空气质量数据的校准

摘要

本文假设国控点数据为真实数据，通过计算自建点数据与理想数据的相关系数分析了自建点与国控点数据的内涵。通过建立 BP 神经网络模型，对某公司自主研发的微型空气质量检测仪测量数据进行了校准，给出的误差值更小污染物预测解决方案。

对于问题一，首先假设国控点检测的数据为准确数据，然后对数据进行预处理，把存在问题的数据，如国控点数据和自建点数据在日期上缺失或多出的数据，以及数值异常的数据剔除掉。然后把国控点和自建点的数据作周平均处理和日平均处理，用 EXCEL 作出对比折线图，通过对比图得出了 PM2.5、PM10、CO、NO、SO2、O3 六种污染物的季节性变化趋势，以及自测点与国控点数据的之间关系。发现自测点在测试 PM2.5、PM10 这两个指标时较准确，而在测试 CO、NO、SO2、O3 时误差较大。

对于问题二，首先把国控点与自建点的数据进行归一化处理，对归一化后的污染物指标数据计算其相互之间的 Pearson 相关系数，进而得到了国测点 6 个指标与自测点 11 个指标之间的相关系数矩阵。通过相关系数矩阵，分析了每一个污染物指标的自测点与国控点之间的相关性，发现 CO、NO、SO2、O3 四个值的相关系数 $r<0.4$，说明自测点对这四个指标的测量不准确。最后通过归一化的误差值与归一化的天气值指标之间的关系图，分析了天气因素对误差之间的关系。

对于问题三，属于回归分析问题，文章采用 BP 神经网络来对数据进行回归分析。为了让神经网络的预测性能更好，在训练神经网络时采用了如下的优化手段：（1）数据归一化处理，减小数据分布偏差；（2）训练集随机选择，增加数据多样性；（3）对预测值采用多次预测求平均值的方法减少预测的误差；（4）优化 BP 网络的隐层数目，使网络性能达到最优。通过训练好的 BP 网络模型，得出了国控值、BP 预测值、自建点测量值三者之间的对比关系图，发现模型对污染物指标的预测能力得到的极大的提升，如 PM2.5 的准确度由自建点的 48.22%提升到 90.35%，PM10 的准确度由自建点的 54.06%提升到 89.37%。

关键词： 空气污染　数据处理　相关系数　均方误差　BP 神经网络

一、问题的重述

大气污染对生态环境和人类健康构成巨大的危害，通过对 PM2.5、PM10、CO、NO2、SO2、O3 浓度的实时监测可以及时掌握空气质量并采取相应的防治措施。虽然国家监测站点（国控点）对“两尘四气”有监测数据，且较为准确，但因为国控点的布控较少，数据发布时间滞后较长且花费较大，并不能给空气质量作出实时的监测和预报。某公司自主研发的微型空气质量仪相对便宜，可以实时网格化监控某一地区的空气质量，并实时监控温度、湿度、风速、气压、降水等气象参数。

由于微型空气质量仪所使用的电化学气体传感器在长时间使用后会产生一定的零点漂移和量程漂移，非常规气态污染物（气）浓度变化对传感器存在交叉干扰，以及天气因数对传感器的影响，在国控点附近所安放的自建点上，同一时间微型空气质量仪所测得的数据与该国控点的数据值有所不同，所以需要利用国控点每小时的数据对附近的自建点数据进行调整。

附件已提供了相关数据。请建立数学模型研究以下问题：

1. 对自建点数据与国控点数据进行探索性数据分析。

2. 对导致自建点数据与国控点数据造成差异的因素进行分析。

3. 利用国控点数据，建立数学模型对自建点数据进行校准。

二、问题分析及模型假设

2.1 对于问题一的分析

在对于自建点数据与国控点数据进行探索性数据分析前，我们首先对数据进行了预处理。分析数据发现国控点的数据和自建点的数据中有个别日期出现了缺失、多出现象，为此先把对应不上的日期剔除掉，以免对分析造成影响。我们还发现在 2018 年 11 月 24 号这一天自建点检测的 SO2 数据异常，经过与国控点同一天 SO2 的浓度对比以及观察前后一个月的浓度变化得知这天明显不符合规律，推测应是该天自建点传感器出故障，或则监测点附近的人类活动造成了数据异常，为不影响分析，应对该天数据进行修正。接着对附件 1 和附件 2 的数据进行处理，由于把污染物每天不同时间段浓度变化作图过于繁杂，且不利于分析此问题，所以我们用污染物的国控点与自建点的浓度的日平均值与周平均值作图比较，通过两者的趋势图探索自建点和国控点之间的规律与联系。

2.2 对于问题二的分析

对于此问题，需要把自建点测得的各种污染物参数在加入天气参数条件下跟国控点参数对比，因为各个污染物数据的单位、数值范围等都不相同为了便于后续的分析与处理，我们把国控点和自建点所测得的所有数据统一进行了归一化处理。

为了准确简单找出各个变量之间的联系，我们引用 Pearson 相关系数来研究问题中各个变量的相关性，以此来衡量变量之间的相关程度。相关系数 r 计算公式如下：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} \qquad (1)$$

r 的取值范围为-1 到 1 之间，即-1≤r≤1，当 r 越趋向 1（ 或-1）代表相关程度越大，越趋向 0 代表相关程度越低，1 为正相关，-1 为负相关。只要掌握了国控点数据与自建点数据之间的线性相关度，便能直观地知道国控点数据和自建点数据差异的因数，并进而进行具体的分析了。

2.3 对于问题三的分析

在前面的问题研究中已经把自建点异常的数据处理了，在利用国控点数据建立模型时无须考虑数据的影响偏差。建立一个模型首先要设计变量，在此问题中，自建点的 11 个测量值都可以认为是影响污染物真实指标的变量，由于因变量太多，加上因变量之间可能存在交叉干扰，因此，采用传统的方法建立模型对自建点数据进行修正难度较大，于是我们考虑用 BP 神经网络来对目标函数进行模拟。文章首先利用自建点与国控点的数据训练神经网络，把自建点数据作为输入，把国控点某个污染物的数据作为输出，并且实验不同的隐层对 bp 网络预测性能的影响。最后我们得到了能通过输入自建点数据预测真实污染物指标的 bp 神经网络。

三、模型假设

1. 假设国控点测得的数据绝对准确。
2. 假设自建点除了零点漂移和量程漂移外不出现其他误差。
3. 假设自建点测量的温度、湿度、风速、气压、降水等气象参数绝对准确

四、符号说明

符号	含义
r	相关常数
L	BP 神经网络隐层数目
MSE	均方误差值
O_t	BP 神经网络预测值
Y_t	国测点污染物数据
P_t	自建点原始数据值
P_bp	BP 网络预测准确率
P_Z	自建点准确率
BZ	比例系数 BZ=MSE_BP/MSE_Z
BZ_J	比例系数 BZ_J=MSE_BP_J /MSE_Z_J

五、模型建立、求解与数据分析

5.1 问题一：对自建点数据与国控点数据进行探索性数据分析

在建模或数据分析前，首先要做的事情就是数据预处理，数据预处理的好坏，很大程度上决定了模型或分析结果的好坏。其中，异常值（outliers）检测是整个数据预处理过程中，十分重要的一环。通过分析自建点与国控点的数据，我们发现自建点与国控点有各别天数（如：国控点 2019-1-12 和自建点 2019-1-14、2019-4-10 天）是没有对应上的，于是我们把这几天数据剔除，这样得到的 206 天的数据就是我们进行分析的基础数据了。

分析数据发现自建点的数据是以分钟为单位记录数据的，所以其数据量远远多于国控点。为了能对比性分析自建点与国控点的数据关系，我们把自建点与国控点的数据进行了均值化处理，分别得到自建点的日均数据和周均数据以及国控点的日均数据和周均数据。

5.1.1 PM2.5 数据分析

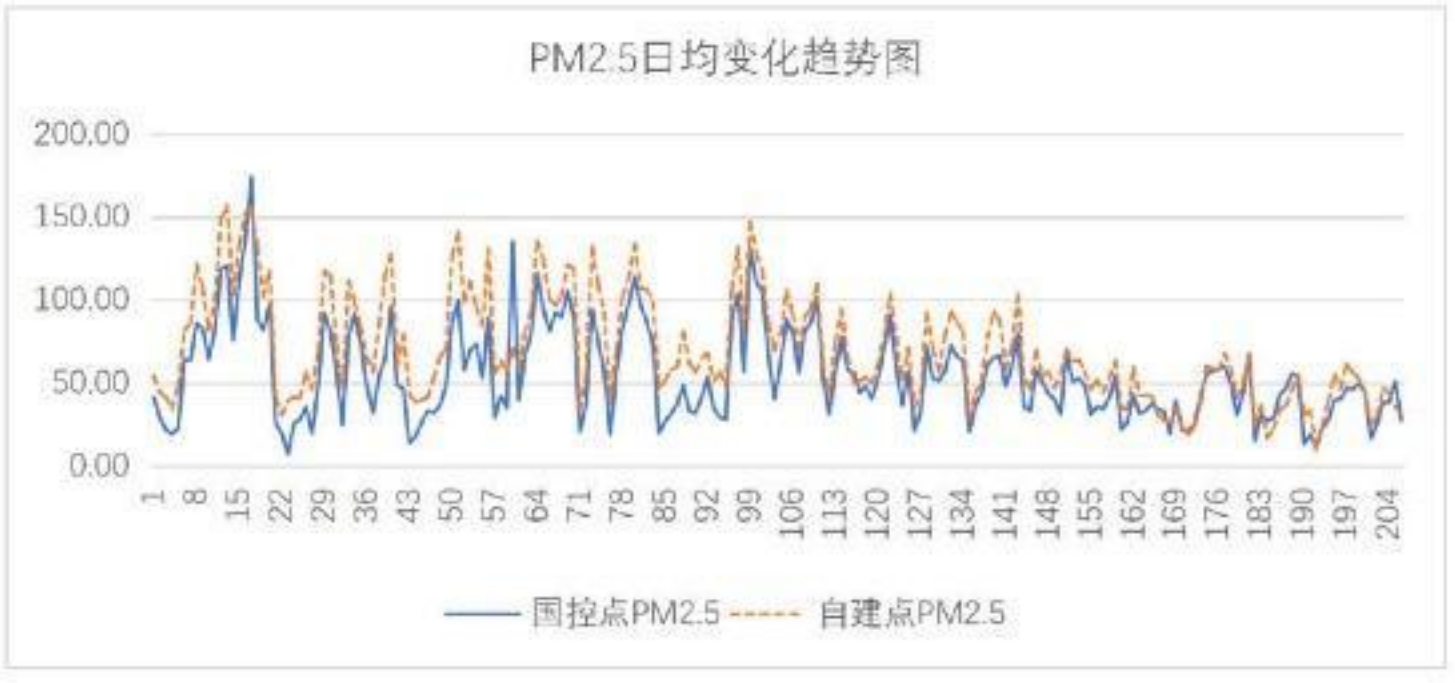


图 1　PM2.5 日均变化趋势图

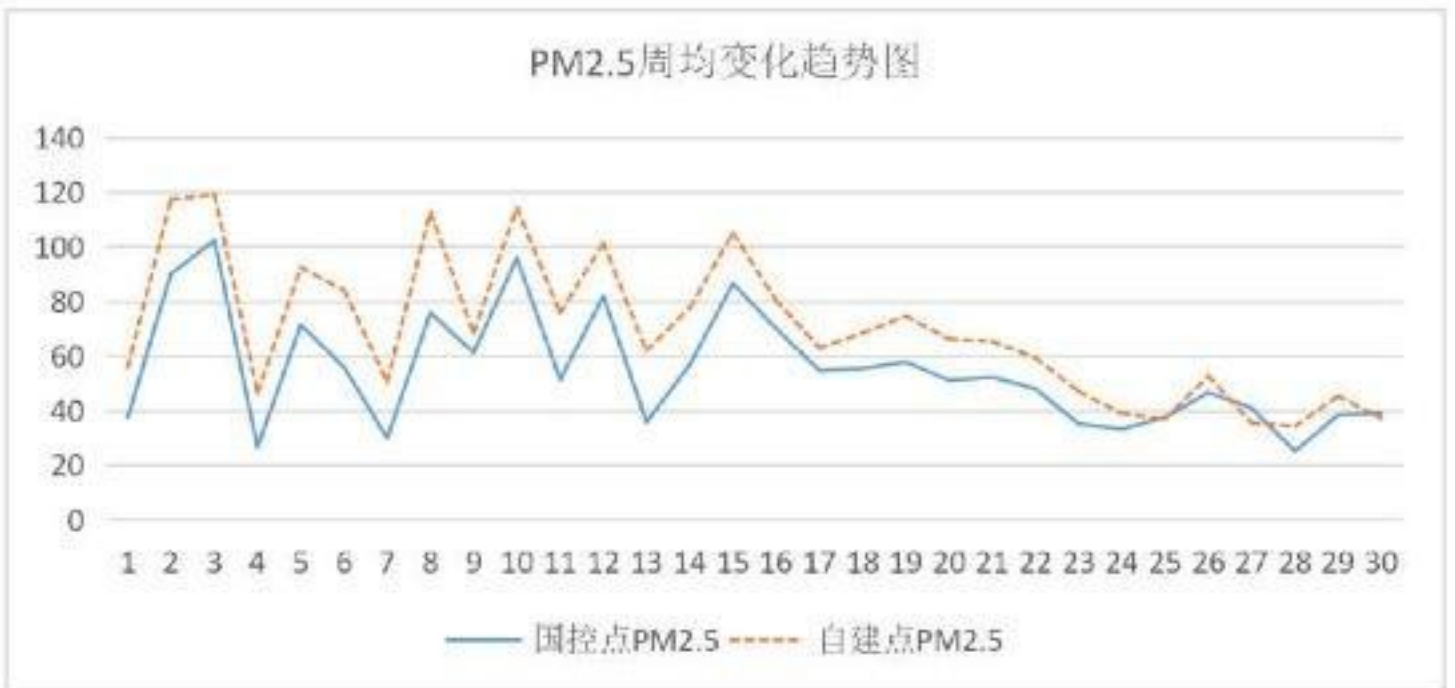


图 2　PM2.5 周均变化趋势图

通过把国控点与自建点的 PM2.5 浓度数据的日平均值和周平均值结合一起分析，并利用 MATLAB 画出对应的折线图（如图 1），我们得到了 PM2.5 日均变化趋势图（图 1）与 PM2.5 周均变化趋势图（图 2）。两个图中，红色虚线为自建点图，蓝色实线为国控点图。

分析图 1 与图 2，不难看出自建点测出的 PM2.5 数值与国控点的变化趋势整体一致，但浓度整体偏高（此处以国控检测值作为准确值）。冬季相比于夏季 PM2.5 的浓度高，2、3 月份浓度有所上升，应与春节放鞭炮有关。

结论一：自建点能较准确的测量出 PM2.5 浓度变化趋势，但测出的 PM2.5 浓度偏高。

结论二：PM2.5 有很大的季节性变化，表现为冬季浓度相对高，夏季浓度相对低。

5.1.2 PM10 数据分析

对比分析方法如 5.1.1 所述，由图易得，PM10 的变化趋势跟 PM2.5 整体相同，符合 PM2.5 的变化特征。如下两图，上面为 PM10 浓度变化日平均趋势图，下面为周平均趋势图，红色虚线为自建点图，蓝色实线为国控点图。

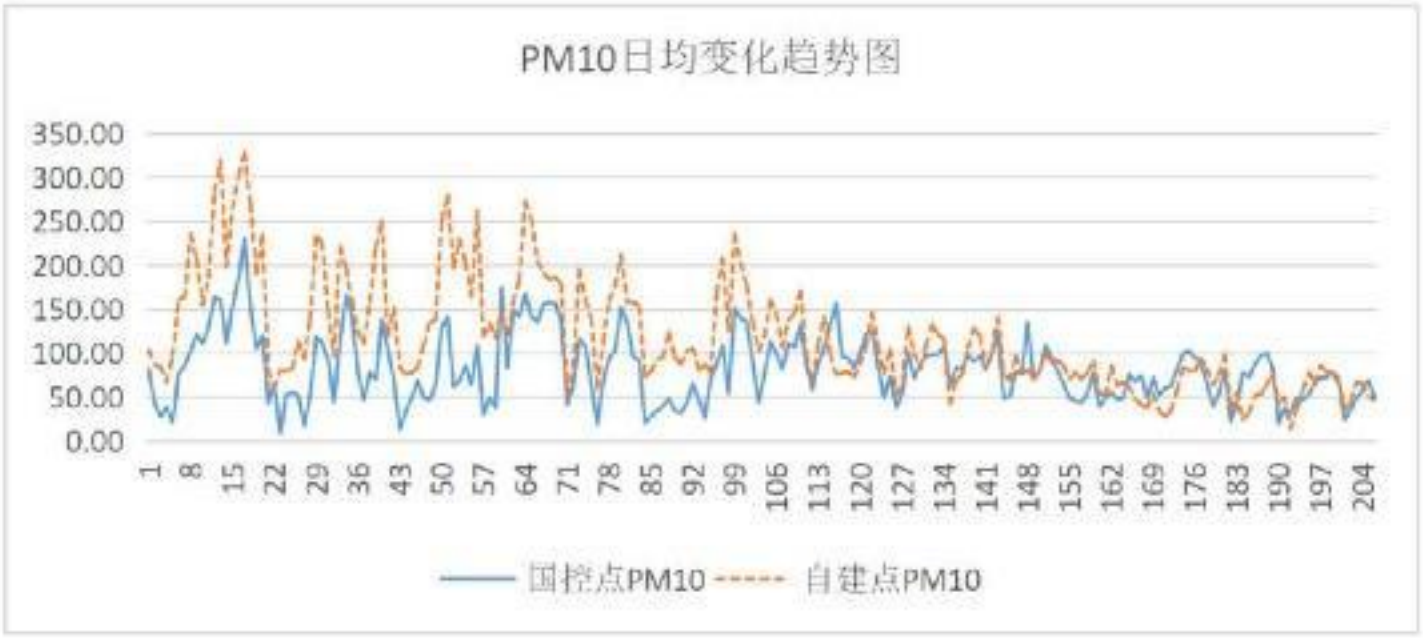


图 3　PM10 日均变化趋势图

图 4　PM10 周均变化趋势图

结论三：PM10 浓度在冬季较高，夏季有所下降。

结论四：自建点的测量值总体准确，冬季符合国控点的变化趋势，夏季对比国控点差异小。

5.1.3　CO 数据分析

图表形象地表达出 CO 浓度也有较大的季节性特征，表现为冬季浓度起伏大，夏季浓度持续较高。自建点测量的 CO 浓度整体变化趋势跟国控点变化趋势大体相同，但仍有很大误差。一月份到四月份自建点对 CO 的检测明显有错，5 月份到 6 月份自建点测量值较接近国控点。可以认为天气对自建点测量 CO 有很大影响。从国控点（蓝色虚线）可以看出 CO 的浓度整体上也是冬季高，夏季相对低。如下两图，上面为 CO 浓度变化日平均趋势图，下面为周平均趋势图，红色虚线为自建点图，蓝色实线为国控点图。

图 5　CO 日均变化趋势图

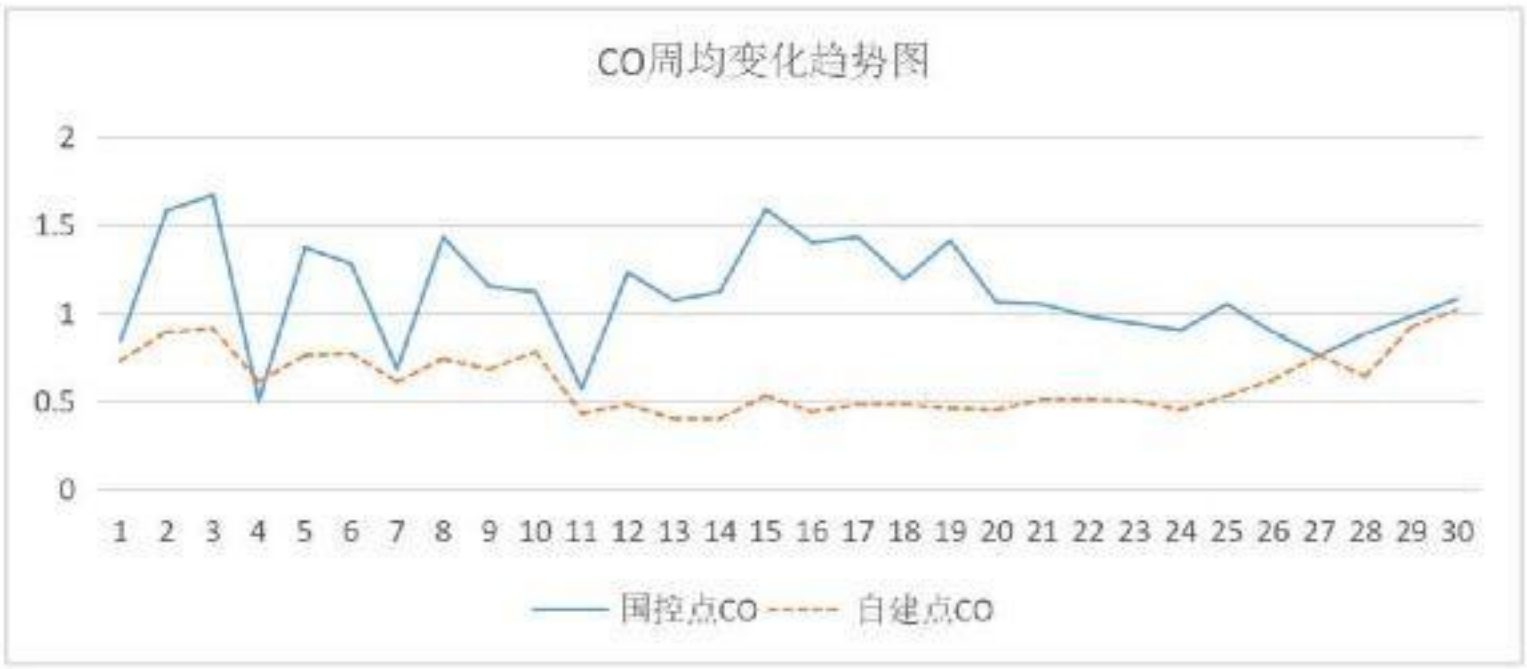


图 6　CO 周均变化趋势图

结论五：CO 浓度在冬季起伏大，在夏季保持高浓度并逐渐降低。

结论六：自建点在 1 月份到 4 月份测量 CO 浓度误差较大。

5.1.4　NO2 数据分析

11 到 12 月份质控点测量的 NO2 浓度明显差别于国控点，1 到 6 月份测量值整体变化趋势与国控点相同。NO2 浓度在 11、12 月份较低，在一月份迅速升高并保持在较高浓度。由于 1 到 6 月份自建点能较准确得测量 NO2 浓度，因此不考虑是地理因数引起的误差。如下两图，上面为 NO2 浓度变化日平均趋势图，下面为周平均趋势图，红色虚线为自建点图，蓝色实线为国控点图。

图 7　NO2 日均变化趋势图

图 8　NO2 周均变化趋势图

结论七：NO2 浓度受温度影响大。

结论八：自建点在温度较低时无法准确测量 NO2 浓度，误差较大。

5.1.5　SO2 数据分析

如下图所示，在 11 月 24 号这一天自建点检测的 SO2 浓度异常高，显然不符合国控点的变化规律，推测可能是传感器当天有问题，或者测试点有人放鞭炮。红色虚线为自建点图，蓝色实线为国控点图。

图 9　SO2 日均变化趋势图

为不影响分析，需要把该天异常数据进行修正，修正方法是把同一天的国控点测得的数据代替自建点的数据，这样既符合整体规律，又能让往后的分析与预测更接近真实情况。根据修正国控点图形得出 SO2 浓度含量在冬天较高并且在一月份浓度迅速降低，夏季保持在低浓度水平，自建点在冬季对 SO2 测量有较大误差。修正后，如下两图，上面为 SO2 浓度变化日平均趋势图，下面为周平均趋势图，红色虚线为自建点图，蓝色实线为国控点图。

图 9　修正后 SO2 日均变化趋势图

图 10　修正后 SO2 周均变化趋势图

结论九：SO2 在 11、12 月浓度较高，并且受温度影响大。

结论十：自建点在冬季无法正确测量 SO2 浓度，整体误差非常大。

5.1.6　O3 数据分析

O3 在 11 到 12 月份浓度较低，从冬季到夏季含量递增。在冬季自建点不能准确测量出 O3 的浓度。如下两图，上面为 O3 浓度变化日平均趋势图，下面为日平均趋势图，红色虚线为自建点图，蓝色实线为国控点图。

图 11　O3 日均变化趋势图

图 12　O3 周均变化趋势图

结论十一：O3 浓度跟温度正相关。

结论十二：温度较低时自建点不能准确测量 O3 浓度，整体误差非常大。

5.2 问题二：对导致自建点数据与国控点数据造成差异的因素进行分析

首先把国控点检测的数据和自建点检测的数据统一进行归一化处理（归一化程序详见附录），通过 Pearson 相关系数计算公式，利用 MATLAB 计算得到如下相关度表 A（计算程序详见附录），表中数据为相关系数 r 的值。

	国控点 PM2.5	国控点 PM10	国控点 CO	国控点 NO2	国控点 SO2	国控点 O3
自建点 PM2.5	**0.9205**	0.7614	0.6678	0.1595	0.3622	-0.4628
自建点 PM10	0.866	**0.7243**	0.6294	0.034	0.4882	-0.5695
自建点 CO	0.3293	0.3639	**0.3355**	-0.1884	0.5824	-0.0449
自建点 NO2	0.3665	0.4272	0.4431	**0.2384**	0.4204	-0.4045
自建点 SO2	0.4587	0.3858	0.4361	-0.049	**0.3935**	-0.2696
自建点 O3	0.15	0.2613	0.197	-0.172	0.4671	**0.2509**
风速	-0.2477	-0.2077	-0.3037	-0.1785	-0.2265	0.1471
压强	0.1595	0.0987	-0.0329	-0.0946	0.2414	-0.5686
降水量	-0.0732	-0.0935	0.093	-0.166	0.2902	-0.111
温度	-0.2084	-0.0688	-0.0589	0.066	-0.1384	0.7046
湿度	0.1167	-0.1817	0.1591	-0.3558	0.1293	-0.5493

表 1 国控点数据与自建点数据的相关性分析表

	自建点 PM2.5	自建点 PM10	自建点 CO	自建点 NO2	自建点 SO2	自建点 O3
风速	-0.2266	-0.2843	-0.3906	-0.1715	-0.4697	-0.2675
压强	0.4919	0.5745	-0.1705	0.2893	0.3206	-0.2247
降水量	0.0516	0.1508	0.3464	0.5341	-0.1566	0.5567
温度	-0.5744	-0.6063	0.2654	-0.1986	-0.3546	0.3636
湿度	0.4793	0.5416	0.0772	0.0554	0.5215	-0.1272

表 2 自建点数据与天气因数的相关性分析表

由表 1 得，自建点与国控点的 PM2.5 的相关系数大于 0.7，为高度线性相关，故自建点数据较接近于国控点，忽略受天气的影响。自建点的 CO、NO2、SO2、O3 与国控点相对应的污染物的相关系数均小于 0.4，即相关度较低，说明其测试误差相当大，此时要考虑风速、压强、降水量、温度、湿度的影响。

由表 2 可知，跟自建点 CO 相关度较高的天气因素有风速、降水量、温度；跟自建点 NO2 相关度较高的因素有压强和降水量；除降水量外其余因素都跟自建点 SO2 相关度较高；除湿度外，O3 对其余因素相关度较高。

把跟 CO 相关度较高的因素数值作归一化周均图，并把国控点与自建点 CO 差值放同一图分析，如图：

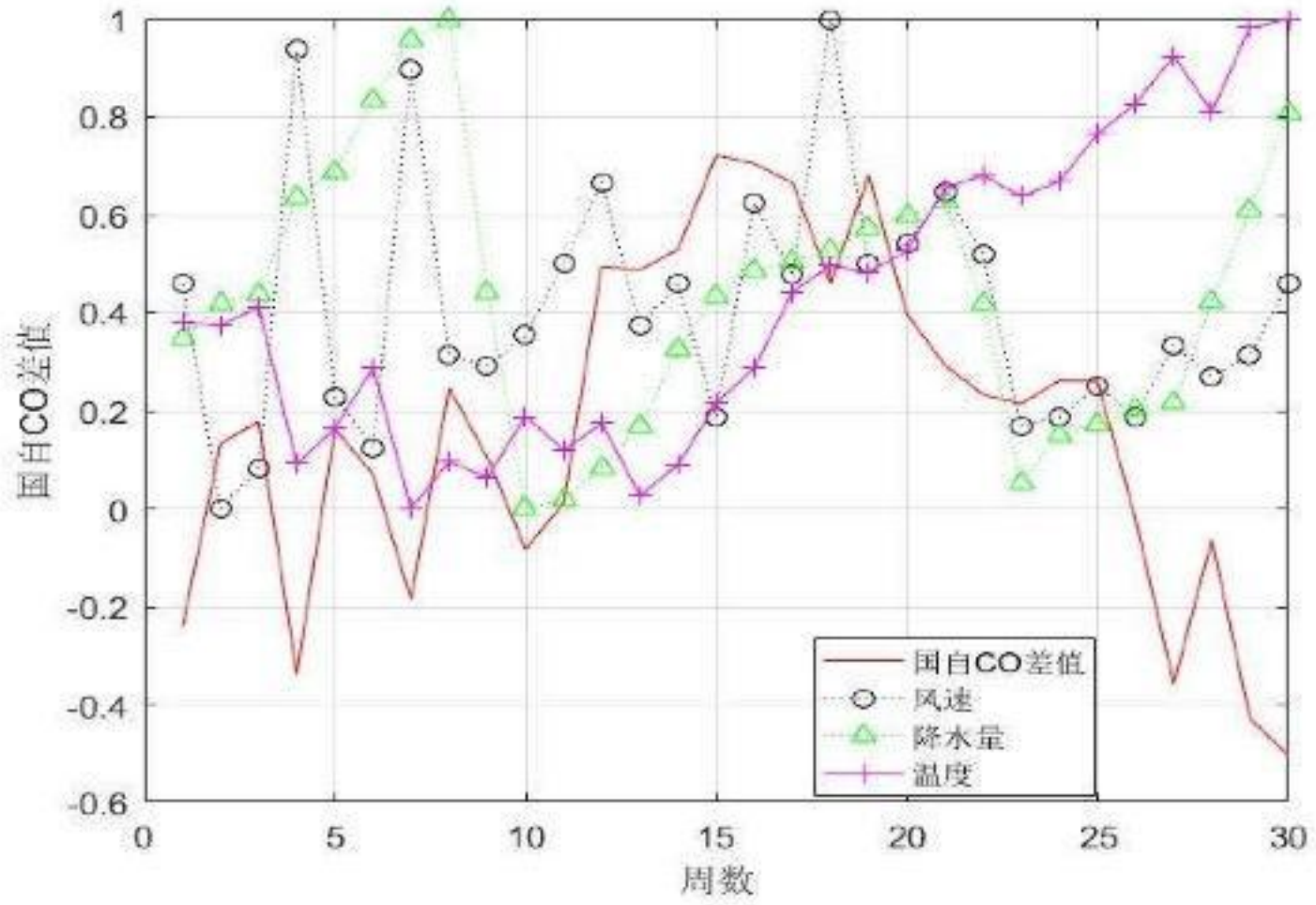


图 13 国自 CO 差值与天气因素变化趋势图

由图 13 可知，风速较大时国控点与自建点 CO 差异较大，降水量增加时，国控点与自建点 CO 差异增大，温度降低时国控点与自建点 CO 差异变大，在 25 到 30 周，风速、降水量、温度都在上升，而国控点与自建点 CO 差异随之变大。从第 1 周到第 8 周降水量一直上升，而国自 CO 差值并不随其变化，相比之下，风速跟温度对差值影响较为敏感

结论:风速与温度对测量差异的影响相比于降水量更敏感更大。

把跟 NO2 相关性较大的因素数值作归一化周均图，把国控点与自建点的 NO2 差值放同一图分析，如图 14：

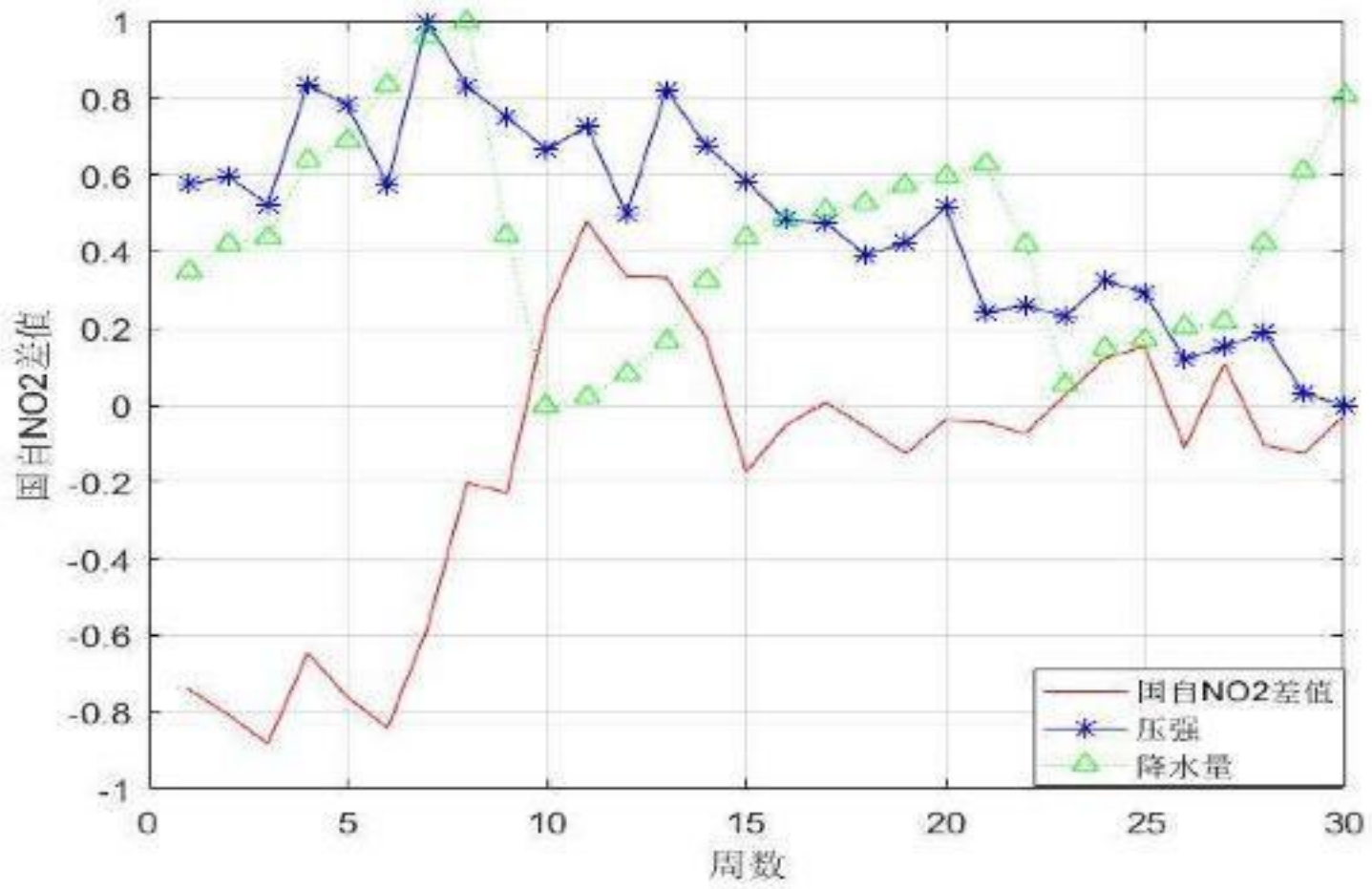


图 14 国自 NO2 差值与天气因素变化趋势图

如果把压强和降水量单独跟国自 NO2 差值比较，很难看出其规律，但是把压强和降水量当成一个整体就会发现，当压强和降水量都都较高时，国自 NO2 的差值较大，当压强和降水量都整体较低时，国自 NO2 的差值变小。

结论：压强与降水量两个天气因数叠加影响时会产生截然不同的结果

把跟 SO2 相关性较大的因素数值作归一化周均图，把国控点与自建点的 SO2 差值放同一图分析，如图 15：

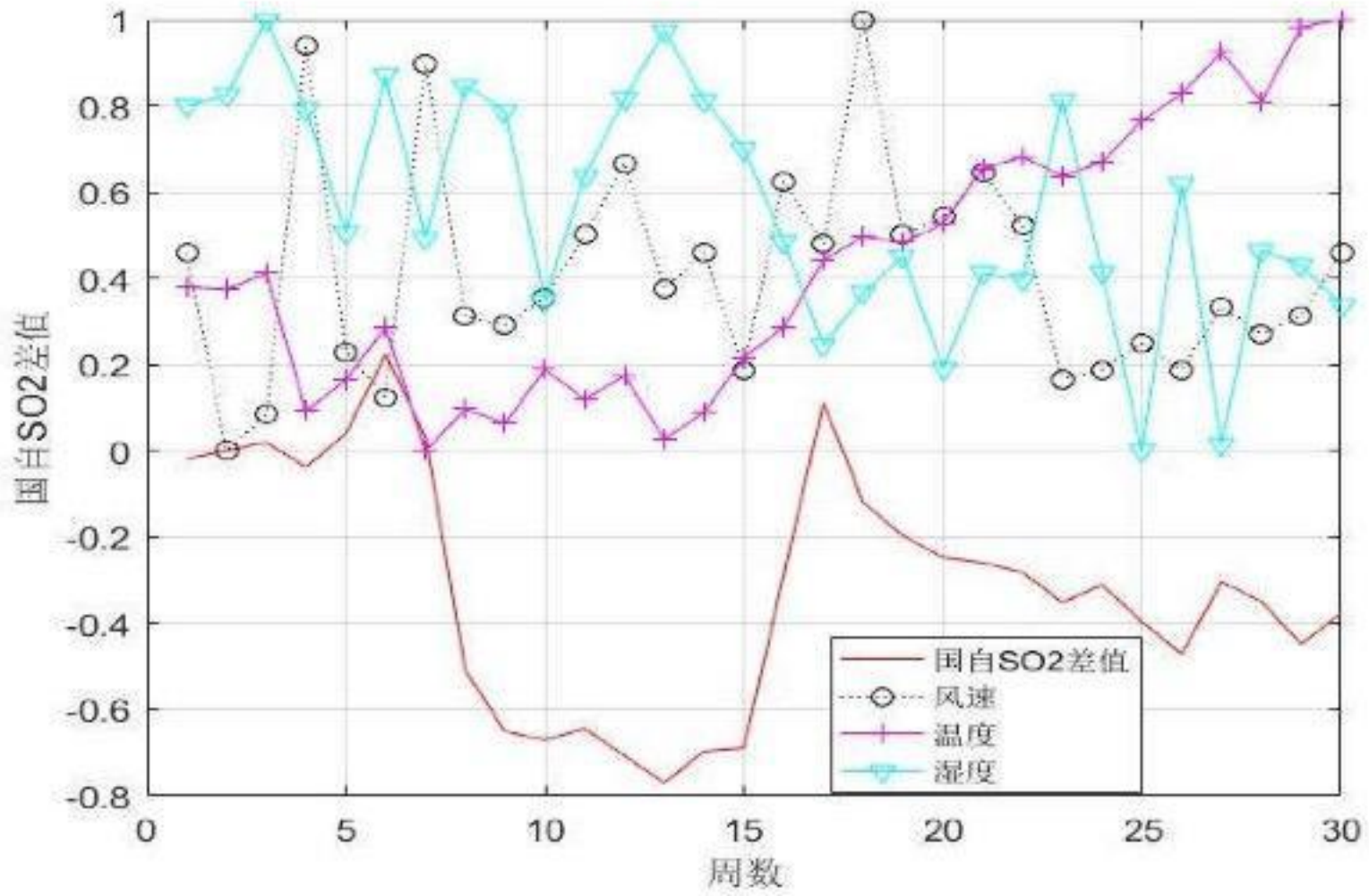


图 15　国自 SO2 差值与天气因素变化趋势图

从 1 到 5 周 SO2 的差值接近 0，而天气因数都有较大起伏，推测是由于各因素的不同影响综合起来就被互相抵消了，第 8 周左右，由于风速、湿度的下降，造成国自 SO2 差值迅速增大，第 15 周后风速上升，差值减小，16 周左右由于湿度、风速下降，造成差值增大。

结论：不同因数综合影响会被抵消，可能会被叠加，使得有时相关性较大的因数对差值影响不大，而相关性较小的因数对差值影响较大。

把跟 SO2 相关性较大的因素数值作归一化周均图，把国控点与自建点的 SO2 差值放同一图分析，如图 16：

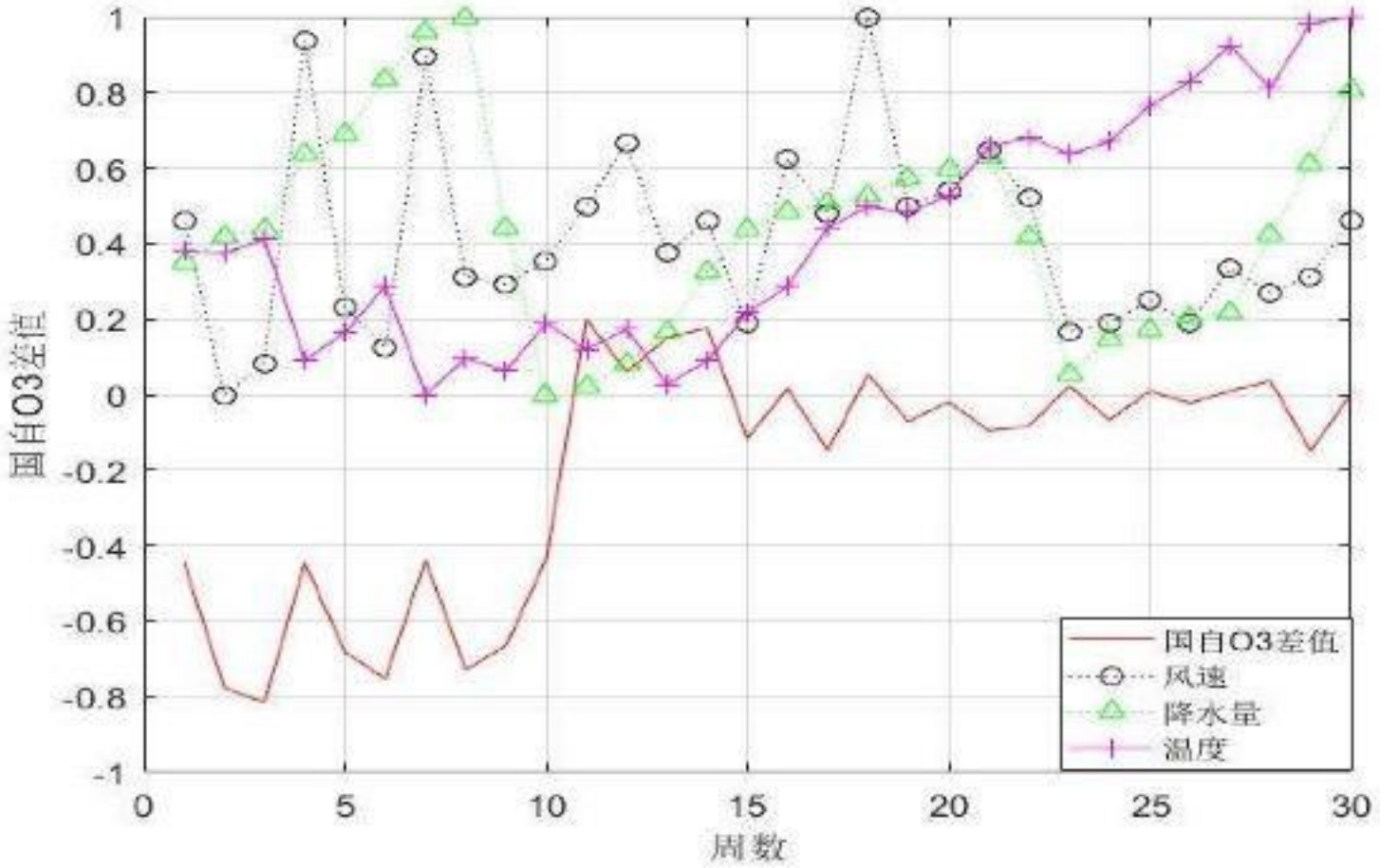


图 16　国自 03 差值与天气因素变化趋势图

在第 1 到第 10 周，气温较低与降水量较高时国自 03 差值较大，在第 10 周到第 20 周，温度降水量风速上升，但 03 差值并不波动，即当因温度降水量风速同时上升时反而对 03 差值影响不大。

结论：当几种相关性较大的因素一起作用时，有时反而影响变小。

最后，因为自建点的 pm2.5 和 pm10 的测量值与国控点高度吻合，相关度高，所以认为自建点的这两个数据是较准确地，因此就不做误差分析了。

5.3 问题三：利用国控点数据，建立数学模型对自建点数据进行校准。

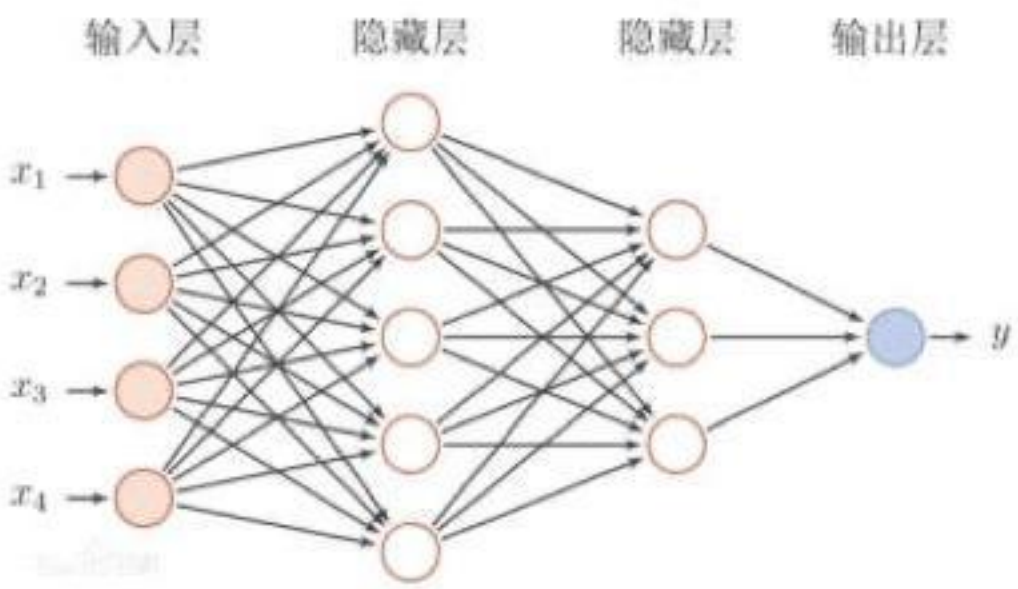
问题三属于回归分析问题，即如何通过自建点的测试数据预测出真实的数据。由表 1 可知国控点数据与自建点的 11 个测量数据（自建点 PM2.5、自建点 PM10、自建点 CO、自建点 NO、自建点 SO2、自建点 03、风速、压强、降水量、温度、湿度）都存在相关性，只是有的相关性强，有的相关性弱。因此可设自建点 11 个测量因子为自变量，真实值（国控点测量值）为因变量，则真实值（国控点测量值）有函数如下：

$$y_i = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_{11} x_{11} + \varepsilon \qquad (2)$$

对上述函数，因为自变量多，且相互之间的作用机制不明确（比如可能存在交叉干扰项），采用传统的方法建立模型对自建点数据进行修正难度较大，于是

我们考虑用 BP 神经网络来对目标函数进行模拟，即把自建点数据作为输入，把国控点某个污染物的数据作为输出，最终得到训练好的 bp 神经网络，就可以通过输入自建点数据预测出污染物的真实指标了。

BP 神经网络是采用反向传播学习算法和非线性可微转移函数的多层网络典型的反向传播算法是梯度下降算法。一个经过训练的 BP 神经网络能够根据输入给出合适的结果，这个特性使得 BP 神经网络很适合采用输入/目标对进行训练。



上图为 BP 神经网络结构，神经网络由 3 个神经元组成，即输入层、隐层、输出层，其中左边为输入层，输入层由大量的数据组成，中间两层为隐层，右边为输出层，输入层与隐层之间的传递函数一般采用激励函数，隐层与输出层之间的传递函数一般也采用激励函数。多层的神经网络理论上可以拟合出任何函数。

5.3.1 BP 神经网络的训练方法

本文为了让 bp 神经网络能更好的预测出真实的污染物指标，在训练神经网络时，采用了如下的优化手段：

（1）数据处理前首先进行数据归一化处理，因为归一化后数据分布偏差较小，对 BP 神经网络训练效果更好；

（2）然后从 206 天中随机选择 176 天的数据作为训练集，剩余 30 天的数据作为测试集，数据的随机选取对 BP 神经网络训练有好处，因为数据多样性得到了保障；

（3）对某个预测值采用多次预测，然后求平均值的方法，其好处是预测结果更加稳健，偶然性误差大大减少；

（4）由于隐层（L）是影响 BP 神经网络预测的重要参数，而参数的选择直接影响 bp 网络预测的错误率，所以本文实验了不同隐层数目对网络性能的影

响，分别让隐层数从 5 至 16 取不同值，得到不同结构的 BP 神经网络，然后对每一个这种神经网络再通过多次运行求均值（20 次）的方法来对污染物数据预测，并得到不同隐层数下的均方差值（BP 神经网络性能指标，此值越小，BP 神经网络性能越好）。

5.3.2 BP 神经网络的性能指标

均方误差是各个数据偏离真实值的距离平方和的平均数（即预测值和真实值一一对应的差值的平方和的均值，其公式为：

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(O_t - Y_t)^2 \qquad (3)$$

其中， O_t为神经网络的预测输出值，Y_t为真实值，N为样本数。

本文以国控点数据作为真实值，对BP神经网络预测数据和自建点原始数据求均方误差值。MSE_BP为归一化处理后BP神经网络预测值与真实输出值的均方差值，显然此值越小，说明BP网络的预测约准确。MSE_Z为归一化处理后自建点数据值与真实输出值的均方差值，此值越大，则表示自建点的测试数据与国控点差异越大。

定义比例系数　　BZ=MSE_BP/MSE_Z　　（4）

系数 BZ 为衡量 BP 神经网络性能指标，BZ 值越小，BP 神经网络性能越好。

MSE_BP_J 为解归一化处理后 BP 神经网络预测值与真实输出值的均方差值,MSE_Z_J 为解归一化处理后自建点数据值与真实输出值的均方差值相应的，定义比例系数　　BZ_J=MSE_BP_J /MSE_Z_J　　（5）

为了更加清晰的表示出 BP 神经网络的预测效果，本文定义了 BP 网络预测准确度 *P*_bp：

$$P\_bp = 1 - \frac{\sum_t |O_t - Y_t|}{\sum_t |Y_t|} \qquad (6)$$

以及自建点准确度 P_Z:

$$P\_Z = 1 - \frac{\sum_t |Z_t - Y_t|}{\sum_t |Y_t|} \qquad (7)$$

即，用 1 减去多次预测的误差与真实值之比，*P*_bp 值与 *P_Z* 值在 0-1 之间，越接近 1 说明 BP 网络预测的准确度越高,越接近 0 说明预测的数值与真实值差异越大。Y_t为真实值，O_t为神经网络的预测输出值，Z_t 为自建点测试值 。

5.3.3　BP 神经网络的结构的确定

根据上述的算法设计，我们利用 matlab2018 自带的神经网络工具箱，从 206

天中随机选择 176 天的数据作为训练集，来训练我们所需的神经网络。在训练时，我们把自建点的每一天的 11 个测试量：自建点 PM2.5、自建点 PM10、自建点 CO、自建点 NO、自建点 SO2、自建点 O3、风速、压强、降水量、温度、湿度，作为网络的输入向量，把国控点的各个测试值分别当成不同神经网络的输出值。

如，其中的预测 pm2.5 的神经网络结构如下图 ：

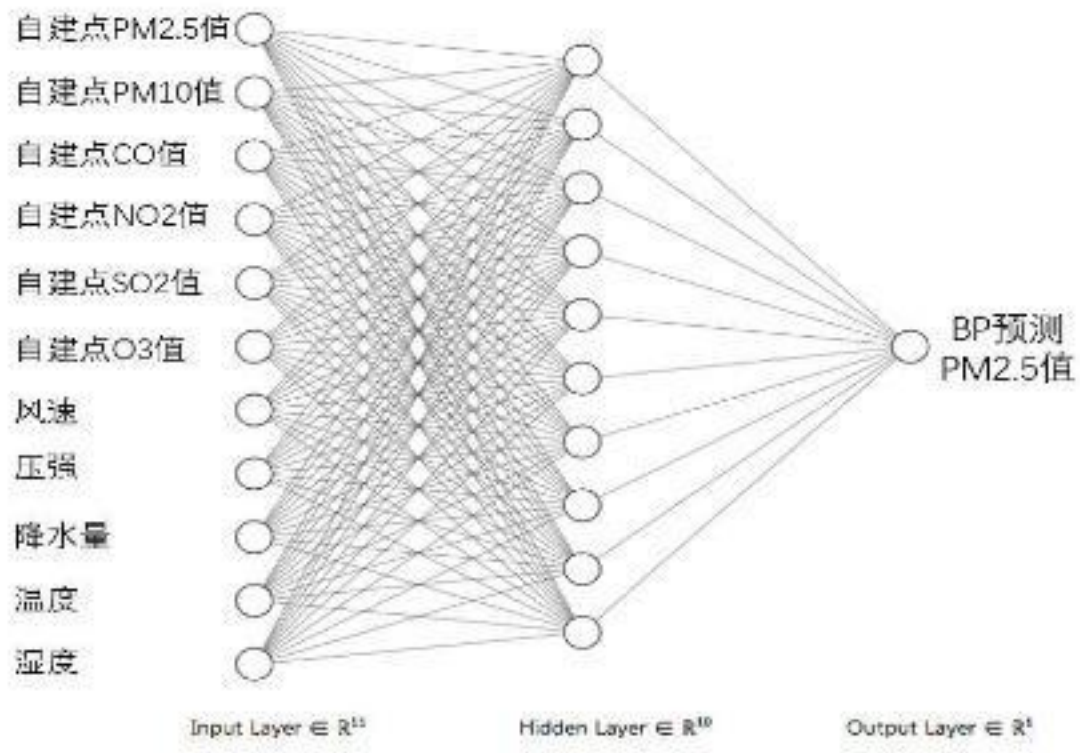


图 17　预测 PM2.5 的 BP 神经网络结构示意图

也就是说对每一个污染物指标，我们都建立一个神经网络，这样我们就需要训练 6 个不同的神经网络。用如上方法训练好 6 个 BP 神经网络之后，用剩余 30 天的数据作为测试集，对我们得到的 BP 神经网络的性能进行测试。

为了得到最优的神经网络结构我们通过程序设计，让 BP 神经网络的隐层数目分别取 5-16，每一层都分别计算这一层的均方差比例系数 BZ，通过此系数来反应 BP 网络的性能。下表为不同隐层数对应点的预测值的均方差比例系数 BZ 值。

隐层数	PM2.5	PM10	CO	NO2	SO2	O3
5	0.0912	0.4727	0.2504	0.2903	0.4357	0.0933
6	0.1056	0.4016	0.1026	0.3554	0.2092	0.1198
7	0.116	0.0986	0.1973	0.2662	0.3887	0.1088
8	0.0989	0.2263	0.2132	0.3939	0.4652	0.0763
9	0.1666	0.2683	0.1601	0.3774	0.3452	0.0877
10	0.0967	0.3249	0.2274	0.3831	0.3764	0.084
11	0.0909	0.2319	0.1241	0.1556	0.6272	0.1844
12	0.2646	0.3363	0.1478	0.2298	0.2887	0.0575
13	0.1482	0.2169	0.1257	0.2324	0.3894	0.1031
14	0.0926	0.2045	0.1269	0.2209	0.4087	0.1369
15	0.0969	0.1901	0.1137	0.3439	0.4939	0.192
16	0.1302	0.3702	0.1946	0.2424	0.3157	0.1611

表 3　不同隐层数的比例系数 BZ 值

例如，由表 3 数据可知，对 PM2.5 而言，比例系数 BZ 值最小时对应的隐层数为 11，此时 BP 神经网络预测性能最好。

同理得出各污染物质指标预测所需的 BP 神经网络的最佳隐层数，如下表：

	PM2.5	PM10	CO	NO2	SO2	O3
最优隐层数	11	7	6	11	6	12

表 4　不同污染空气的最优隐层数

5.3.4　BP 神经网络模型的对预测性能分析

根据上面的实验，我们分别得出了 PM2.5、PM10、CO、NO2、SO2、O3 这 6 个 污染指标的最佳 BP 神经网络结构（隐层数目）。下面开始对 BP 网络的性能进行测试，测试的方法是，对某一个污染指标来说，比如 PM2.5,我们把 30 个随机选择的自建点的 11 个数据作为测试集样本然后输入训练好的 BP 神经网络，进而得到 PM2.5 的 BP 神经网络预测值，然后我们把此值和自建点原始数据值以及国控点测量值三者进行比较，做出下列的对比图。

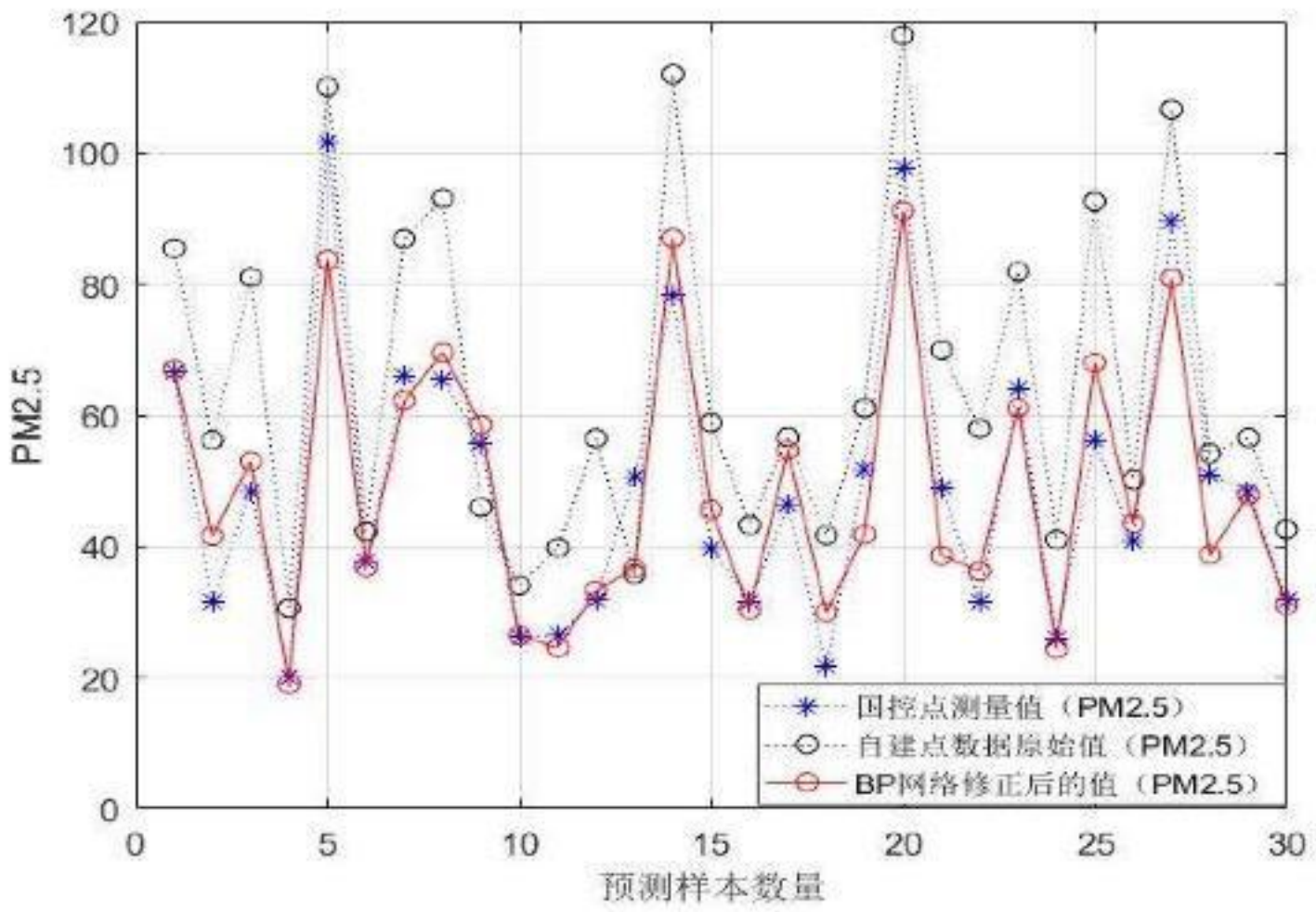


图 18　PM2.5 预测效果对比图

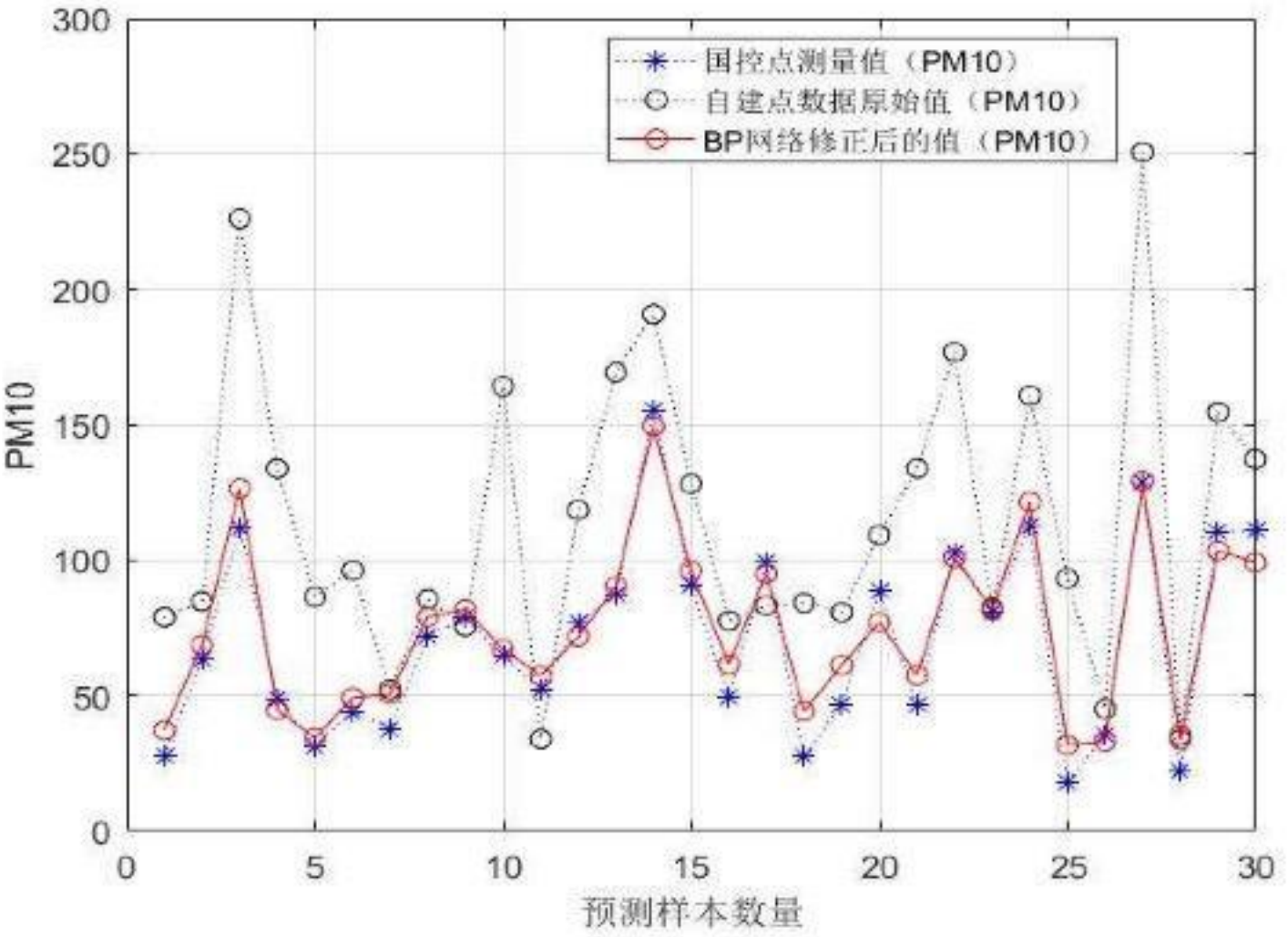


图 19　PM10 预测效果对比图

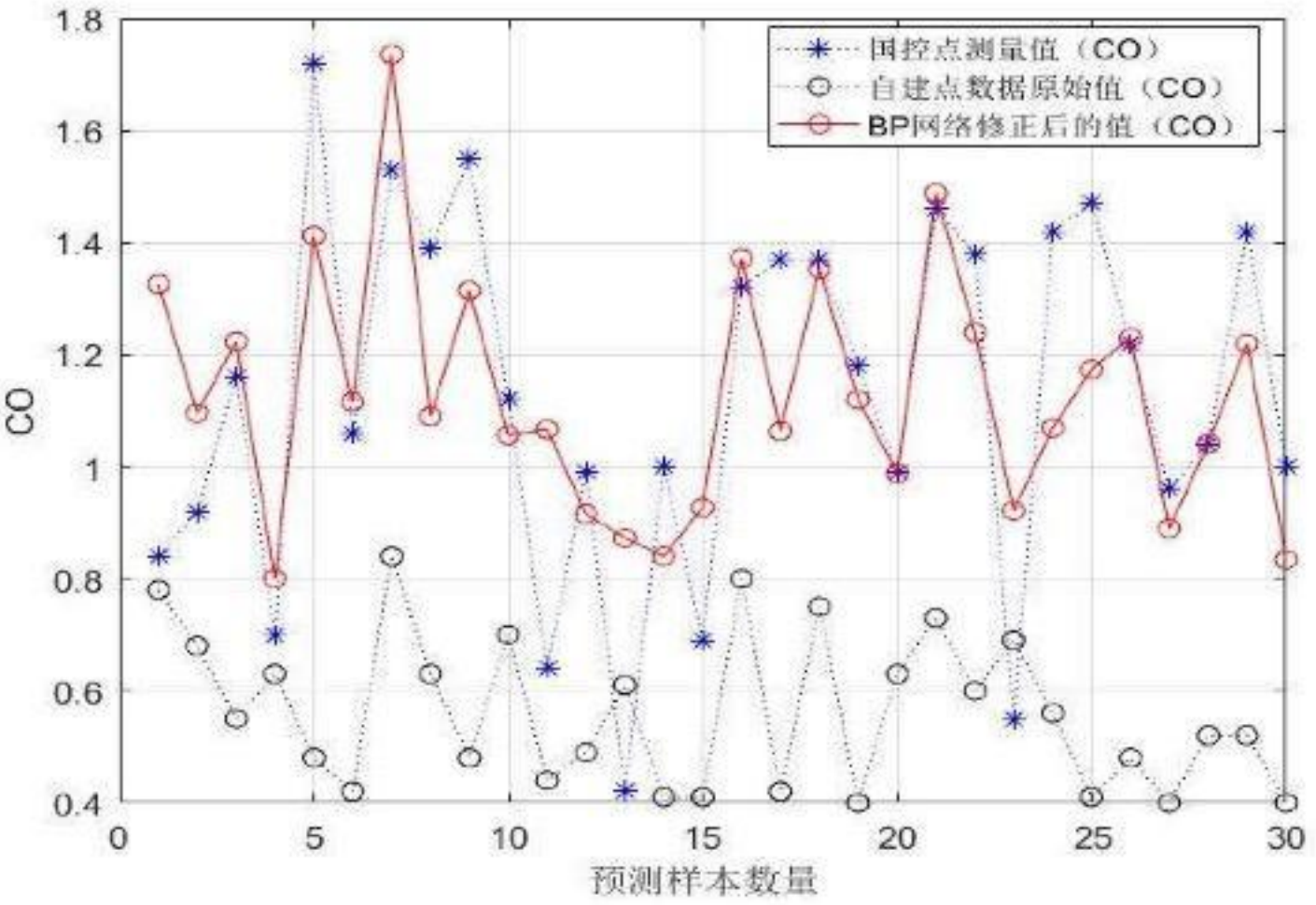


图 20　CO 预测效果对比图

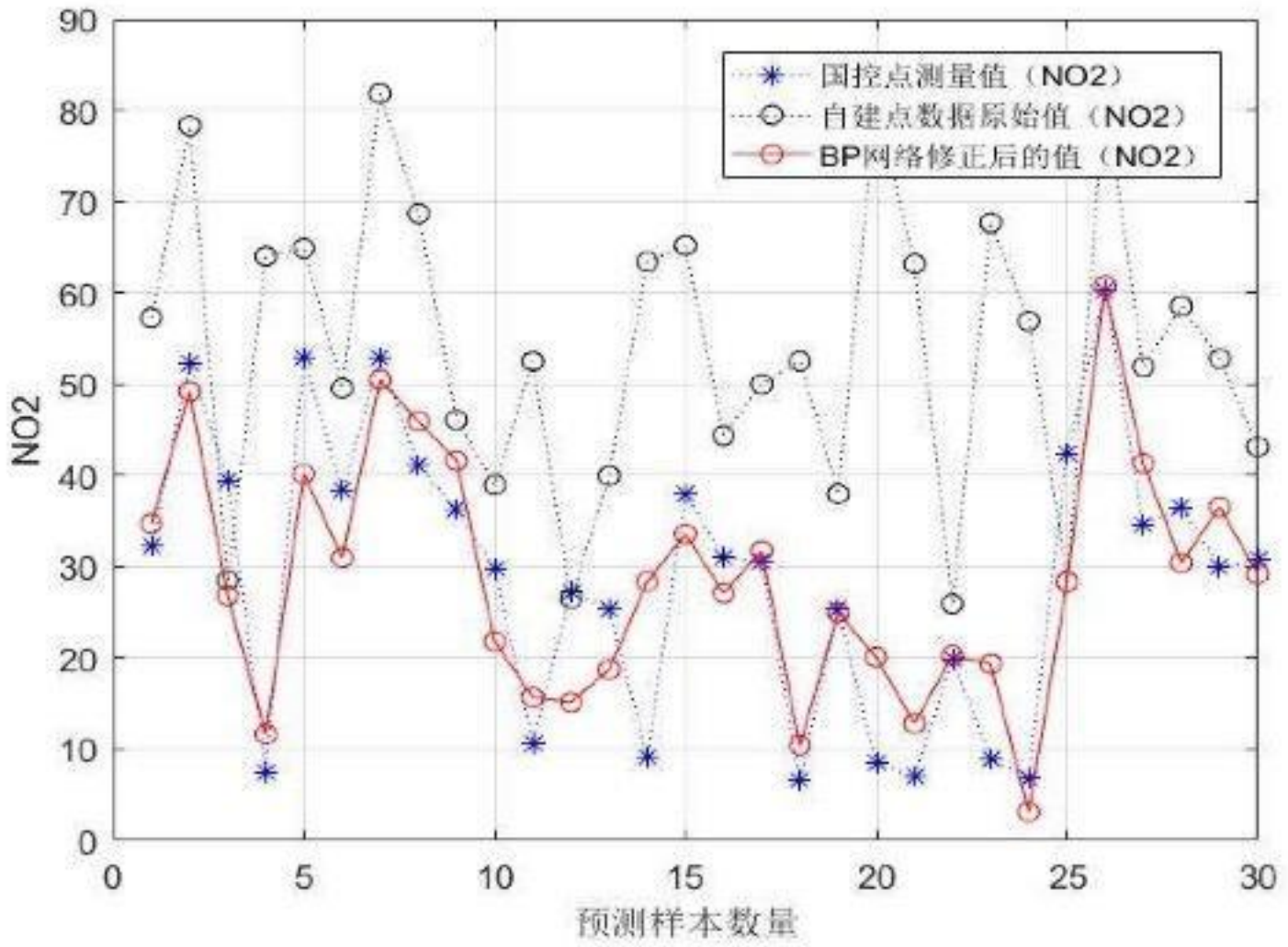


图 21　NO2 预测效果对比图

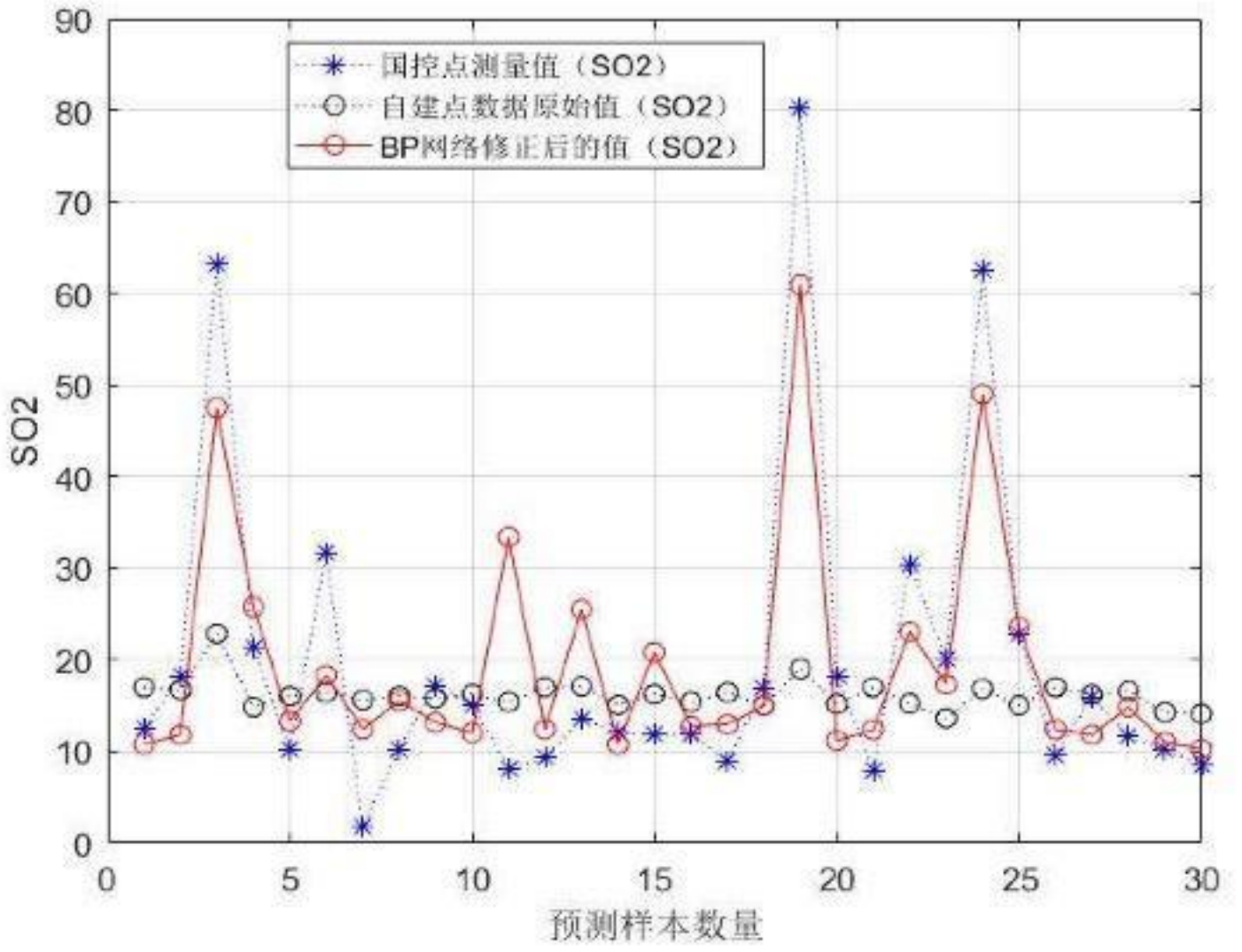


图 22　SO2 预测效果对比图

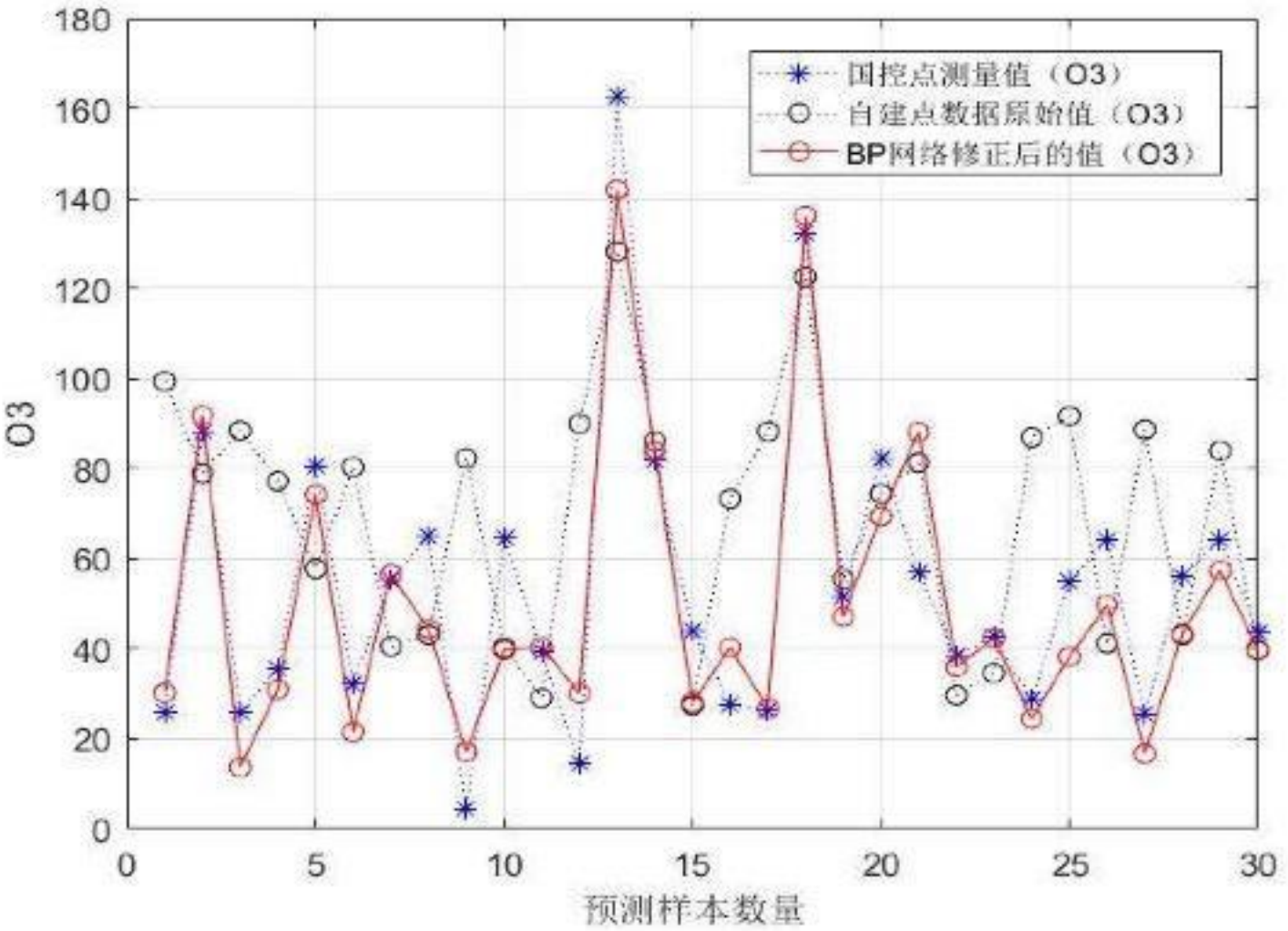


图 23　O3 预测效果对比图

由图 20 至图 25 各自对照可知：BP 神经网络修正后的污染物指标预测值远远优于自建点原始数据值，BP 神经网络的预测值更接近与真实值（即国控点测量值），说明我们的模型非常成功，可以非常好的对自建点数据进行校准。

为了更准确比较本文的 BP 模型对预测性能的提升效果，本文分别求出了神经网络预测的 PM2.5、PM10、CO、NO2、SO2、O3 这 6 个污染指标与国测数值的均方差值 MSE_BP 以及均方差比例系数 BZ，以及自测点这六个指标的数值与国测数值均方差值 MSE_Z 以及均方差比例系数 BZ

归一化后的数据我们记为 MSE_BP、MSW_Z、BZ，数据解归一化后的数据我们记为 MSE_BP1、MSE_Z1、BZ_J。

自建点的准确率记为 P_Z，BP 网络预测的准确率记为 P_bp。

		PM2.5	PM10	CO	NO2	SO2	O3
均方差（归一化）	MSE_BP	0.0019	0.0015	0.0098	0.0098	0.0118	0.0051
	MSE_Z	0.0209	0.0153	0.0951	0.0627	0.0565	0.0886
	BZ	0.0909	0.0986	0.1026	0.1556	0.2092	0.0575
均方差（解归一化）	MSE_BP_J	52.6175	74.4271	0.0529	59.4366	76.2294	154.2694
	MSE_Z_J	360.5526	3185.2	0.4378	1053.5	293.5084	1508.1
	BZ_J	0.1459	0.0234	0.1209	0.0564	0.2597	0.1023

自建点准确率 P_Z	48.22%	54.06%	67.21%	34.97%	0%	0%
BP 网络预测准确率 P_bp	90.35%	89.37%	85.76%	75.85%	71.53%	84.44%

表 5　BP 神经网络预测性能分析表

通过表 5，发现通过训练好的 BP 网络模型，污染物指标的预测能力得到的极大的提升。如 PM2.5 的准确度由自建点的 48.22%提升到 90.35%，PM10 的准确度由自建点的 54.06%提升到 89.37%，SO2 的准确度由 0%提升到 71.53%，可见我们的模型相当成功。

六、模型的评价与改进

本文假假设国控点数据为真实数据，通过计算自建点数据与理想数据的相关系数分析了自建点与国控点数据的内涵。通过计算国测点 6 个指标与自测点 11 个指标之间的相关系数矩阵，分析了每一个污染物指标的自测点与国控点之间的相关性。通过优化的 BP 神经网络来对自建点数据进行了修正和预测，得出了国控值、BP 预测值、自建点测量值三者之间的对比关系图，发现模型对污染物指标的预测能力得到的极大的提升。由于训练 BP 神经网络的数据是经过日均值处理的，所以数据不能最精确的表达出真实情况，因此 BP 神经网络预测出的数据仍然有少许的误差。进一步改进则需要用更多更密集的数据对 BP 神经网络训练和改进。

七、参考文献

1、韩中庚，宋明武，邵广纪，数学建模竞赛一获奖论文精选与点评

2、王连堂，《数学建模》

3、冯杰，数学建模原理与案例，北京：科学出版社，2007。

4、韩中庚，数学建模方法及其应用,北京：高等教育出版社，2004。

5、百张怡文，胡静宜，王冉．基于神经网络的 PM_(2.5)预测模型[J]．江苏师范大学学报(自然科学版)，2015(1):63-65.

6、https://blog.csdn.net/weixin_40909201/article/details/82113053

7、田静毅[1]，范泽宣[2]，孙丽华[3]．基于 BP 神经网络的空气质量预测与分析[J]．辽宁科技大学学报，2015(2):131-136.

附录

pm25 不同隐层数目试验

```
clc
clear all
load data.mat   %载入 206 行 17 列的总数据

%记录归一化的 MSE_bp,MSE_Z,bili 和解归一化的 MSE_bp1,MSE_Z1,bili1
%初始值置零
JFC=zeros(6,12);

for yc=5:16       %试验不同 隐层 数目对结构的影响
    yc                %显示层数

%初始化各个参数相量（数据归一化到 0-1 之间）
[G_pm25_d_N,ps_G_pm25_d_N]=mapminmax(data(:,1)',0,1);
[G_pm10_d_N,ps_G_pm10_d_N]=mapminmax(data(:,2)',0,1);
[G_CO_d_N,ps_G_CO_d_N]=mapminmax(data(:,3)',0,1);
[G_NO2_d_N,ps_G_NO2_d_N]=mapminmax(data(:,4)',0,1);
[G_SO2_d_N,ps_G_SO2_d_N]=mapminmax(data(:,5)',0,1);
[G_O3_d_N,ps_G_O3_d_N]=mapminmax(data(:,6)',0,1);

[Z_pm25_d_N,ps_Z_pm25_d_N]=mapminmax(data(:,7)',0,1);
[Z_pm10_d_N,ps_Z_pm10_d_N]=mapminmax(data(:,8)',0,1);
[Z_CO_d_N,ps_Z_CO_d_N]=mapminmax(data(:,9)',0,1);
[Z_NO2_d_N,ps_Z_NO2_d_N]=mapminmax(data(:,10)',0,1);
[Z_SO2_d_N,ps_Z_SO2_d_N]=mapminmax(data(:,11)',0,1);
[Z_O3_d_N,ps_Z_O3_d_N]=mapminmax(data(:,12)',0,1);
```

```
Feng_d_N=mapminmax(data(:,13)',0,1);
Ya_d_N=mapminmax(data(:,14)',0,1);
Shui_d_N=mapminmax(data(:,15)',0,1);
Wen_d_N=mapminmax(data(:,16)',0,1);
Shi_d_N=mapminmax(data(:,17)',0,1);
```

```
%计算国测点各测量值与自测点的 11 个测量值的相关系数：
%|r|<0.4 为低度线性相关；0.4≤|r|<0.7 为显著性相关；0.7≤|r|<1 为高度线性相关。
%相关系数函数 corr 计算的相量必须是列向量
```

```
r1_1=corr(G_pm25_d_N',Z_pm25_d_N','type','pearson');
r1_2=corr(G_pm25_d_N',Z_pm10_d_N','type','pearson');
r1_3=corr(G_pm25_d_N',Z_CO_d_N','type','pearson');
r1_4=corr(G_pm25_d_N',Z_NO2_d_N','type','pearson');
r1_5=corr(G_pm25_d_N',Z_SO2_d_N','type','pearson');
r1_6=corr(G_pm25_d_N',Z_O3_d_N','type','pearson');
r1_7=corr(G_pm25_d_N',Feng_d_N','type','pearson');
r1_8=corr(G_pm25_d_N',Ya_d_N','type','pearson');
r1_9=corr(G_pm25_d_N',Shui_d_N','type','pearson');
r1_10=corr(G_pm25_d_N',Wen_d_N','type','pearson');
r1_11=corr(G_pm25_d_N',Shi_d_N','type','pearson');

r2_1=corr(G_pm10_d_N',Z_pm25_d_N','type','pearson');
r2_2=corr(G_pm10_d_N',Z_pm10_d_N','type','pearson');
r2_3=corr(G_pm10_d_N',Z_CO_d_N','type','pearson');
r2_4=corr(G_pm10_d_N',Z_NO2_d_N','type','pearson');
r2_5=corr(G_pm10_d_N',Z_SO2_d_N','type','pearson');
r2_6=corr(G_pm10_d_N',Z_O3_d_N','type','pearson');
r2_7=corr(G_pm10_d_N',Feng_d_N','type','pearson');
```

```
r2_8=corr(G_pm10_d_N',Ya_d_N','type','pearson');
r2_9=corr(G_pm10_d_N',Shui_d_N','type','pearson');
r2_10=corr(G_pm10_d_N',Wen_d_N','type','pearson');
r2_11=corr(G_pm10_d_N',Shi_d_N','type','pearson');

r3_1=corr(G_CO_d_N',Z_pm25_d_N','type','pearson');
r3_2=corr(G_CO_d_N',Z_pm10_d_N','type','pearson');
r3_3=corr(G_CO_d_N',Z_CO_d_N','type','pearson');
r3_4=corr(G_CO_d_N',Z_NO2_d_N','type','pearson');
r3_5=corr(G_CO_d_N',Z_SO2_d_N','type','pearson');
r3_6=corr(G_CO_d_N',Z_O3_d_N','type','pearson');
r3_7=corr(G_CO_d_N',Feng_d_N','type','pearson');
r3_8=corr(G_CO_d_N',Ya_d_N','type','pearson');
r3_9=corr(G_CO_d_N',Shui_d_N','type','pearson');
r3_10=corr(G_CO_d_N',Wen_d_N','type','pearson');
r3_11=corr(G_CO_d_N',Shi_d_N','type','pearson');

r4_1=corr(G_NO2_d_N',Z_pm25_d_N','type','pearson');
r4_2=corr(G_NO2_d_N',Z_pm10_d_N','type','pearson');
r4_3=corr(G_NO2_d_N',Z_CO_d_N','type','pearson');
r4_4=corr(G_NO2_d_N',Z_NO2_d_N','type','pearson');
r4_5=corr(G_NO2_d_N',Z_SO2_d_N','type','pearson');
r4_6=corr(G_NO2_d_N',Z_O3_d_N','type','pearson');
r4_7=corr(G_NO2_d_N',Feng_d_N','type','pearson');
r4_8=corr(G_NO2_d_N',Ya_d_N','type','pearson');
r4_9=corr(G_NO2_d_N',Shui_d_N','type','pearson');
r4_10=corr(G_NO2_d_N',Wen_d_N','type','pearson');
r4_11=corr(G_NO2_d_N',Shi_d_N','type','pearson');
```

```
r5_1=corr(G_SO2_d_N',Z_pm25_d_N','type','pearson');
r5_2=corr(G_SO2_d_N',Z_pm10_d_N','type','pearson');
r5_3=corr(G_SO2_d_N',Z_CO_d_N','type','pearson');
r5_4=corr(G_SO2_d_N',Z_NO2_d_N','type','pearson');
r5_5=corr(G_SO2_d_N',Z_SO2_d_N','type','pearson');
r5_6=corr(G_SO2_d_N',Z_O3_d_N','type','pearson');
r5_7=corr(G_SO2_d_N',Feng_d_N','type','pearson');
r5_8=corr(G_SO2_d_N',Ya_d_N','type','pearson');
r5_9=corr(G_SO2_d_N',Shui_d_N','type','pearson');
r5_10=corr(G_SO2_d_N',Wen_d_N','type','pearson');
r5_11=corr(G_SO2_d_N',Shi_d_N','type','pearson');

r6_1=corr(G_O3_d_N',Z_pm25_d_N','type','pearson');
r6_2=corr(G_O3_d_N',Z_pm10_d_N','type','pearson');
r6_3=corr(G_O3_d_N',Z_CO_d_N','type','pearson');
r6_4=corr(G_O3_d_N',Z_NO2_d_N','type','pearson');
r6_5=corr(G_O3_d_N',Z_SO2_d_N','type','pearson');
r6_6=corr(G_O3_d_N',Z_O3_d_N','type','pearson');
r6_7=corr(G_O3_d_N',Feng_d_N','type','pearson');
r6_8=corr(G_O3_d_N',Ya_d_N','type','pearson');
r6_9=corr(G_O3_d_N',Shui_d_N','type','pearson');
r6_10=corr(G_O3_d_N',Wen_d_N','type','pearson');
r6_11=corr(G_O3_d_N',Shi_d_N','type','pearson');


%pm25_BP 网络

% 训练集
a=randperm(206);      %把 1-206，这 206 个整数打乱顺序。
B=a(1:176);           %随机选 176 个训练集
```

```
C=a(177:206);          %剩余的随机 30 个测试集


P_train                                                                    =
    [Z_pm25_d_N(B);Z_pm10_d_N(B);Z_CO_d_N(B);Z_NO2_d_N(B);Z_SO2_d_N(B);Z_O3_d
    _N(B);Feng_d_N(B);Ya_d_N(B);Shui_d_N(B);Wen_d_N(B);Shi_d_N(B)];
T_train = G_pm25_d_N(B);

% 测试集—— 30 个样本，测试点多一些，从图上更好说明我们 Bp 网络的优越性
P_test                                                                     =
    [Z_pm25_d_N(C);Z_pm10_d_N(C);Z_CO_d_N(C);Z_NO2_d_N(C);Z_SO2_d_N(C);Z_O3_d
    _N(C);Feng_d_N(C);Ya_d_N(C);Shui_d_N(C);Wen_d_N(C);Shi_d_N(C)];
T_test = G_pm25_d_N(C);
N = size(P_test,2);

T_sim_bp=zeros(1,N); %预测值的初始化，都设置为 0

for i=1:20              %循环 20 次，把 20 次 BP 网络预测的值求和后取平均值

    % 创建网络
        net = newff(P_train,T_train,yc);    %yc 为隐层的层数
    net.trainFcn = 'traingda';          %traingda 表示自适应调整学习率的梯度
下降反向传播算法训练函数
 % 设置训练参数
    net.trainParam.epochs =600;
    net.trainParam.goal = 1e-5;
    net.trainParam.lr = 0.001;
 % 训练网络
    net = train(net,P_train,T_train);
 %仿真测试
```

```
        T_sim_bp_i= sim(net,P_test);       %T_sim_bp_i 为循环时 T_sim_bp 每次得到
    的临时值
        T_sim_bp=T_sim_bp+T_sim_bp_i;   %把每次循环时 T_sim_bp 的值累加起来;

end

T_sim_bp=T_sim_bp/20;                          %把累加的 T_sim_bp，再求平均值;

%相对误差 error
error_bp = abs(T_sim_bp - T_test)./T_test;

% 结果对比
result_bp = [T_test' T_sim_bp'  error_bp'];     % 测试值，预测值，误差的结果对
    比

PM25_MSE_BP=0;  %BP 网络预测数据的均方误差初始化
PM25_MSE_Z=0;   %自测点原始数据的均方误差初始化

Z_pm25_d_N_L=Z_pm25_d_N(C);  %非常重要，如果换成 pm10,这里需要修改

for i=1:N
    PM25_MSE_BP=PM25_MSE_BP+(T_test(i)-T_sim_bp(i))^2;
    PM25_MSE_Z=PM25_MSE_Z+(T_test(i)-Z_pm25_d_N_L(i))^2;
end

PM25_MSE_BP=PM25_MSE_BP/N     %BP 网络预测数据的均方误差
PM25_MSE_Z=PM25_MSE_Z/N       %自测点原始数据的均方误差
bili=PM25_MSE_BP/PM25_MSE_Z              %此值越小，说明 bp 网络的预测效果越好。

p=1-sum(abs(T_sim_bp- T_test))/sum( T_test)
```

```
JFC(1,yc-4)=PM25_MSE_BP;
JFC(2,yc-4)=PM25_MSE_Z;
JFC(3,yc-4)=bili;
%--------------------------------------------------

%% 绘图(归一化数据)
figure
plot(1:N,T_test,'b:*' , 1:N,Z_pm25_d_N(C),'k:o',1:N,T_sim_bp,'r-o')     %把国
    测的值，自建点原值测量值，BP 网络修正的值，三者一起画出来。
legend('国控点测量值（PM2.5）','自建点数据原始值（PM2.5）','BP 网络修正后的值
    （PM2.5）')
xlabel('预测样本数量')
ylabel('PM2.5')
grid

end
```