基于 PSO-BP 神经网络预测广州市日均 PM_{10} 浓度[1]

南方医科大学　　林愿仪、林伟俊、尹安琪

摘　要

背景：可吸入颗粒物 PM_{10} 是指悬浮在空气中，空气动力学当量直径小于或等于 $10\mu m$ 的颗粒物，其浓度的升高会给人群健康造成很大的危害。对 PM_{10} 浓度进行预测预报可以为环境管理决策提供依据, 同时有助于市民及时采取相应的防控措施降低污染的影响。

目的：拟用 2008 年-2011 年广州市前一天的日均 PM_{10} 浓度结合同期的气象等因素，建立 PSO-BP 神经网络模型，对 PM_{10} 浓度进行预测。

方法：将广州市 2008 年-2011 年的 PM_{10} 浓度和气象资料数据分为训练样本和测试样本。在训练样本中运用逐步回归法筛选出影响 PM_{10} 浓度预测的主要因素，建立多元线性回归模型和 PSO-BP 神经网络模型对 PM_{10} 浓度进行预测。利用测试样本检验两模型的预测效果，并用 RMSE、MAE、MAPE、PMAD、R^2 等指标对两模型的预测效果进行比较，以说明 PSO-BP 神经网络模型的预测效果。本研究主要采用 SPSS 20.0 统计软件对数据进行气象因素等变量的筛选、描述性分析和多元线性回归模型的建立；采用 Matlab 2014a 统计软件建立 PSO-BP 神经网络模型。

结果：运用逐步回归法筛选出前一天的 PM_{10}、极大风速、最小相对湿度、日平均气温、能见度共 5 个主要影响变量，其中前一天的 PM_{10} 浓度对模型的贡献最大，其次是极大风速。对 PSO-BP 神经网络模型与多元线性回归模型测试样本的预测效果进行比较，多元线性回归模型的 RMSE、MAE、MAPE、PMAD、R^2 分别为 23.215、18.806、0.195、0.196 和 0.764；PSO-BP 神经网络模型的 RMSE、MAE、MAPE、PMAD、R^2 分别为 21.776、17.165、0.169、0.179 和 0.800。可见 PSO-BP 神经网络模型预测效果更优，模型的拟合效果与实际数据的误差更小。

[1] 注:该论文获得由中国统计教育学会举办的“2015 年（第四届）全国大学生统计建模大赛”大数据统计建模类本科生组一等奖。

结论：利用 PSO-BP 神经网络模型预测广州市未来一天的日均 PM_{10} 浓度效果较好，与多元线性回归模型相比其误差更小，预测效果更优，可为环境管理决策提供依据。

关键词：PM_{10}　多元线性回归　PSO-BP 神经网络　气象因素　广州市

引　言

（一）研究背景和目的

1.　研究背景

随着工业的发展和城市进程的加剧，我国乃至全球的大气环境污染形势愈趋严峻，人类健康因而受到巨大威胁。世界卫生组织（WHO）最新估计数据显示：每年有700万例的过早死亡与空气污染有关[1]，而流行病学研究表明，随着大气中的悬浮颗粒物（Particulate Matter，PM）浓度的升高,人体的呼吸道症状会加剧，因上呼吸道疾病就诊或住院的人数也会增加[2]，同时也会引起人体肺功能的降低及心肺疾病死亡率的增加[3,4]。而随着科研工作的深入，人们逐渐认识到直径小于或等于10μm的颗粒物（PM_{10}）是导致城市人群患病率和死亡率增加的主要因素[5]。因此，如何及时，准确的预测PM_{10}的浓度，为环境管理决策提供信息成为大家十分关注的问题。

研究表明，大气污染物与特定的气象因素有着密切的关系，气象因素往往制约着大气污染物的稀释、扩散、输送和转化，进而影响大气污染的浓度和分布[6,7]。在不同的气象条件下，同一污染源排放所造成的空气污染物浓度可相差几十倍甚至几百倍[8]。因此，运用气象因素等对PM_{10}浓度进行预测有着重要的意义。

广州市是广东省重要的政治和文化中心，其迅速发展的同时也带来了严重的环境污染问题。2008年广州市人民政府发布了空气污染综合整治实施方案[9]，有效的改善了空气质量，但日均PM_{10}浓度还保持在较高的水平，2008年-2011年期间有3.765%的日子PM_{10}浓度超标（按我国标准日均浓度150μg/m3计算），其防治问题依然值得探索。

2.　研究目的

近年来，随着计算机和信息技术的快速发展，使许多以前难以预测的空气质量预测逐渐成为可能。我国目前已有空气质量形式预报[10]，能够预测未来2-3日的空气质量情况，但其只将我国划分为京津冀、长三角、珠三角等三大区域，范围过大不够精准，且其对空气质量情况只能给出简单评价，未能提供具体大气颗粒物浓度指标。因此，建立一个对特定城市的PM_{10}浓度的预测模型，为环境管理决策提供准确、全面、及时的环境污染水平信息，对环境污染问题的治理和改善提供有力的工具显得尤为重要。广州市是珠江三角洲的重点经济发展城市，环境污染问题尤为严重，如何建立适合广州市的PM_{10}浓度预测模型是我们需要解决的问题。

（二） 研究现况

目前，国内有许多学者致力于研究城市空气污染浓度预测模式。吴嘉荣[11]用线性回归法建立了城市环境空气质量预报模式，表明了前一日PM_{10}浓度和气象因素与第二日的PM_{10}存在相关关系，对PM_{10}浓度进行了简单预测，但未进行预测效果评价。李祚泳等率先将神经网络应用于空气污染预测的探索性研究，预测了SO_2的浓度，并指出BP网络的预测精度优于模糊识别模型的预测精度[12]。周国亮等[13]利用BP神经网络对空气质量级别等计数资料作出了预测，准确率较高。石灵芝等[14]对长沙市PM_{10}每小时浓度进行预测，尽管检验预测时间较短（2008-01-05至2008-01-09共5天），但预测效果较好，整体R^2达到0.62。于宗艳等[15]利用免疫粒子群优化算法得到了空气质量评价模型，但未对空气质量作出预测。

在国外的研究中，Misiti M,等[16]人对每日PM_{10}浓度建立了混合线性回归进行预测，Thomas S等[17]人考虑了气象因素的滞后效应建立了多元线性模型，并用神经网络较好的预测了$PM_{2.5}$的浓度（R^2=0.79）。同样，Ul-Saufie AZ等[18]人利用其他污染物变量和气象变量也分别建立了多元线性回归模型和人工神经网络模型对PM_{10}浓度进行较为准确的预测。Jef H等[19]人在建立神经网络模型预测未来一天的PM_{10}浓度时加入了新的自变量边界层高度，但发现神经网络未能从中提取到有用的信息，因而并没有提高预测精度。W. Z. LU等[20]人提出了PSO-BP模型预测空气质量的可行性，但未进行预测效果评价。

综上所述，现有文献通过对PM_{10}浓度与气象因素等的分析，建立多元线性回归模型和神经网络模型等模型对PM_{10}浓度进行了不同程度的预测。但存在以下不足之处：一是部分国内研究只限于用当天的气象数据预测当天的颗粒物浓度，而在现实生活中当天的气象数据往往未能提前获取，因而其预测应用的意义不大；二是部分研究只根据经验理论对气象因素进行选择，没有运用科学的方法对气象因素等进行筛选并运用适合当地气象影响因素建立预测模型；另外，不同模型的

预测效果如何尚有待比较, 尤其是复杂的神经网络方法是否优于传统的多元线性模型有必要予以探讨，为该方法的推广应用提供参考。

（三） 本文研究思路与创新之处

1. 研究思路

基于以上研究目的与研究现况本研究提出如下（图 1）研究思路：以 2008 年-2011 年广州市地区内 9 个空气质量监测站点（天河职幼，市检测站，市 86 中，市 5 中，麓湖，花都师范，广雅中学，广东商学院，番禺中学等）的 PM_{10} 日平均浓度和气象因素等为研究对象，应用 BP 神经网络与粒子群优化算法（PSO）结合的 PSO-BP 神经网络模型预测未来一天的 PM_{10} 日平均浓度，同时与多元线性回归模型预测结果进行比较，从而说明 PSO-BP 神经网络模型对 PM_{10} 日平均浓度的预测效果，并为预测和控制空气污染提供一系列科学的依据。

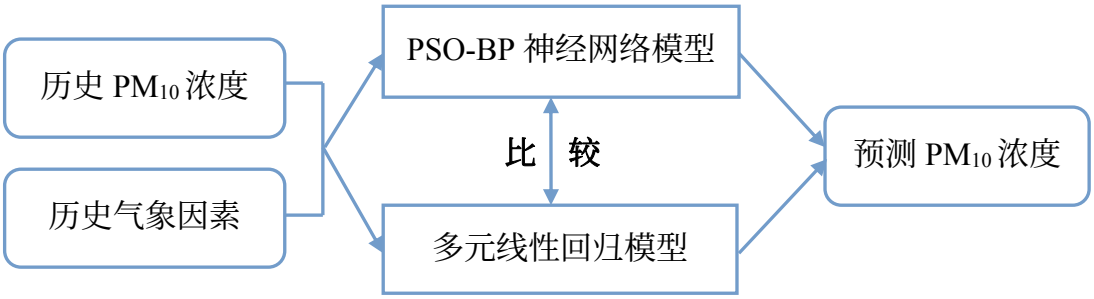

图 1 本文研究思路

2. 创新之处

（1）本研究以 2008 年-2011 年长时间段的日均 PM_{10} 浓度和气象因素等历史数据为基础, 运用前一日的 PM_{10} 浓度和气象数据建立适合广州市的 PM_{10} 预测模型，对日均 PM_{10} 浓度进行提前一天的预测。

（2）在 BP 神经网络的基础上加上 PSO 算法构建 PSO-BP 神经网络模型对日均 PM_{10} 浓度进行预测，更好的减小了预测误差。

（3）气象因素等变量主要运用了科学的逐步回归方法进行选择，筛选出影响 PM_{10} 浓度预测的主要气象因素等，进而建立多元线性回归模型和 PSO-BP 神经网络模型，并进行两模型的比较，更有力的说明重点研究模型 PSO-BP 神经网络模型的预测效果。

一. 方 法

（一）研究方法的基本原理

1. 多元线性回归模型

多元线性回归模型是探讨一个变量和多个变量之间的关系的常用方法，主要以多个自变量的最优组合共同预测或估计因变量，其在环境空气污染研究中也常被使用。多元线性回归模型的主要形式如下：

$$Y_{\mathrm{i}} = \beta_0 + \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + \ldots + \beta_{pi} X_{pi} + \varepsilon_i \tag{1}$$

其中，Y 是自变量（预测变量），β_0 是常数，β_1，$\beta_2 \cdots\cdots \beta_{\mathrm{p}}$ 是自变量 X_1，$X_2 \cdots\cdots X_p$ 的回归系数，ε 是残差（观测值与预测值的差值）。回归系数 β_0，β_1，$\beta_2 \cdots\cdots \beta_{\mathrm{p}}$ 常用最小二乘法求得[18, 21]。

2. BP 神经网络

线性模型只能解决线性可分问题，而 BP 神经网络属于多层感知器（Multi-layer Perceptrons，MLP）的一种，能够解决预测中的线性不可分问题。多层感知器除了输入层和输出层外，还具有若干隐含层。上下层之间实现全连接，而每层单元之间无连接。大部分情况下多层感知器采用误差反向传播（Back Propagation）的算法进行权值调整，即当一学习样本提供给网络之后， 神经元的激活值从输入层经中间层向输出层传播，在输出层的各个神经元获得网络的输入响应。随后， 按照减小目标输出与实际误差的方向，从输出层经过中间层逐层修正各层的连接权值，最后回到输入层。具体的方法如下：

（1）输入信息的顺向传播

隐含层中第 i 个神经元的输出为：

$$a_{1i} = f_1\left[\sum_{j=1}^{r} \omega_{1ij} P_j + \theta_{1i}\right],\ \mathrm{i}=1,2,\cdots,\mathrm{S1} \tag{2}$$

输出层中第 k 个神经元的输出为：

$$a_{2k} = f_2\left[\sum_{j=1}^{S1} \omega_{2ki} a_{1i} + \theta_{2k}\right],\ \mathrm{k}=1,2,\cdots,\mathrm{S2} \tag{3}$$

误差函数为：

$$E(\omega, B) = \frac{1}{2}\sum_{j=1}^{S1}(t_k - a_{2k})^2 \tag{4}$$

式中 S1、S2 分别为隐含层、输出层的神经元个数, k 为迭代次数[22]。

（2）误差函数的反向传播

输出层的权值变化, 对从第 i 个输入到第 k 个输出的权值有：

$$\sum \omega_{2ki} = -\eta\frac{\partial E}{\partial \Delta\omega_{2ki}} = -\eta\frac{\partial E}{\partial a_{2k}}\frac{\partial a_{2k}}{\partial \Delta\omega_{2ki}} = \eta(t_k - a_{2k})f_2^{'}a_{1i} = \eta\delta_{ki}a_{1i} \tag{5}$$

式中: $\delta_{ki} = (t_k - a_{2k})f_2^{'} = e_k f_2^{'}, e_k = t_k - a_{2k}$ ；

同理可得:

$$\Delta b_{2ki} = -\frac{\partial E}{\partial \Delta\theta_{2ki}} = -\eta\frac{\partial E}{\partial a_{2k}}\frac{\partial a_{2k}}{\partial \Delta\theta_{2ki}} = \eta(t_k - a_{2k})f_2^{'} = \eta\delta_{ki} \tag{6}$$

隐含层的权值变化, 对从第 j 个输入到第 i 个输出的权值有：

$$\Delta\omega_{1ij} = -\eta\frac{\partial E}{\partial \Delta\omega_{1ij}} = -\eta\frac{\partial E}{\partial a_{2k}}\frac{\partial a_{2k}}{\partial a_{1i}}\frac{\partial a_{1i}}{\partial \Delta\omega_{1ij}} = -\eta\sum(t_k - a_{2k})f_2^{'}\omega_{2ki}f_1^{'}p_j = \eta\delta_{ki}P_j \tag{7}$$

同理可得：

$$\Delta\theta_{1i} = \eta\delta_{ij} \tag{8}$$

式中负号表示梯度下降, 常数 $\eta(0 < \eta < 1)$ 表示比例系数, 即学习率[23, 24]。

（3）模型表达式

基于本研究，单个输出节点（PM_{10}）的 BP 神经网络模型的构建可简化为下式表示：

$$\hat{y}(I) = A_2(\sum_{i=1}^{N_i}\omega_m^2 \cdot (A_1(\sum_{m=1}^{N_m}\omega_{im}^1 \cdot x_i(I) + b_m^1)) + b_o^2) \tag{9}$$

其中 A_1 和 A_2 分别为隐含层和输出层的传递函数； ω_{im}^1 和 ω_m^2 分别表示为输入层 i 个节点到隐含层 m 个节点的权重和隐含层 m 个节点到单个输出节点的权重；

b_m^1和b_o^2分别表示第 m 个隐含层节点偏倚和输出层的偏倚；N_m和N_i分别表示输入层和隐含层节点数。

建立 BP 神经网络的参数选择具体分为 6 个：

（1）网络层数。

BP 网络可以包含一到多个隐含层。不过，理论上证明单个隐含层网络可以通过适当增加神经元节点的个数实现任意非线性映射。

（2）隐含层节点数。

目前并没有一个理想的解析式可以用来确定合理的神经元个数。通常做法是采用经验公式给出估计值：

$$\sum_{i=0}^{n} C_M^i > k \tag{10}$$

$$M = \sqrt{n+m} + a \tag{11}$$

$$M = \log_2 n \tag{12}$$

其中，k 为样本数，M 为隐含层神经元个数，n 为输入层神经元个数，m 为输出层神经元个数，a 是[0,10]之间的常数。若 i>M，规定$C_M^i = 0$。

（3）BP 学习时权值的初始值确定。

初始值过大过小都会影响学习速度，经验值为$(-2.4/F, 2.4/F)$或$(-3/\sqrt{F}, 3/\sqrt{F})$之间，其中 F 为权值输入端神经元个数。另外，为避免每一步权值的调整方向是同向的，应将初始权值设为随机数，本文也取初始权值和阈值为[0,1]之间的随机数[25]。

（4）传递函数的选择。

传递函数必须可微。一般隐含层使用 Sigmoid 函数，而输出层为线性函数。如果输出层也采用 Sigmoid 函数，则输出值将会被限制在(0,1)或(-1,1)之间。Sigmoid 函数又可分为 Log-Sigmoid 函数和 Tan-Sigmoid 函数。

（5）学习速度的选定。

学习速度不能选太大，否则算法不收敛。也不能太小，会使训练时间太长。一般选择 0.01~0.1 之间的值。

（6）训练方法的选择。

BP 修正权值的方式有两种：串行方式和批量方式。在串行方式中，每一个输入被作用于网络后，权重和阈值被更新一次。在批量方式中，所有的输入被应用于网络后，权重和阈值才被更新一次。使用批量方式不需要为每一层的权重和阈值设定训练函数，而只需为整个网络指定一个训练函数，使用起来相对方便，而且许多改进的快速训练算法只能采用批量方式，在这里我们只用批量方式。

确定以上参数后便可进行网络训练。

3. 粒子群算法（PSO）

粒子群算法，也称粒子群优化算法（PSO），是近年来发展起来的一种新的进化算法[26]。其源于生物社会学家对鸟群、鱼群或者昆虫捕食行为的研究，是一种实现简单、全局搜索能力强且性能优越的启发式搜索技术。鸟类捕食时，每只鸟找到食物最简单有效的方法就是搜索当前距离最近食物的鸟的周围区域，可视鸟群为粒子群，将食物视为全局最优解，将鸟群捕获食物的过程等价于粒子群寻找全局最优解的过程[27]。

在 PSO 算法中，每粒子都代表极值优化问题的一个潜在最优解，用位置、速度和适应度值三项指标表示该粒子的特征，适应度值由适应度函数计算得到，其值的好坏表示粒子的优劣。粒子在解空间中运动，通过跟踪个体极值 Pbest 和群体极值 Gbest 更新个体位置，个体极值 Pbest 是指个体所经历位置中计算得到的适应度值最优位置，群体极值是指种群中的所有粒子搜索到的适应度最优位置。粒子每更新一次位置，就计算一次适应度值，并且通过比较新粒子的适应度值和个体极值，群体极值的适应度值更新个体极值 Pbest 和群体极值 Gbest。

在 D 维搜索空间中，有 m 个粒子，其中第 i 个粒子的位置是 $\vec{x}_i=(x_{i1},x_{i2},...,x_{iD})$，也代表问题的一个潜在解 i=1,2,……，m,其速度为 $\vec{v}_i=(v_{i1},v_{i2},...,v_{iD})$。将 $\vec{x}_i$ 带入目标函数可计算出其适应值。记第 i 个例子搜索到的最优位置为 $\vec{p}_i=(p_{i1},p_{i2},...,p_{iD})$，整个粒子群搜索到的最优位置为 $\vec{p}_g=(p_{g1},p_{g2},...,p_{gD})$。粒子状态更新操作如下：

$$v_{id}^{k+1}=wv_{id}^{k}+c_1r_1(p_{id}^{k}-x_{id}^{k})+c_2r_2(p_{gd}^{k}-x_{id}^{k}) \tag{13}$$

$$x_{id}^{k+1}=x_{id}^{k}+v_{id}^{k+1} \tag{14}$$

其中，i=1，……，m,d=1,……，D；w 是非负常数，称为惯性因子。w 也可以随着迭代线性地减小；学习因子 c1 和 c2 是非负常数；r1 和 r2 是介于[0，1]之间的随机数；$v_{id} \in [-v_{\max}, v_{\max}]$,vmax 是常数。

迭代中止条件一般选为最大迭代次数和粒子群迄今为止搜索到的最优位置满足适应阈值[28]。

4. PSO-BP 神经网络模型

由于 BP 算法收敛速度慢而且极易陷入局部最优，在应用中网络结构的确定基本依赖经验，主要是采用递增或递减的试探方法来确定的网络隐节点，这些缺陷使得神经网络的训练样本和测试样本的输出具有不一致性和不可预测性，极大的限制了神经网络在实际预报中的应用[29]。

为了避免 BP 神经网络陷入局部极小值和增加其泛化性能，提供预测精度，采用 PSO 算法优化 BP 神经网络的权值和阈值。PSO 的适应度函数为神经网络的输出误差，公式为：

$$f_i = \frac{1}{n_i} \sum_{q=1}^{n_i} (O_{iq} - T_{iq})^2 \tag{15}$$

其中，n_i 为训练样本的个数，Q_{iq}、T_{iq} 分别为训练样本 q 在第 i 粒子的位置所确定的网络权值和阈值下的网络实际输出和期望输出[30]。

PSO-BP 神经网络的具体流程如图 2 所示：

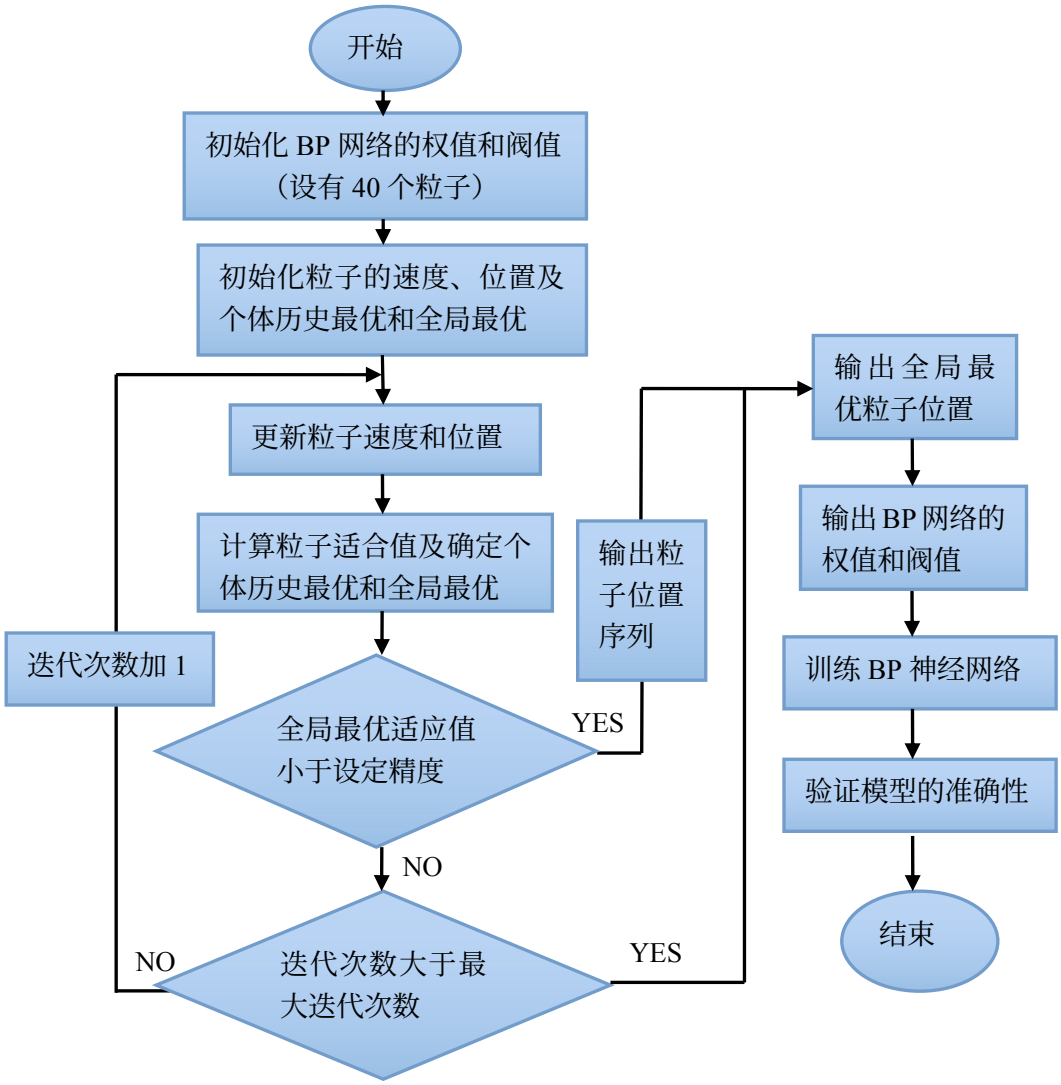

图 2　PSO-BP 神经网络流程图

由图 2 可知，PSO-BP 神经网络算法的具体步骤为：

（1）初始化 BP 神经网络和粒子群。

根据样本数据设计 BP 网络的输入、输出和隐含层神经元数目、学习函数及训练函数；根据粒子群的规模，按照个体结构产生一定数目的粒子群，其中不同的个体代表神经网络的 1 组不同的权值。同时，初始化粒子的速度、位置、个体历史最优 p_i、全局最优 p_g、迭代误差精度和最大迭代次数等[26]。

（2）迭代与更新。

利用式（13）（14）更新粒子的速度和位置，并用式（15）计算粒子的适应值。判断当前迭代次数是否大于最大迭代次数或当前最优适应值小于设定精度，若是满足条件，则输出全局最优粒子位置及 BP 网络的权值和阈值。

（3）训练 BP 网络

根据输出的 BP 网络权值和阈值训练 BP 神经网络，并运用测试样本对其进行检验，PSO-BP 神经网络完成。

（二） 统计分析软件

本研究主要采用 SPSS 20.0 统计软件对数据进行气象因素等变量的筛选、描述性分析和多元线性回归模型的建立；采用 Matlab 2014a 统计软件建立 BP 神经网络模型。本研究所用软件均为正版权威统计软件。

（三） 数据的来源

1. 地面 PM_{10} 数据：从广州市环境保护局官网[31]获得 2008-2011 年广州市 9 个监测站点（天河职幼，市检测站，市 86 中，市 5 中，麓湖，花都师范，广雅中学，广东商学院，番禺中学等）的日均 PM_{10} 浓度数据。9 个站点的具体分布如表 1 和图 3 所示。

2. 气象数据和能见度数据：从中国气象科学数据共享服务网获得广州市 2008 年-2011 年日均降水量、风速、风向、气压、气温、水汽压、相对湿度、日照时数等气象因素数据；从 Weather underground 网站获得 2008-2011 年能见度数据[32, 33]。

图 3　广州市空气监测点位图

表 1　广州市空气监测点位

测点名称	经纬度	所属行政区
广雅中学	E：113 ° 14'01" N：23 ° 08'31"	荔湾区
市 5 中	E：113 ° 15'35" N：23 ° 06'15"	海珠区
市监测站	E：113 ° 15'35" N：23 ° 07'59"	越秀区
天河职幼	E：113 ° 19'02" N：23 ° 08'09"	天河区
麓湖	E：113 ° 16'50" N：23 ° 09'25"	天河区
广东商学院	E：113 ° 21'12" N：23 ° 05'31"	海珠区
市 86 中	E：113 ° 25'54" N：23 ° 06'18"	黄埔区
番禺中学	E：113 ° 21'14" N：22 ° 57'05"	番禺区
花都师范	E：113 ° 12'40" N：23 ° 23'30"	花都区

（四） 数据的处理与变量的选择

1.　数据集的划分

在本研究中，我们建立了一个训练样本（包含 1430 行数据）用来建立模型以及一个测试样本（包含 31 行数据）用来检验模型的预测效果，具体分配如下：

（1）训练样本：训练样本的数据从 2008 年 1 月 1 日-2011 年 11 月 30 日，共 1430 行数据，其中自变量数据从 2008 年 1 月 1 日-2011 年 11 月 29 日，预测的 PM_{10} 数据从 2008 年 1 月 2 日-2011 年 11 月 30 日。

（2）测试样本：测试样本的数据从 2011 年 12 月 1 日-2011 年 12 月 31 日，共 31 行数据，其中自变量数据从 2011 年 11 月 30 日-2011 年 12 月 30 日，预测的 PM_{10} 数据从 2011 年 12 月 1 日-2011 年 12 月 31 日。

2.　数据预处理

本次研究意在建立适合广州市的 PM_{10} 预测模型，因而我们对 9 个站点的日均 PM_{10} 浓度数据求平均值来代表广州市的 PM_{10} 污染水平。同时，本文主要考虑前一天的 PM_{10} 浓度和气象因素与预测的 PM_{10} 浓度的相关关系。

3. 变量的选择

（1）因变量：预测的日均 PM_{10} 浓度；

（2）自变量：经过逐步回归方法筛选出与预测日期相对应的前一天气象因素等变量——PM_{10}（$PM_{10\,t-1}$）、极大风速（$JDFS_{t-1}$）、最小相对湿度（$MinRH_{t-1}$）、日平均气温（$Tamp_{t-1}$）、能见度（See_{t-1}）共 5 个主要影响变量。

（五）模型的构建

1. 多元线性回归模型

由筛选出的气象因素等自变量和预测的日均 PM_{10} 浓度因变量构建多元线性回归模型：

$$PM_{10t} = \beta_0 + \beta_1 PM_{10t-1} + \beta_2 JDFS_{t-1} + \beta_3 MinRH_{t-1} + \beta_4 Tamp_{t-1} + \beta_6 See_{t-1} + \varepsilon \qquad (16)$$

具体建模思路如下图 4 所示[34]：

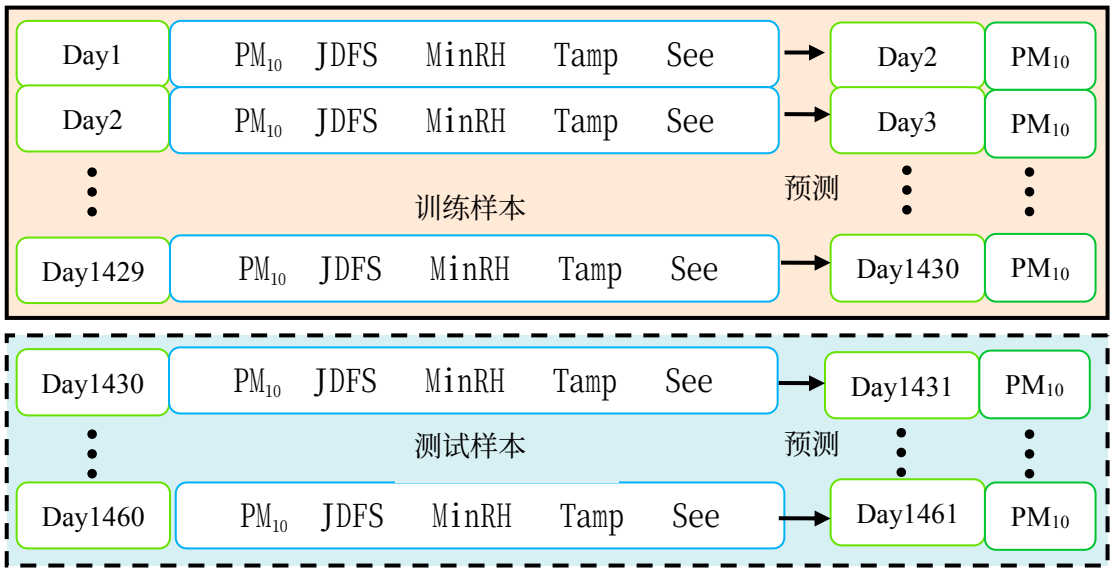

图 4 多元线性回归模型建模思路

2. PSO-BP 神经网络模型

本文专注于神经网络模型跟线性模型的预测效果作比较，因此我们采用跟线性模型相同的数据，并把经过逐步回归筛选出的 5 个主要影响因素作为 BP 神经网络的输入层节点。我们采用普遍的归一化处理数据，相应地，对训练后输出数据进行反归一化处理。随后运用经验公式对模型的参数进行设定：

（1） BP 神经网络具体参数设定

①隐含层传递函数选择双曲正切函数（Tan-Sigmoid）。在本研究中，相对于

使用 Log-Sigmoid 作传输函数能更好地减少预测误差。

②输出层传递函数选择线性函数，以保持输出的范围。

③训练方法采用 LM（Levenberg-Marquardt）算法，有更快收敛速度。

④ 训练的终止采用早终止（Early Stopping）技术。当误差函数（MSE）达到10^{-3}收敛即终止训练，避免了过度拟合的发生。

⑤隐含层节点数由经验公式（11）得出，并选择 M=11。

因此，建立输入节点有 5 个，隐含节点 11 个，输出节点 1 个的 BP 神经网络模型。具体建模思路如下图 5 所示：

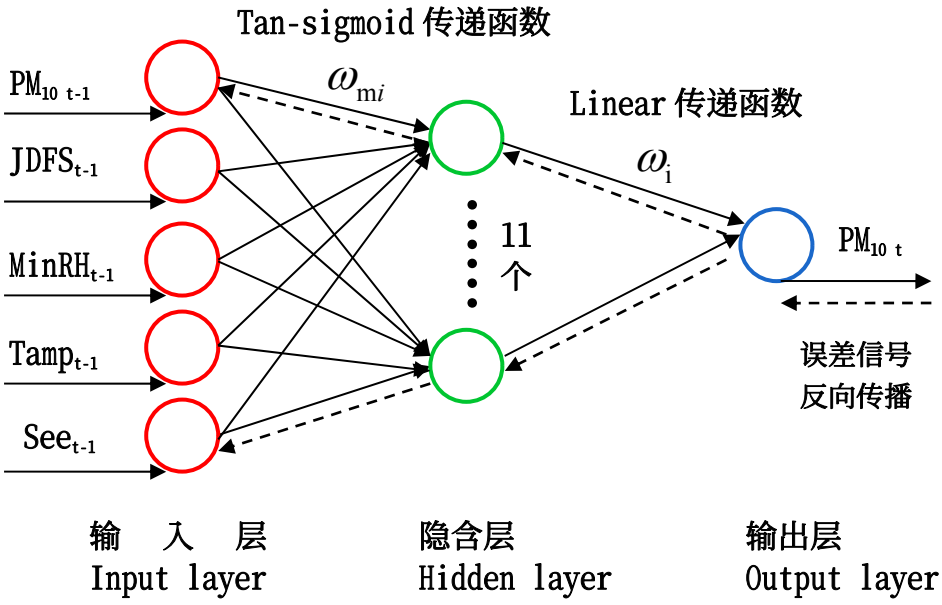

图5 BP神经网络模型建模思路

（2） PSO 具体参数设定

①群体规模 N。即粒子个数，一般情况下取 20~40。本文取 N=40。

②惯性权重ω。较大的ω可以加强 PSO 的全局搜索能力，较小的ω能加强局部搜索能力，而动态ω能够获得比固定值ω稳健和更好的寻优效果。因此采用基

于迭代次数的惯性权重因子递减方法以保证收敛，式子表达如下：

$$\omega(\mathrm{i}) = \omega_{\max} - ((\omega_{\max} - \omega_{\min}) / i_{\max}) \cdot i \qquad (17)$$

其中，ω表示惯性权重因子，ωmax表示最大惯性权重因子，ωmin表示最小惯性权重因子，i_{max}表示最大迭代次数，i 取值范围为[1：i_{max}][35]。本文取ωmax=0.900，ωmin=0.300，i_{max}=100。

③学习因子 c_1 和 c_2。c_1 和 c_2 通常等于 2，文献中也有其他取值[36]。本文取 c_1=2.800，c_2=1.300。

④粒子最大速度 V_{max}。它决定粒子在一个循环中最大的移动距离，通常设定为粒子的范围宽度。本文取经验值 V_{max}=1。

⑤终止条件。达到最大迭代次数或全局最优位置满足适应阈值。

二．结　果

（一）基本统计描述

1. 与预测的PM_{10}浓度相对应的前一日自变量描述性分析

2008年-2011年的PM_{10}浓度和气象环境自变量的描述性分析如下表2所示：

表2　模型自变量的描述性统计量

变量	N	最小值	最大值	均值	标准差
PM_{10}（$\mu g/m^3$）	1461	7.100	284.700	70.430	36.983

极大风速（0.1m/s）	1461	22.000	167.000	62.323	21.790
最小相对湿度（1%）	1461	12.000	96.000	51.906	16.161
日平均气温（0.1℃）	1461	54.000	335.000	223.637	64.857
能见度（km）	1461	1.000	10.000	6.329	2.029

2. 预测的日均PM_{10}浓度与时间的描述性分析

如表 3 所示，2008—2011 年年均 PM_{10} 浓度一直维持在 70□g/m^3 左右的较高水平，2009 年的 PM_{10} 浓度的最大值甚至达到了 284.700□g/m^3。而由 2008-2011 年日均 PM_{10} 浓度的时间序列图（图 6）可知，PM_{10} 浓度呈现冬春季高，夏秋季低的季节特征。广州市属于亚热带海洋性季风气候，受季风的影响，夏季来自海洋的暖气流形成高温、高湿、多雨的气候，有利于污染物的扩散与沉降。冬季，来自北方大陆的冷风形成低温、干燥、少雨的气候，不利于污染物的扩散，且冬季较易出现逆温，逆温层的出现会直接影响污染物的扩散，导致污染物质量浓度的增加[37]。

表3 预测的PM_{10}浓度年变化描述性统计量

年份	N	最小值	最大值	均值	标准差
2008	365	12.100	210.200	71.410	39.380
2009	365	7.100	284.700	70.450	43.722
2010	365	11.800	208.000	69.681	32.563
2011	365	14.200	208.400	70.250	30.947

图6 2008-2011年日均PM_{10}浓度图

3. 预测的日均PM_{10}浓度与气象因素等自变量的关系图

为了更好的研究PM_{10}浓度和气象因素等滞后效应对预测PM_{10}浓度的影响，我们做了如图7所示的散点图，并计算了它们之间的相关系数（表4）。由表4的Pearson相关系数值可知，前一天的PM_{10}浓度对预测当天的PM_{10}浓度影响作用最大，其次是极大风速。

表4 预测的PM_{10}与模型中自变量的相关系数

自变量	Person 相关系数	P 值
PM_{10}	0.661	<0.001
极大风速	-0.412	<0.001
最小相对湿度	-0.395	<0.001
日平均气温	-0.208	<0.001
能见度	-0.362	<0.001

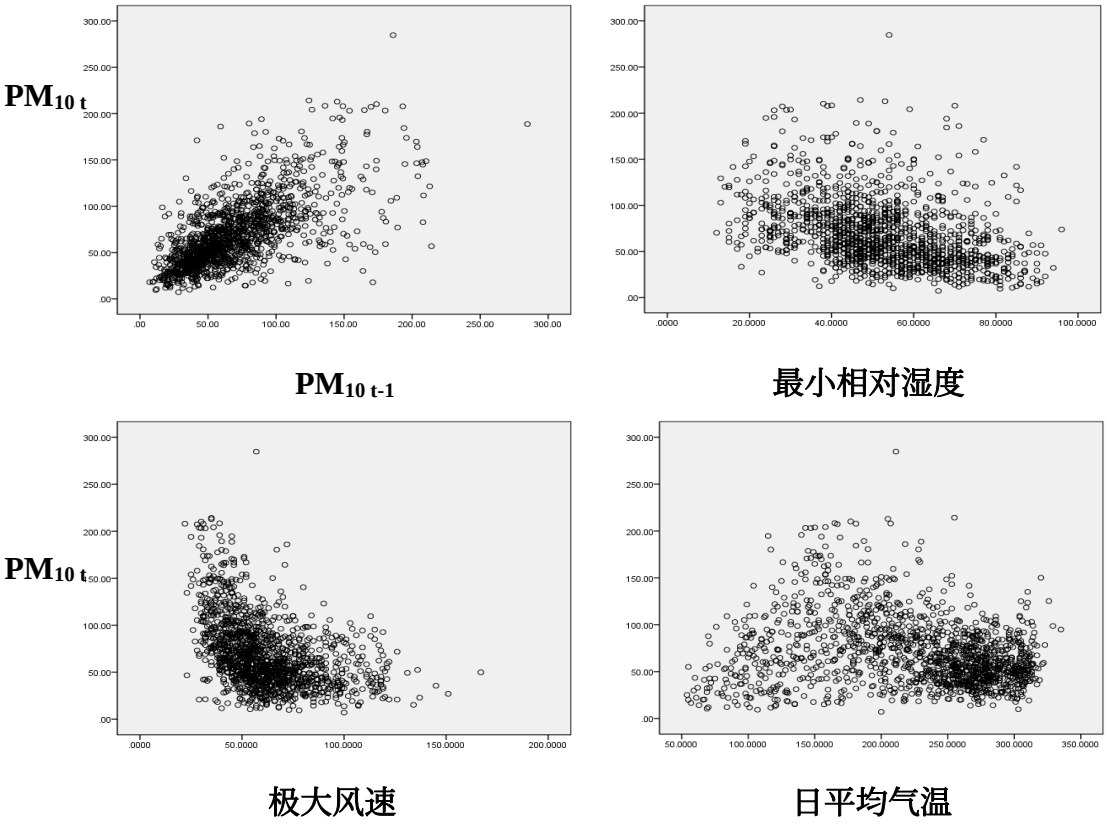

图7 预测的$PM_{10\,t}$与各自变量的散点图

（二） 模型的结果

1. 多元线性回归预测模型

本文主要运用最小二乘法对多元线性回归模型的参数进行估计，最终模型的参数估计如下表 5 所示：

表 5 多元线性回归模型自变量的参数估计

变量	参数估计	标准误差	t 值	P 值

常数项	123.758	5.784	21.397	<0.001
	0.446	0.022	20.197	<0.001
PM_{10}（$\mu g/m^3$）				
极大风速（0.1m/s）	-0.430	0.034	-12.770	<0.001
最小相对湿度（1%）	-0.582	0.045	-12.836	<0.001
日平均气温（0.1℃）	-0.066	0.010	-6.535	<0.001
能见度（km）	-2.078	0.408	-5.099	<0.001

由表 5 可知，模型中自变量的参数估计均有统计学意义（P<0.001），因此可构建广州市 PM_{10} 浓度预测的多元线性回归模型为：

$$PM_{10t} = 123.758 + 0.446PM_{10t-1} - 0.430JDFS_{t-1} - 0.582MinRH_{t-1}$$

$$- 0.066Tamp_{t-1} - 2.078See_{t-1} + \varepsilon \quad (18)$$

我们用建立好的多元线性回归模型（式 18）对 2011 年 12 月的日均 PM_{10} 浓度进行预测，预测结果如下图 8 所示。

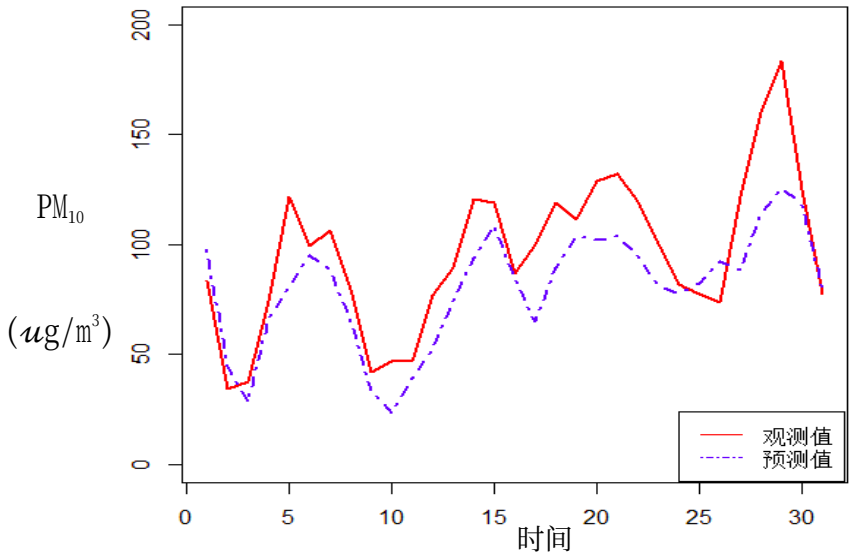

图8 多元线性回归模型预测值与观测值

2. PSO-BP神经网络预测模型

根据所设定的参数，我们用 PSO-BP 预测模型对 31 天（2011 年 12 月）的日均 PM_{10} 浓度进行预测，预测结果如下图 9 所示。可以看出，PSO-BP 预测模型能较好的预测日均 PM_{10} 的浓度趋势。

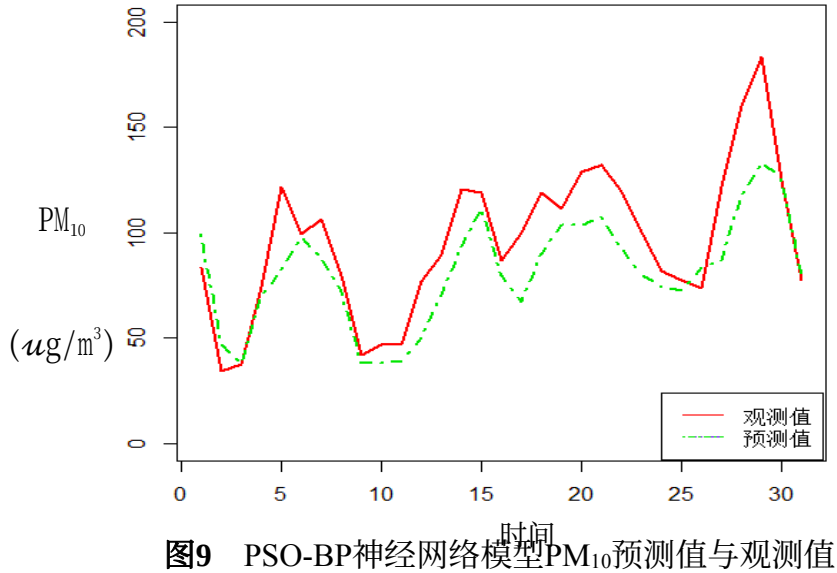

图9 PSO-BP神经网络模型PM_{10}预测值与观测值

（三）模型预测效果比较

利用纳入相同的变量的两个所建预测模型对 31 天（2011 年 12 月）的 PM_{10} 浓度进行了预测并比较，数据结果如下表 6 所示（PM_{10} 真实值和预测值单位均为：ug/m^3）：

表 6 模型预测结果比较

日期	实测值	预测值		相对误差(%)	
		多元回归	PSO-BP	多元回归	PSO-BP
1 日	83.300	97.670	99.480	17.251	19.424
2 日	34.400	44.890	46.680	30.494	35.698
3 日	37.600	28.460	38.240	24.309	1.702
4 日	74.200	65.960	69.810	11.105	5.916
5 日	121.600	80.820	82.200	33.536	32.401
6 日	99.300	94.790	97.630	4.542	1.682
7 日	106.700	88.240	88.150	17.301	17.385
8 日	79.700	66.010	71.860	17.177	9.837
9 日	42.000	33.910	37.990	19.262	9.548

10 日	46.900	23.380	38.460	50.149	17.996
11 日	47.100	39.390	39.100	16.369	16.985
12 日	76.900	52.660	50.150	31.521	34.785
13 日	89.600	74.140	70.190	17.254	21.663
14 日	120.700	93.870	93.640	22.229	22.419
15 日	118.900	108.510	111.440	8.738	6.274
16 日	86.700	85.250	79.790	1.672	7.970
17 日	100．000	65.140	67.180	34.860	32.820
18 日	119.300	89.510	90.060	24.971	24.510
19 日	111.600	103.580	103.660	7.186	7.115
20 日	128.900	102.270	103.730	20.659	19.527
21 日	132.400	104.010	107.590	21.443	18.739
22 日	119.800	94.870	92.400	20.810	22.871
23 日	100.700	81.140	80.060	19.424	20.497
24 日	81.800	77.470	74.000	5.293	9.535
25 日	77.700	82.300	72.600	5.920	6.564
26 日	73.800	92.320	83.790	25.095	13.537
27 日	122.000	88.250	87.050	27.664	28.648
28 日	160.200	114.410	117.30	28.583	26.779
29 日	183.600	125.790	133.090	31.487	27.511
30 日	125.100	118.720	126.320	5.100	0.975
31 日	77.800	80.050	80.440	2.892	3.393
均值	96.140	80.570	81.740	19.493	16.925

根据上表我们对两模型作图比较（图 10），由图可以看出，PSO-BP 神经网络比线性回归模型好，尤其是局部极小值部分更为精确。但两模型在某些段上还有一定的误差，原因可能是其他环境因素对 PM_{10} 有一定影响但未发现并收集纳入模型所致。本文着重于模型的比较，模型虽尚待完善但足以说明 PSO-BP 比线性模型预测效果好。

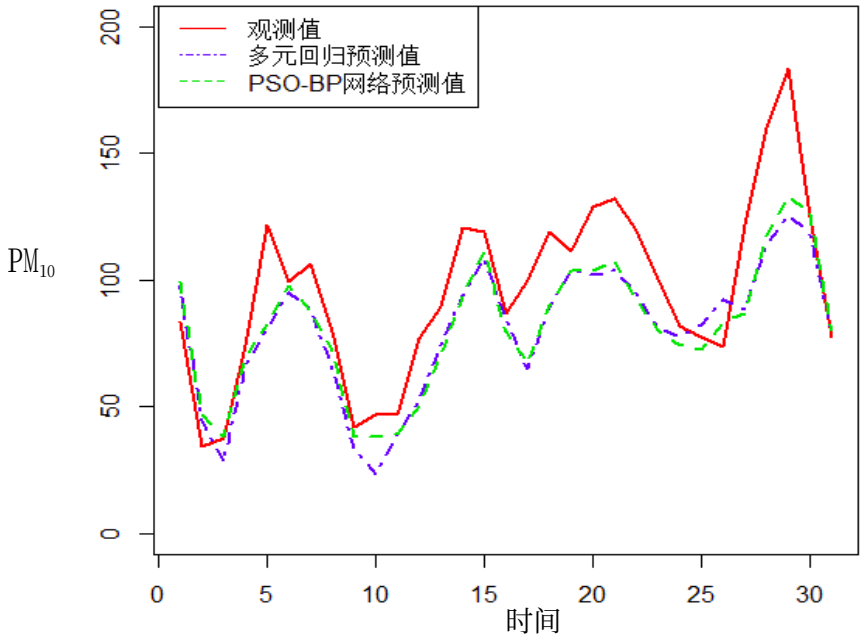

图10 两模型的预测值与观测值

为进一步评价线性回归模型和 PSO-BP 预测模型在研究日均 PM_{10} 质量浓度空间预测中的精度差，本研究分别计算了两模型的测试样本的均方根误（RMSE，$\mu g/m^3$）、平均绝对误差（MAE，$\mu g/m^3$）、平均绝对百分比误差（MAPE，%）、平均绝对偏差百分比（PMAD，%）和相关系数（R^2）等 5 个评价指标。

$$RMSE=\sqrt{\frac{\sum_{i}^{N}(P_i-O_i)^2}{N}} \tag{19}$$

$$MAE=\frac{\sum_{i}^{N}|P_i-O_i|}{N} \tag{20}$$

$$MAPE=\frac{\sum_{i}^{N}\left|\frac{O_i-P_i}{O_i}\right|}{N} \tag{21}$$

$$PMAD=\frac{\sum_{i}^{N}|P_i-O_i|}{\sum_{i}^{N}|O_i|} \tag{22}$$

$$R^2=\frac{\left[\sum_{i}^{N}(P_i-\overline{P}_i)(O_i-\overline{O}_i)\right]^2}{\sum_{i}^{N}(P_i-\overline{P}_i)^2\cdot\sum_{i}^{N}(P_i-\overline{P}_i)^2} \tag{23}$$

其中，P_i为预测值，O_i为观测值。

具体结果如下表 7 所示：

表7 模型预测效果比较

	RMSE	MAE	MAPE	PMAD	R^2
PSO-BP	21.776	17.165	0.169	0.179	0.800
线性回归	23.215	18.806	0.195	0.196	0.764

由结果发现, PSO-BP神经网络模型预测平均相对误差为16.925%, 均方根误、平均绝对误差、平均绝对百分比误差、平均绝对偏差百分比均比多元线性回归模型的小，且R^2比多元线性回归模型的高。因此，PSO—BP神经网络模型较普通多元线性回归模型有更好的预测效果，该模型具有良好的可靠性，能够较为准确地预测未来一段时间的日均PM_{10}浓度，为PM_{10}的预防、治理提供依据[38]。

三．讨　论

在本文研究的模型中纳入了与预测的PM_{10}浓度相对应的前一日的PM_{10}浓度($\mu g/m^3$)、极大风速(0.1m/s）、最小相对湿度(%)、日平均气温 (0.1℃)、能见度（km）5个变量，我们分析这些因素与预测的PM_{10}浓度的关系如下：

（一） PM_{10}浓度

前一日的PM_{10}浓度对预测当日的PM_{10}浓度有较强的影响。PM_{10}长期飘浮在空中，若无强风、降雨等气象条件，PM_{10}不易扩散或沉降，若遇到相对湿度较高的天气，还易形成雾，并与空气中的灰尘、汽车尾气等结合，形成二次颗粒物，加重污染。因此，PM_{10}浓度的滞后效应需要引起重视。

（二）极大风速

我们发现前一日的极大风速对PM_{10}浓度有直接影响，体现在对颗粒物有传播和扩散作用。广州市年平均极大风速为6.163m/s，超过了一般风速的阈值(6～7m／s)[39]。一般来说，地表的尘埃由于过大的风速扬起，导致PM_{10}浓度的增加，可表现出风速与颗粒物浓度呈现正相关性。而我们考虑的是极大风速的滞后效应，前一日的极大风速于当日效应减弱，低于阈值，此时，风速越大，越利于大气颗

粒物的稀释扩散，这与朱倩茹等[40]于2013年广州地区的研究结果一致。而长时间静风则不利于大气中PM_{10}浓度的扩散，使其积聚，浓度升高。

（三）最小相对湿度

广州地处南亚热带,属南亚热带典型的海洋季风气候，雨量充沛，年平均最小相对湿度达52.202%，相对较高，而最小相对湿度往往存在滞后效应，前一天的最小相对湿度与相应的预测的PM_{10}浓度的相关性为-0.395。较高的湿度对大气中悬浮颗粒物有凝结作用，空气中的水分使得PM_{10}粘在一起，形成较重的颗粒，最后沉淀下来，使得PM_{10}浓度降低。

（四）日平均气温

广州全年日平均气温较高，达22.550℃， PM_{10}浓度与日平均气温存在较弱的负相关， 随着气温的升高，PM_{10}浓度降低，可能的机理在于，近地面地表温度较高时，大气对流作用加强，混合层的高度也会随之增加，有利于降低大气稳定度，促进PM_{10}的稀释扩散[41]。

（五）能见度

大气中PM_{10}浓度的升高会导致大气能见度下降，能见度的降低会给社会造成极大的经济损失和健康安全隐患。能见度与颗粒物中化学组分关系密切，其中，PM_{10}中硫酸盐浓度对能见度影响最大[42]。风速低，湿度大已形成稳定的大气结构，不利于PM_{10}的稀释扩散，由此造成的雾或雾霾天气将会影响能见度。此外，大气中其他污染物如汽车尾气、烟、尘等加强了对光的吸收、散射作用,使来自物体的光信号减弱，也造成能见度的降低。

四．启示和建议

（一）加强组织管理部门的内部联系

加强卫生部门的监测预测，及时根据气象部门和环保部门提供的数据对颗粒物浓度进行预测，协调各组织各部门的工作，提前做好预防准备工作，使污染危害减小到最低。

（二）加强对污染源的监督管理力度

颗粒物浓度不断地升高其主要原因是工业及交通的大量污染物的排放，因此需要严格控制污染源的排放。加强对重点工业企业在线监测，加大监管力度，严查违法排污。联动珠三角相关城市，共同防治区域空气污染。

（三）加强对市民的环保健康教育

社会环境很大一部分都是受人类影响的，要创造出适合人类生活、工作的环境，协调人与自然的关系，环保健康教育必不可少。要积极开展环保健康讲座，大力宣传环保健康知识，加强人们的环保意识。

五．研究的不足和展望

（一）由于未收集到广州市其他站点的日均 PM_{10} 浓度数据，以天河职幼，市检测站，市 86 中，市 5 中，麓湖，花都师范，广雅中学，广东商学院，番禺中学 9 个站点所得的日均 PM_{10} 浓度代替广州市的日均 PM_{10} 浓度。希望在日后的研究中能找到更多广州市监测站点的数据，使数据更加精确，有利于后面的研究。

（二）本文主要分析气象因素对 PM_{10} 浓度的影响，在模型中暂未考虑到加入其他空气污染物变量。希望在日后的研究中加入如 SO_2 等的空气污染物建立预测模型，优化我们现有的模型。

（三）本研究中运用 PSO-BP 神经网络模型虽已较好的预测到日均 PM_{10} 浓度，但其误差还相对较大，模型还需要进一步的改进，使预测的精度更高。

参考文献

[1] 世界卫生组织（WHO）官网. http://www.who.int/mediacentre/news/releases/2014/air-pollution/zh/

[2] Effects of Air Pollutants on upper respiratory tract and eye symptoms[J].

[3] Gilmour P S, Brown D M, Lindsay T G, et al. Adverse health effects of PM10 particles: involvement of iron in generation of hydroxyl radical[J]. Occup Environ Med. 1996, 53(12): 817-822.

[4] Pope C R, Bates D V, Raizenne M E. Health effects of particulate air pollution: time for reassessment?[J]. Environ Health Perspect. 1995, 103(5): 472-480.

[5] 熊生龙. 贵阳空气中 PM_（10）浓度的神经网络模拟预测与运用[D]. 浙江大学, 2006.

[6] 陈巧俊，王雪梅，吴志勇，等. 珠三角城市扩张对春季主要气象参数和 03 浓度的影响[J]. 热带气象学报. 2012, 28(3): 356-366.

[7] 冯建军, 沈家芬, 梁任重, 等. 广州市 PM_(10)与气象要素的关系分析[J]. 中国环境监测. 2009(01): 78-82.

[8] 朱敏，陈海宇，张晓君. 上海市青浦区 PM_10_污染状况及其与气象要素的关系_朱敏[J]. 中国环境管理. 2012(1): 7-11.

[9] 印发广州市 2008-2010 年空气污染综合整治实施方案的通知[J]. 广州政报. 2008(11): 13-39.

[10] 中国环境监测总站. http://www.cnemc.cn/publish/totalWebSite/news/news_43317.html

[11] 吴嘉荣. 用线性回归法建立城市环境空气质量预报模式[J]. 引进与咨询. 2005(12): 29-30.

[12] 曹兰. 空气中 PM_(10)浓度的 BP 神经网络预报研究[J]. 污染防治技术. 2010(01): 18-21.

[13] 周国亮, 刘希玉, 武鲁英. BP 神经网络模型在空气质量级别评价中的应用[J]. 计算机工程与设计. 2009(02): 392-394.

[14] 石灵芝，邓启红，路蝉，等. 基于 BP 人工神经网络的大气颗粒物 PMlo 质量浓度预测[J]. 中南大学学报. 2012, 43(5): 1969-1974.

[15] 于宗艳，韩连涛. 免疫粒子群算法优化的环境空气质量评价方法[J]. 环境工程学报. 2013(11): 4486-4490.

[16] Misiti M, Misiti Y, Poggi J, et al. Mixture of linear regression models for short term PM10 forecasting in Haute Normandie (France)[J]. Case Studies In Business, Industry And Government Statistics. 2015, 6(1): 47-60.

[17] Thomas S, Jacko R B. Model for Forecasting Expressway Fine Particulate Matter and Carbon Monoxide Concentration: Application of Regression and Neural Network Models[J]. Journal of the Air & Waste Management Association. 2007, 57(4): 480-488.

[18] Ul-Saufie A Z. Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM10 Concentration Level Based On Gaseous And Meteorological Parameters[J]. International Journal of Applied Science and Technology. 2011, 1(4).

[19] Hooyberghs J, Mensink C, Dumont G, et al. A neural network forecast for daily average PM 10 concentrations in Belgium[J]. Atmospheric Environment. 2005,

39(18): 3279-3289.

[20] W.Z. LU, H.Y.FAN, A.Y.T. LEUNG, et al. ANALYSIS OF POLLUTANT LEVELS IN CENTRAL HONG KONG APPLYING NEURAL NETWORKMETHOD WITH PARTICLE SWARM OPTIMIZATION[J]. Environmental Monitoring and Assessment. 2002(79): 217-230.

[21] 马雁军, 杨洪斌, 张云海. 空气污染预测与地面气象要素应用[J]. 气象科技. 2004(02): 123-125.

[22] 武常芳, 张承中, 邢诒, 等. 基于 B-P 神经网络优化算法的城市环境空气中 PM_(10)浓度预测模型[J]. 环境保护科学. 2008(01): 1-3.

[23] 陈明. MATLAB 神经网络原理与实例精解[M]. 清华大学出版社, 2013.

[24] 黄淮滨, 张宏东, 张传刚. 基于 BP 神经网络和 ARIMA 组合模型的空气环境 API 预测[J].

[25] 高隽. 人工神经网络原理及仿真实例[M]. 机械工业出版社, 2005.

[26] 李慧民, 李振雷, 何荣军, 等. 基于粒子群算法和 BP 神经网络的冲击危险性评估[J]. 采矿与安全工程学报. 2014(02): 203-207.

[27] 李强, 周轲新. 基于 PSO-BP 算法的压力传感器温度补偿研究[J]. 电子学报. 2015(02): 412-416.

[28] 王岁花, 冯乃勤, 李爱国. 基于粒子群优化的 BP 网络学习算法[J]. 计算机应用与软件. 2003(08): 74-76.

[29] 吴建生, 刘丽萍, 金龙. 粒子群-神经网络集成学习算法气象预报建模研究[J]. 热带气象学报. 2008(06): 679-686.

[30] 王爱萍, 江丽. 基于 PSO 的 BP 神经网络学习算法[J]. 计算机工程. 2012, 38(21): 193-196.

[31] 广州市环境保护局官网. http://www.gzepb.gov.cn/

[32] 中国气象科学数据共享服务网. http://www.escience.gov.cn/metdata/page/index.html

[33] Weather underground 网站. http://simplifiedchinese.wunderground.com/

[34] Zhou Q, Jiang H, Wang J, et al. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network[J].

Science of The Total Environment. 2014, 496: 264-274.

[35] 洪亮，李瑞娟. 基于粒子群算法优化 BP 神经网络的色彩空间转换[J]. 包装工程. 2014(09): 105-109.

[36] 韩力群. 人工神经网络教程（第一版）[M]. 北京邮电大学出版社, 2006.

[37] Meng H J Z M. Inter-comparison of seasonal variability and nonlinear trend between AERONET aerosol optical depth and PM10 mass concentrations in Hong Kong[J]. 中国科学：地球科学英文版. 2014, 57(11): 2606-2615.

[38] 武建辉，王国立. BP 神经网络与多重线性回归模型在煤工尘肺发病工龄预测中的比较研究[J]. 中国煤炭工业医学杂志. 2014(12): 1992-1995.

[39] 唐猛. 长沙市颗粒物 PM10 浓度统计学分布特性与预测[D]. 中南大学, 2010.

[40] 朱倩茹, 刘永红, 徐伟嘉, 等. 广州 PM_(2.5)污染特征及影响因素分析[J]. 中国环境监测. 2013(02): 15-21.

[41] 陈阳，曾钰，张琴，等. 气象因素对长沙市 PM_(2.5)周期性变化规律的影响分析[J]. 四川环境. 2014(06): 81-87.

[42] 刘爱霞，韩素芹，姚青，等. 2011 年秋冬季天津 PM_2_5_组分特征及其对能见度的影响_刘爱霞[J]. 气象与环境学报. 2013, 29(2): 42-47.

附录

附录1　2008-2011年9个站点日均PM10数据.xls

附录2　2008-2011年广州市日均PM10数据+气象数据.sav

附录3　matlab分析中运用的数据（data）.xls

附录4　PSO-BP matlab程序.m

附录5　逐步回归后筛选出的变量.sav