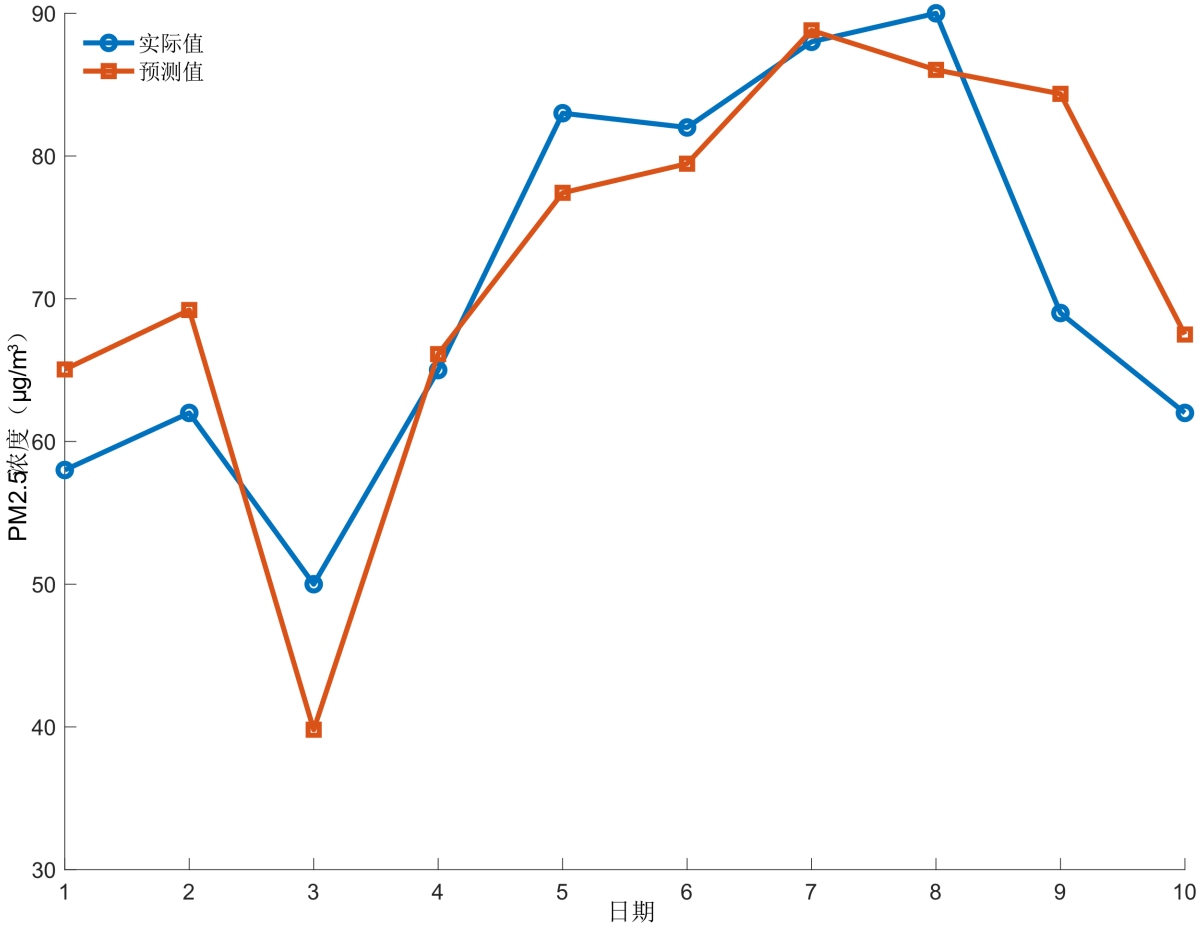

Figure 6. BP neural network test effect chart
图 6. BP 网络测试效果图

得到最佳隐含层节点数为 7。由图 5 训练效果图可以看到均方误差为 354.68，$R^2 = 0.885$ ($p < 0.001$)，预测整体趋势与实际值大致符合，但几处低值与高值预测曲线并未符合。

Table 4. Absolute error and relative error of BP network prediction model test set
表 4. BP 网络预测模型测试集绝对误差和相对误差

时间	实际值	预测值	绝对预报误差	相对预报误差
1 月 1 日	58.00	65.05	−7.05	−12.15%
1 月 2 日	62.00	69.21	−7.21	−11.63%
1 月 3 日	50.00	39.80	10.20	20.39%
1 月 4 日	65.00	66.12	−1.12	−1.73%
1 月 5 日	83.00	77.42	5.58	6.72%
1 月 6 日	82.00	79.46	2.54	3.09%
1 月 7 日	88.00	88.80	−0.80	−0.91%
1 月 8 日	90.00	86.04	3.96	4.40%
1 月 9 日	69.00	84.36	−15.36	−22.26%
1 月 10 日	62.00	67.49	−5.49	−8.86%

由图 6 与表 4 可以看出均方误差为 52.68，$R^2 = 0.74$ ($p < 0.001$)。预报值在实际值曲线上下波动，随着预报时间的增长，预报效果变差，最大相对误差为 20.26%，但是相较于多元线性预报模型，拟合程度更高，均方误差更小，所以在相同的数据条件下，BP 神经网络的预测效果更好。

将输入层到隐含层的权值归一化后得到输入层到隐含层各项因子的权重，即可得到各项因子对预测模型的贡献率。

3. 模型改进

由图 5，训练效果图可以看出，对于污染物浓度高值的日期预测效果并不好，且测试效果显示随着预报时间延长，预测误差逐渐增大，预测效果逐渐变差。所以，改建方案主要涵盖以下几点：1) 更长更精细的时间周期的数据会得出更好的结果；2) 带入更多的变量，如边界层高度资料，污染物排放量资料；3) 污染物浓度的预测不应至限于局部，还应考虑到其他地区的输送，以及地形的影响。

5. 结论与分析

1) 四川主要污染物浓度($PM_{2.5}$ 和 PM_{10})与天气因素的相关性在不同地区有较大差别。总体来看，主要空气污染物浓度与温度、气压、降水浓度有显著相关关系。

2) 在相同数据条件下，神经网络预报模型的预报结果要优于多元线性预报模型。

基金项目

2018~2020 年校级重点高等教育人才培养质量和教学改革项目(JY2018069)；《大气探测学》精品在线开放课程建设项目(BKJX2019062)。

参考文献

[1] 张继娟, 魏世强. 我国城市大气污染现状与特点[J]. 四川环境, 2006(3): 104-108+112.

[2] 龙亚萍, 李立华. 四川省山地旅游气候资源评价[J]. 山地学报, 2018, 36(1): 116-124.

[3] 廖志恒, 孙家仁, 范绍佳, 等. 2006~2012 年珠三角地区空气污染变化特征及影响因素[J]. 中国环境科学, 2015, 35(2): 329-336.

[4] 王式功, 杨德保, 尚可政, 等. 兰州市城区冬半年低空风特征及其与空气污染的关系[J]. 兰州大学学报, 1997(3): 101-106+108-109.

[5] 郭倩, 汪嘉杨, 周子航, 等. 成都市一次典型空气重污染过程特征及成因分析[J]. 环境科学学报, 2018, 38(2): 629-639.

[6] 张娟, 刘志红, 段伯隆, 等. 2014 年成都市大气污染特征及气象因子分析[J]. 四川环境, 2016, 35(6): 79-88.

[7] 郭立平, 乔林, 石茗化, 等. 河北廊坊市连续重污染天气的气象条件分析[J]. 干旱气象, 2015, 33(3): 497-504.

[8] 王明洁, 贺佳佳, 王书欣, 等. 基于 AQI 的深圳大气污染特征及其典型环流形势分析[J]. 生态环境学报, 2018, 27(2): 268-275.

[9] 王自发, 李杰, 王哲, 等. 2013 年 1 月我国中东部强霾污染的数值模拟和防控对策[J]. 中国科学: 地球科学, 2014, 44(1): 3-14.

[10] 郭建平, 吴业荣, 张小曳, 等. BP 网络框架下 MODIS 气溶胶光学厚度产品估算中国东部 PM_(2.5) [J]. 环境科学, 2013, 34(3): 817-825.

[11] 杨旭, 康延臻, 王式功, 等. 郑州市大气污染特征及其与气象条件的关系[J]. 兰州大学学报(自然科学版), 2017, 53(3): 348-354.