

基于气象参数的某城市空气质量 二次预测模型

王校举

上海理工大学机械工程学院, 上海

收稿日期: 2023年2月28日; 录用日期: 2023年5月10日; 发布日期: 2023年5月17日

摘要

本文主要研究了空气质量预报问题, 广泛收集了各类相关数据, 剔除异常数据并利用平均差值补全缺失数据得到计算数据。对气象参数进行合理分类, 并计算相关系数。通过超限学习机神经网络建立预测模型预测各污染物的浓度, 在一次预测的基础上采用遗传算法对模型的参数进行优化和再预测, 使预测出的空气质量指数(AQI)的相对误差最大值最小, 首要污染物误差最小。通过对比分析预测值验证了二次模型的合理性。

关键词

相关系数, 超限学习机, 遗传算法, 空气质量指数

A Secondary Prediction Model of Air Quality in a City Based on Meteorological Parameters

Xiaoju Wang

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 28th, 2023; accepted: May 10th, 2023; published: May 17th, 2023

Abstract

This paper mainly studies the problem of air quality forecasting, collects various relevant data extensively, removes abnormal data, and uses the average difference to complete the missing data to obtain the calculated data. The meteorological parameters are reasonably classified and the

correlation coefficients are calculated. A prediction model is established to predict the concentration of each pollutant through an ELM neural network, and the genetic algorithm is used to optimize and re-predict the parameters of the model on the basis of one prediction, so as to maximize the relative error of the predicted air quality index (AQI). The smallest value is the smallest primary pollutant error. The reasonableness of the quadratic model is verified by comparing and analyzing the predicted values.

Keywords

Correlation Coefficient, ELM, Genetic Algorithm, Air Quality Index

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

大气污染系指由于人类活动或自然过程引起某些物质进入大气中, 呈现足够的浓度, 达到了足够的时间, 并因此危害了人体的舒适、健康和福利或危害了生态环境[1]。目前, 每个城市空气质量监测点都收集了大量的关于气象参数的数据, 这些大量的数据使得以数据为驱动方式对空气质量的预测报告成为了可能。空气质量主要受到二氧化硫(SO₂)、二氧化氮(NO₂)、粒径小于 10 μm 的颗粒物(PM₁₀)、粒径小于 2.5 μm 的颗粒物(PM_{2.5})、臭氧(O₃)、一氧化碳(CO)六种污染物的影响[2], 有研究指出 PM_{2.5} 与温度、风速、降水等等因素有关, 其中受风速影响较大[3]。唐之享对空气质量预测的目标分析, 建立了统计模型方式, 实现了对空气质量预测的整体框架模型[4]。赵晓阳通过实验对比验证了 LSTM 神经网络预测的可靠性[5]。

本文试图给出了一些分类模型的正式描述, 并调查了该模型用于生成, 测试和确认假设的可行方式, 结合当前预测模型的一些特点, 用于预测实际过程中的空气质量变化问题。通过建立相关空气质量预测模型, 来对未来空气质量变化进行预测。为更好地同时预测监测点的空气质量变化情况, 在一次预测模型的基础上梳理收集统计变量数据, 建立可表征空气质量二次预报数学模型。通过对分析数据和分类模型的协议偏差分数来评估模型的可信度以及首要污染物预测准确度, 利用已建立二次预测模型来预测 2021 年 7 月 13 日至 7 月 15 日 6 种常规污染物的单日浓度值并计算相应的 AQI 和首要污染物。

2. 数据来源与数据预处理

2.1. 数据来源

本次预测数据来源于某城市监测点所记录的从 2019/4/16 日到 2021/7/12 日的 6 种污染物的浓度和 5 种自然因素以及 2020-7-23 日到 2021-7-13 日的一次预报数据。

为了便于问题的研究, 对于本文当中的某些条件进行合理的假设及简化:

- 1) 假设评价质量的各个指标间的相互作用关系忽略不计;
- 2) 假设空气质量只和下文提到的气象参数有关;
- 3) 假设本次所搜集的数据能够客观的反应当天污染物浓度的实际情况;
- 4) 假设在预测模型中, 未来预测日期内没有发生重大的自然灾害或自然突变;

根据《环境空气质量指数(AQI)技术规定(试行)》(HJ633-2012), 空气质量指数(AQI)可用于判别空气质量等级。首先需得到各项污染物的空气质量分指数(IAQI), 其计算公式如下:

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} \cdot C_p - BP_{Lo} + IAQI_{Lo} \quad (1)$$

空气质量指数(IAQ)取各分指数的最大值, 即:

$$IAQ = \{ \max \{ IAQI_1, IAQI_2, IAQI_3, \dots, IAQI_n \} \} \quad (2)$$

其中, $IAQI_p$: 污染物P的空气质量分指数, 结果取整数; C_p : 污染物P的质量浓度; BP_{Hi}, BP_{Lo} : 与 C_p 相近的污染物浓度限值的高位值与低位值; $IAQI_{Hi}, IAQI_{Lo}$: 与 BP_{Hi}, BP_{Lo} 对应的空气质量分指数; $IAQI_1, IAQI_2, IAQI_3, \dots, IAQI_n$: 各污染物项目的分布指数。在本文中, IAQ 的计算仅与公式(3)中的六种污染物有关; 计算公式如下:

$$IAQ = \{ \max \{ IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{PM_{10}}, IAQI_{PM_{2.5}}, IAQI_{O_3}, IAQI_{CO} \} \} \quad (3)$$

2.2. 数据预处理

在实际情况中, 由于一些突发情况如: 仪器故障、传输异常等不可控的因素会导致空气质量数据存在丢失的情况, 这样不但会丢失有效信息还会增加模型的不稳定性。因此根据空气质量数据具有一定的时间序列性, 选用差值法对缺失数据进行补充, 公式如下所示[6]:

$$f(x_i) = \begin{cases} x_{i-1}, & x_i = \text{nan}, x_{i-1} \neq \text{nan} \\ x_i, & x_{i-1} \neq \text{nan} \end{cases} \quad (4)$$

式中: 若当前时刻 x_i 为空而前一时刻不为空, 则补充前一时刻值。

3. 基于灰色关联分析方法模型的气象条件分类

3.1. 相关因素影响机理及权重估计

气象条件物理特征的短期不平衡状态, 与气候不同, 天气具有不稳定性, 天气的这种不稳定性可以表征为随机变化或者以某些规律逐步变化, 为研究复杂天气模型的“不稳定模态特征”, 需对天气变化的各类子集因素进行分析以及归类。

3.2. 基于逐步回归方法的变量选择

逐步回归是实现变量选择的一种方法, 基本思路为: 先确定一个初始子集, 然后每次从子集外影响显著的变量中引入一个对 y 影响最大的, 再对原来子集中的变量进行检验, 从影响不显著的变量中剔除一个影响最小的, 直到不能引入和剔除为止。使用逐步回归有两点值得注意, 一是要适当地选定引入变量的显著性水平 α_{in} 和剔除变量的显著性水平 α_{out} , 显然 α_{in} 越大, 引入的变量越多; α_{out} 越大, 剔除的变量越少。二是由于各个变量之间的相关性, 一个新的变量引入后, 会使原来认为显著的某个变量变得不显著, 从而被剔除, 所以在最初选择变量时应尽量选择相互独立性强的变量[7]。其中, 自变量 X_1 、 X_2 、 X_3 、 X_4 、 X_5 分别指代温度、湿度、气压、风速和风向。通过剔除不显著变量, 得到了新的统计结果, 虽然剩余标准差 $RMSE$ 变化不大, 但是统计量 F 的值明显变大, 因此新的回归模型更好一些。其中, 红色表示被删除的不显著变量, 蓝色表示保留的变量。图 1~5 为各污染物的回归交互图。

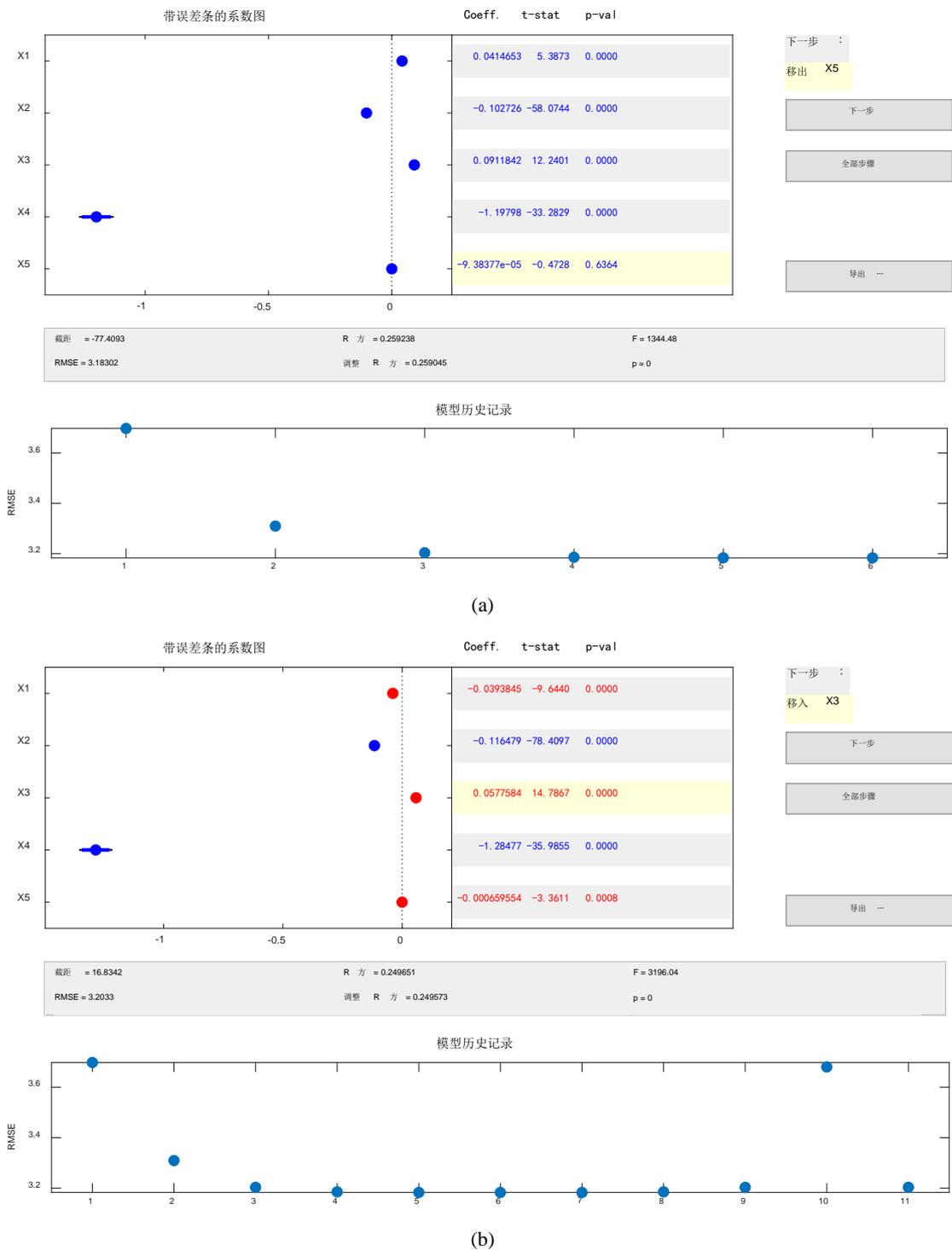


Figure 1. SO₂ gradually returns to the interactive screen. (a) Insignificant independent variables are not removed; (b) Insignificant independent variables have been removed

图 1. SO₂ 逐步回归交互画面。(a) 未剔除不显著自变量；(b) 已剔除不显著自变量

观察图中信息，可知风速的变化对 SO₂ 的浓度影响最大，并且随风速的增大，SO₂ 浓度减小。另一个保留的变量为湿度，其对 SO₂ 浓度呈现较小的负相关。

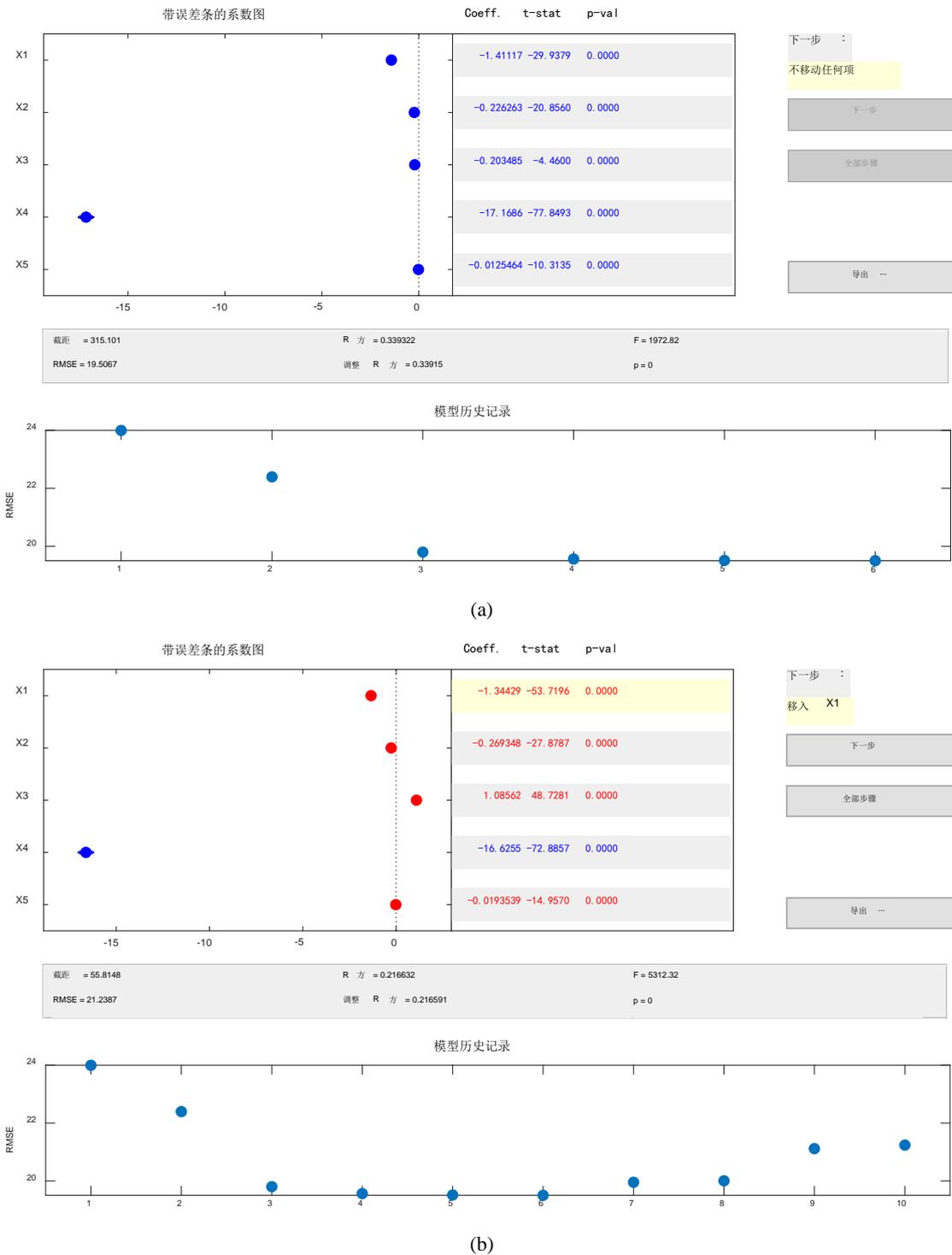


Figure 2. NO₂ gradually returns to the interactive screen. (a) Insignificant independent variables are not removed; (b) Insignificant independent variables have been removed
图 2. NO₂ 逐步回归交互画。(a) 未剔除不显著自变量；(b) 已剔除不显著自变量

其中，自变量含义和剔除方法与上述相同。观察图中信息，可知风速的变化对 NO₂ 的浓度影响最大，并且随风速的增大，NO₂ 浓度减小。

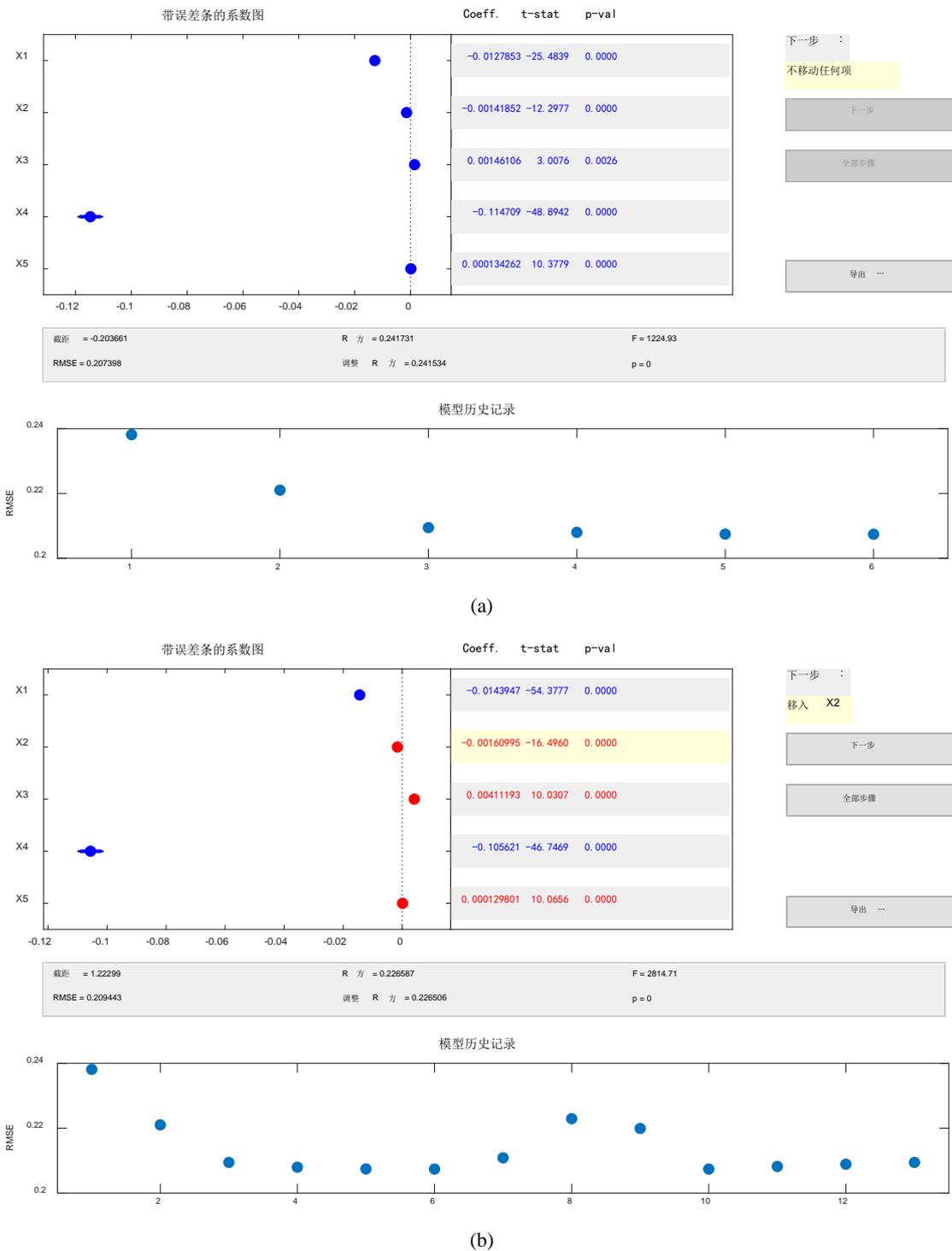


Figure 3. CO gradually returns to the interactive screen. (a) Insignificant independent variables are not removed; (b) Insignificant independent variables have been removed
图 3. CO 逐步回归交互画。(a) 未剔除不显著自变量；(b) 已剔除不显著自变量

观察图中信息，可知风速的变化对 CO 的浓度影响最大，并且随风速的增大，CO 浓度减小。另一个保留的变量为温度，其对 CO 浓度呈现较小的负相关。

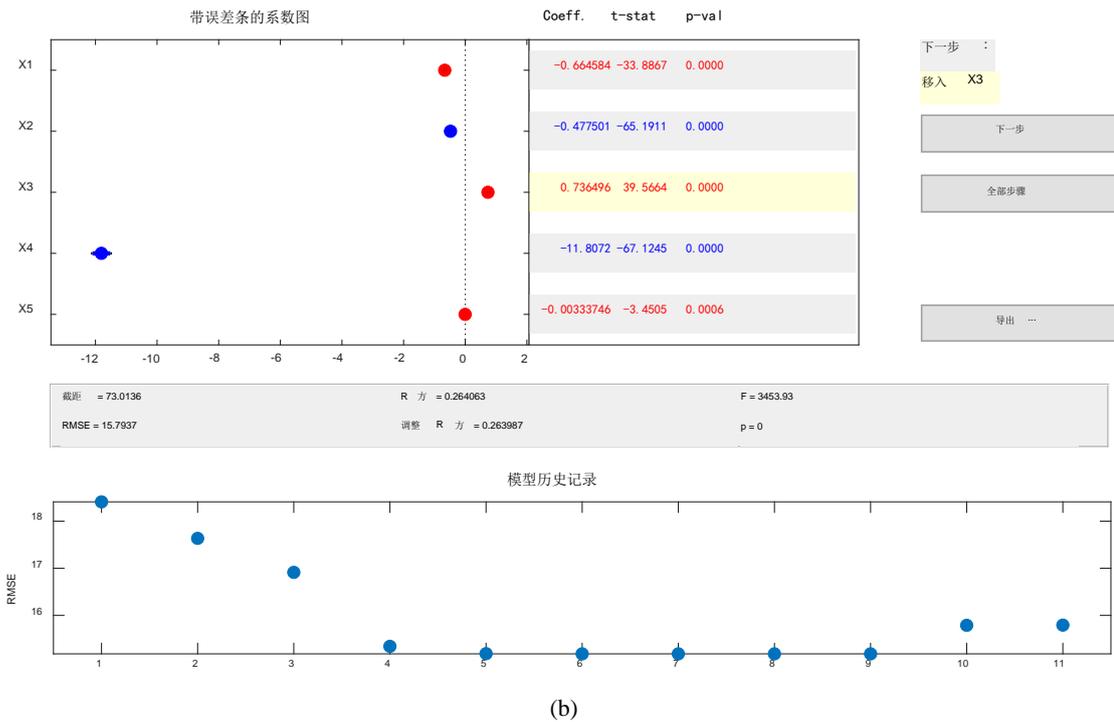
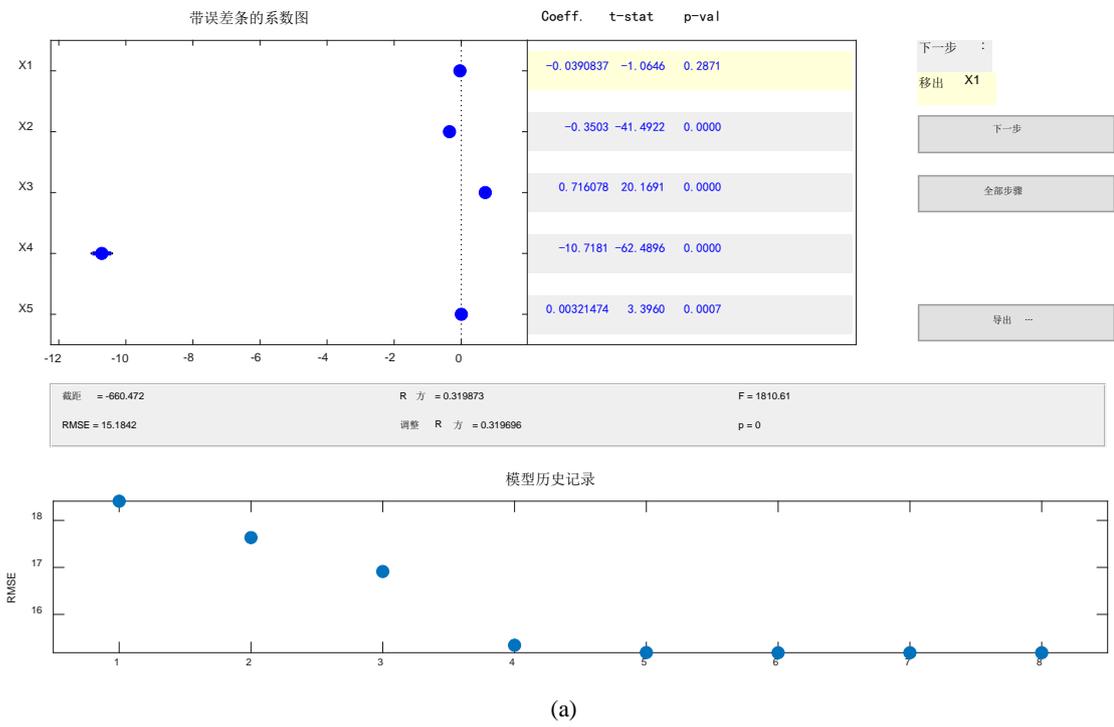


Figure 4. $PM_{2.5}$ gradually returns to the interactive screen. (a) Insignificant independent variables are not removed; (b) Insignificant independent variables have been removed
图 4. $PM_{2.5}$ 逐步回归交互画面。(a) 未剔除不显著自变量; (b) 已剔除不显著自变量

观察图中信息, 可知风速的变化对 PM_{10} 的浓度影响最大, 并且随风速的增大, PM_{10} 浓度减小。另一个保留的变量为湿度, 其对 PM_{10} 浓度呈现较小的负相关。

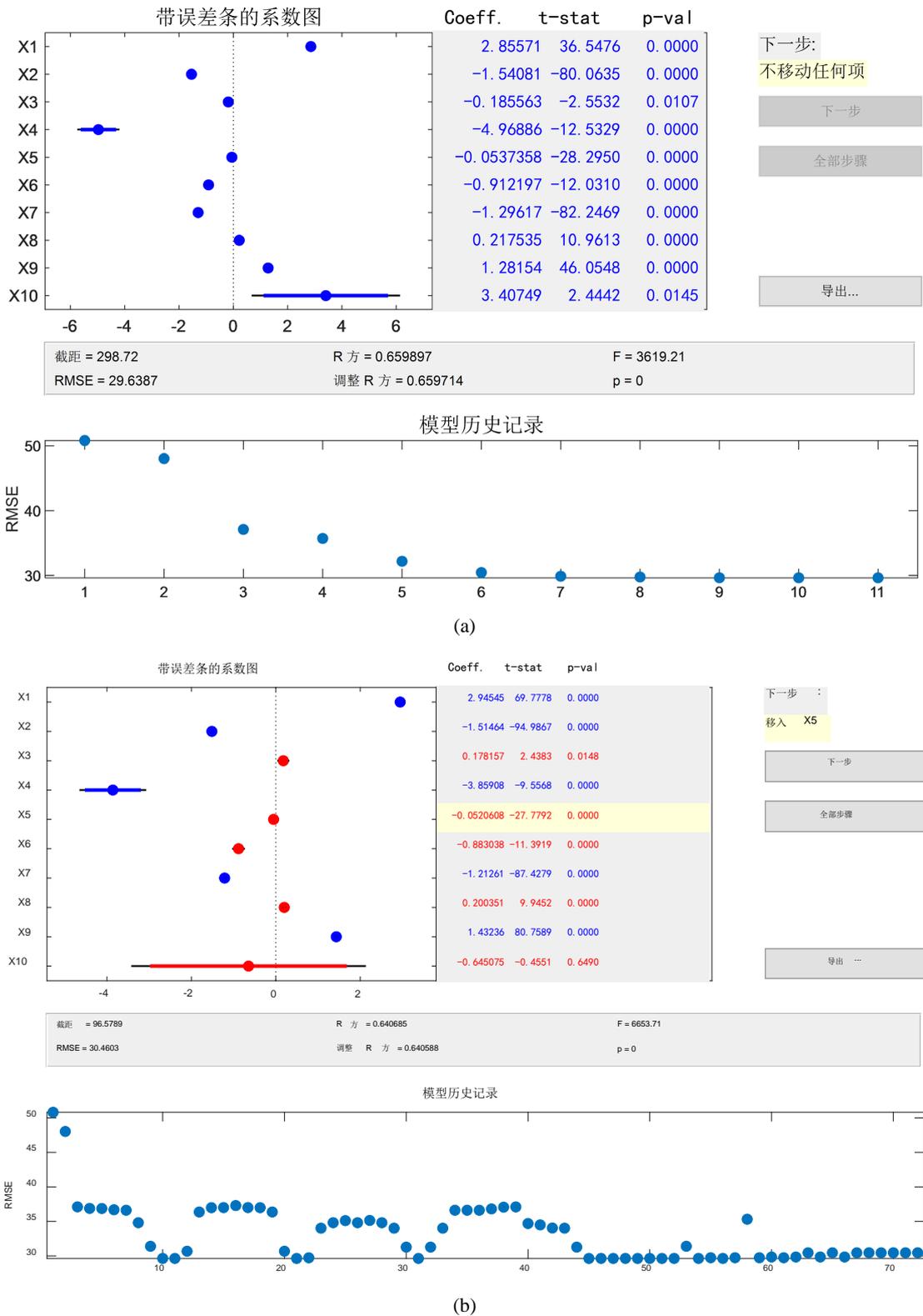


Figure 5. O₃ gradually returns to the interactive screen. (a) Insignificant independent variables are not removed; (b) Insignificant independent variables have been removed
图 5. O₃ 逐步回归交互画面。(a) 未剔除不显著自变量; (b) 已剔除不显著自变量

臭氧浓度预报是六项污染物预报中较难的一项，其原因在于：作为六项污染物中唯一的二次污染物，臭氧并非来自污染源的直接排放，而是在大气中经过一系列化学及光化学反应生成的。所以，可以通过引入更多自变量来分析其形成原因。其中，新添加的自变量 X_6 、 X_7 、 X_8 、 X_9 、 X_{10} 分别指代 SO_2 浓度、 NO_2 浓度、 PM_{10} 浓度、 $PM_{2.5}$ 浓度和 CO 浓度。观察图中信息，可知风速的变化对 O_3 的浓度影响最大，呈现高度负相关；温度的变化对 O_3 浓度呈正相关；湿度 NO_2 浓度呈较小的负相关。

又由于使用逐步回归的方法所得到的交互画面只有对污染物浓度影响最大的变量显示最为显著，而其它不显著的变量无法清楚了解其关联度。因此，使用灰色关联度分析法来求解其变量与污染物之间的关联度，可以验证其结果与逐步回归的结果是否一致。

4. 灰色关联分析方法模型的建立

灰色关联度分析法是灰色系统理论中一种定量描述因素间发展势态的相似或相异程度的量化比较方法。灰色关联度分析法的步骤与模型如下[7]。

分别选取大气污染物 SO_2 、 NO_2 和 PM_{10} 的日平均浓度数列为参考数列，气象参数数列为比较数列，其中参考数列记 $X_0(k)$ ，比较数列记为 $X(k)$ 。

由于气象参数数列中变量的量纲不同，为消除量纲对分析结果的影响，需要进行无量纲化处理。常用的有标准化、初始化、极差法和最大值化等，在此也选用极差法进行处理，公式如下：

令

$$x'_{ij} = \frac{x_{ij} - m_j}{M_j - m_j} \quad (i=1, 2, \dots, 55; j=1, 2, 3, 4) \quad (5)$$

其中 $M_j = \max_{1 \leq i \leq 55} \{x_{ij}\}$, $m_j = \min_{1 \leq i \leq 55} \{x_{ij}\}$ ($j=1, 2, 3$)，则 $x'_{ij} \in [0, 1]$ 是无量纲的指标观测值。

3) 计算关联度函数，公式为：

$$\eta_i(k) = \frac{\min_i \min_k |x_i(k) - x_0(k)| + \rho \max_i \max_k |x_i(k) - x_0(k)|}{|x_i(k) - x_0(k)| + \rho \max_i \max_k |x_i(k) - x_0(k)|} \quad (6)$$

其中， $|x_i(k) - x_0(k)|$ 为 $x_0(k)$ 和 $x_i(k)$ 第 k 个点的绝对误差； $\min_i \min_k |x_i(k) - x_0(k)|$ 为两级最小差；为分辨率， $0 \ll 1$ ，一般取 0.5；越大，分辨率越小，越小分辨率越大。

4) 计算关联度，其公式为：

$$r_i = \frac{1}{n} \sum_{k=1}^n \eta_i(k) \quad (7)$$

其中， r_i 即为 x_i 对 x_0 的关联度。

Table 1. Correlation coefficients between pollutant concentrations and meteorological parameters
表 1. 污染物浓度与气象参数的关联系数

污染物	温度	湿度	气压	风速	风向
SO_2	0.5653	0.5272	0.6406	0.8172	0.6567
NO_2	0.5531	0.5144	0.6232	0.7658	0.6523
PM_{10}	0.5904	0.5538	0.6742	0.7914	0.6594
$PM_{2.5}$	0.5628	0.5315	0.6326	0.7834	0.6551
O_3	0.5439	0.5202	0.6144	0.7984	0.6529
$CO_{2.5}$	0.6073	0.569	0.697	0.7931	0.6544

Table 2. Correlation coefficient of O₃ concentration with meteorological parameters and other pollutant concentrations
表 2. O₃ 浓度与气象参数和其余污染物浓度的关联系数

	SO ₂	NO ₂	PM ₁₀	PM _{2.5}	CO	温度	湿度	气压	风速	风向
O ₃	0.821522	0.765595	0.791107	0.784968	0.797863	0.583785	0.559918	0.666962	1	0.67377

灰色关联度分析法是灰色系统理论中一种定量描述因素间发展势态的相似或相异程度的量化比较方法[7]。它的基本思想是根据序列曲线几何形状的相似程度来判断其联系是否紧密。一般地，曲线越接近，相应序列之间的关联度就越大，反之就越小。根据上文计算出的污染物浓度与气象参数的关联系数和 O₃ 浓度与气象参数和其余污染物浓度的关联系数如表 1 和表 2 所示。

通过上述表格，可知对污染物浓度影响最大的参数为风速，这与逐步回归方法所得结果全吻合，而且还获得了其他影响不显著参数的关联系数。

5. 基于极限学习机神经网络模型的污染物预测

5.1. 模型建立

对于一个特定的时间维度来说，在不考虑各相互影响的情况下，时间维度越靠近的污染物浓度与气象数据越有利于预测当下时间点的空气质量。在按各污染物浓度在空气中含量进行计算时，可最大限度减小不确定因素天气状况带来的影响，从而寻找到可用的潜在规律。因此，模型的建立及求解可以分为以下四步，如图 6 所示。

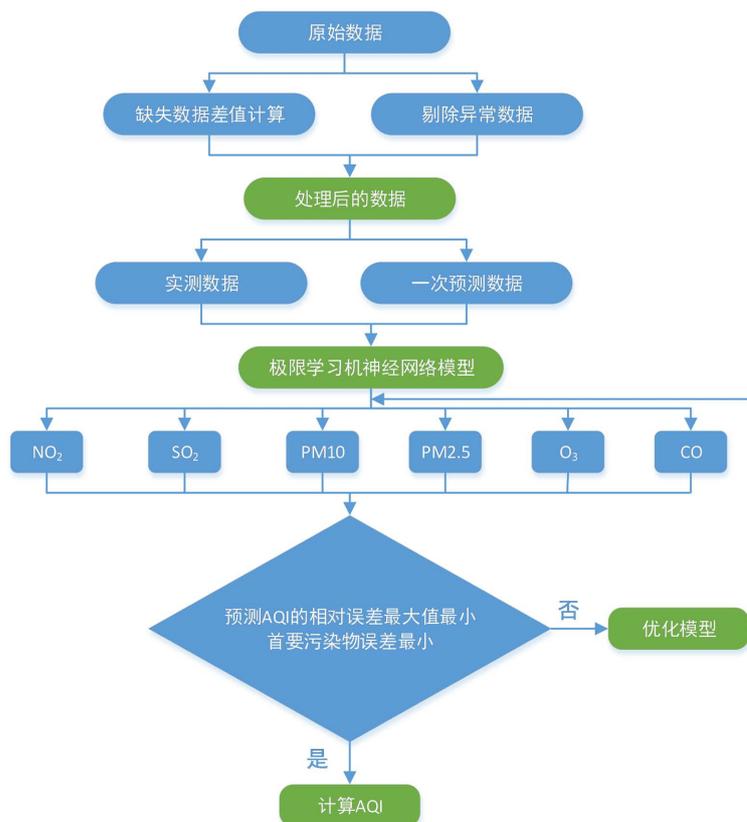


Figure 6. Model establishment and analysis diagram
图 6. 模型建立及分析图

5.2. 预测数据的选择

考虑到同一时间维度监测点的预测准确程度，与时间段越接近，一次预报的越准确。因此，首先利用编程软件寻找距离每天的 24 小时的预测结果最近的运行日来计算结果，当天没有运行预报的，自动向上一天运行日寻找当日的预报结果，以此来达到获得连续 1 小时一次划分的预报结果。

为了保证预测模型具有一定的鲁棒性，只采用一次预报数据和实测数据按时间对其的部分。采用 2020 年 7 月 23 日到 2021 年 7 月 13 日的数据。

5.3. 构造极限学习机神经网络模型

5.3.1. 极限学习机原理概述

典型的单隐含层前馈神经网络结构如图 7 所示，由输入层、隐含层和输出层组成，输入层与隐含层、隐含层与输出层神经元间全连接。其中，输入层有 n 个神经元，对应 n 个输入变量，隐含层有 1 个神经元；输出层有 m 个神经元，对应 m 个输出变量。为不失一般性，设输入层与隐含层间的连接权值 w 为[8]：

$$w = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{l1} & w_{l2} & \cdots & w_{ln} \end{bmatrix}$$

其中， w_{ij} 表示输入层第 i 个神经元与隐含层第 j 个神经元间的连接权值。

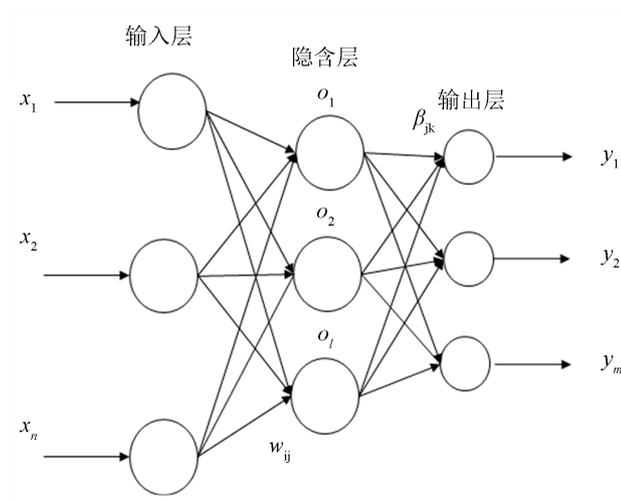


Figure 7. Extreme learning machine neural network model
图 7. 极限学习机神经网络模型

设隐含层与输出层间的连接权值为 β

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{l1} & \beta_{l2} & \cdots & \beta_{lm} \end{bmatrix}$$

其中， β_{jk} 表示隐含层第 j 个神经元与输出层第 k 个神经元间的连接权值。

设隐含层神经元的阈值 b 为：

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_l \end{bmatrix}$$

设具有 Q 个样本的训练集输入矩阵 X 和输出矩阵 Y 分别为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1Q} \\ x_{21} & x_{22} & \cdots & x_{2Q} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nQ} \end{bmatrix}$$

设隐含层神经元的激活函数为 $g(x)$ ，则由图 1 可得，网络的输出 T 为：

$$T = [t_1, \dots, t_Q]_{m \times Q}, t_j = [t_{1j}, \dots, t_{mj}]^T = \begin{bmatrix} \sum_{i=1}^l \beta_{i1} g(w_i x_j + b_i) \\ \sum_{i=1}^l \beta_{i2} g(w_i x_j + b_i) \\ \cdots \\ \sum_{i=1}^l \beta_{im} g(w_i x_j + b_i) \end{bmatrix}, (j=1, 2, \dots, Q) \quad (8)$$

上式可表示为： $H\beta = T'$ 。

其中， T' 为矩阵 T 的转置； H 称为神经网络的隐含层输出矩阵，具体形式如下：

$$H(w_1, \dots, w_l, b_1, \dots, b_l, x_1, \dots, x_Q) = \begin{bmatrix} g(w_1 \times x_1 + b_1) & g(w_2 \times x_1 + b_2) & \cdots & g(w_l \times x_1 + b_l) \\ g(w_1 \times x_2 + b_1) & g(w_2 \times x_2 + b_2) & \cdots & g(w_l \times x_2 + b_l) \\ \cdots & \cdots & \cdots & \cdots \\ g(w_1 \times x_Q + b_1) & g(w_2 \times x_Q + b_2) & \cdots & g(w_l \times x_Q + b_l) \end{bmatrix}_{Q \times l} \quad (9)$$

5.3.2. 优化模型的结果

由于客观现象存在不确定性和获取观测数据时受条件和环境的制约，从数据到模型往往要经过多次反复探索，同时为了所求 AQI 的相对误差最大值最小以及首要污染物误差最小，因而必须优化模型的结构。

Step1: 构造适应度，我们以 AQI 和首要污染物的误差这两指标的组合作为适应度，随机设置一组要优化的参数，这里是惩罚因子和核参数。

Step2: 把数据分为 153 天训练和剩余天数预测两部分，用训练样本和设置的参数取训练模型，然后用测试样本的输入与预测获得预测结果。统计 AQI 的最大相对误差和首要污染物的平均相对误差。

Step3: 以上面描述的适应度函数为适应度采用遗传算法中参数惩罚因子，核参数作为被优化的参数去优化[9]。优化的适应化结果，如图 8 所示。

5.3.3. 极限学习机神经网络模型求解

结合上文可知，在监测点的数据进行预处理之后，符合极限学习机神经网络模型要求，因此我们可以依照此模型求解，具体操作步骤如下：

Step1: 以当前时刻的一次预报量的全部数据为输入影响参数。

Step2: 以本小时之前五个时刻的实际预测数据为输入影响参数，假设当前要预测的时刻是 t ，那么 $t-1$ ， $t-2$ ， $t-3$ ， $t-4$ ， $t-5$ ，这 5 个历史时刻的实际采集值(包括污染物和天气)也要作为输入，获得监测点的训练结果如图 9 所示。

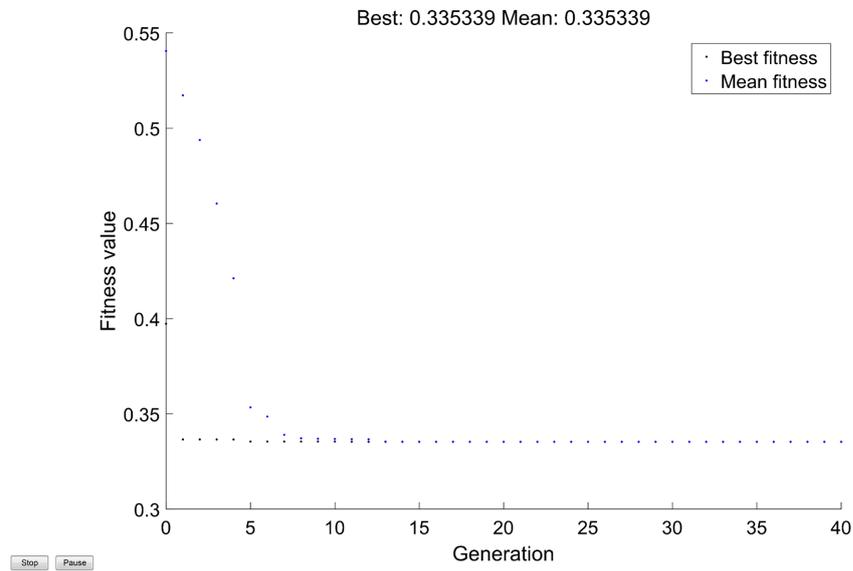


Figure 8. Optimal fitness change results
图 8. 优化的适应度变化结果

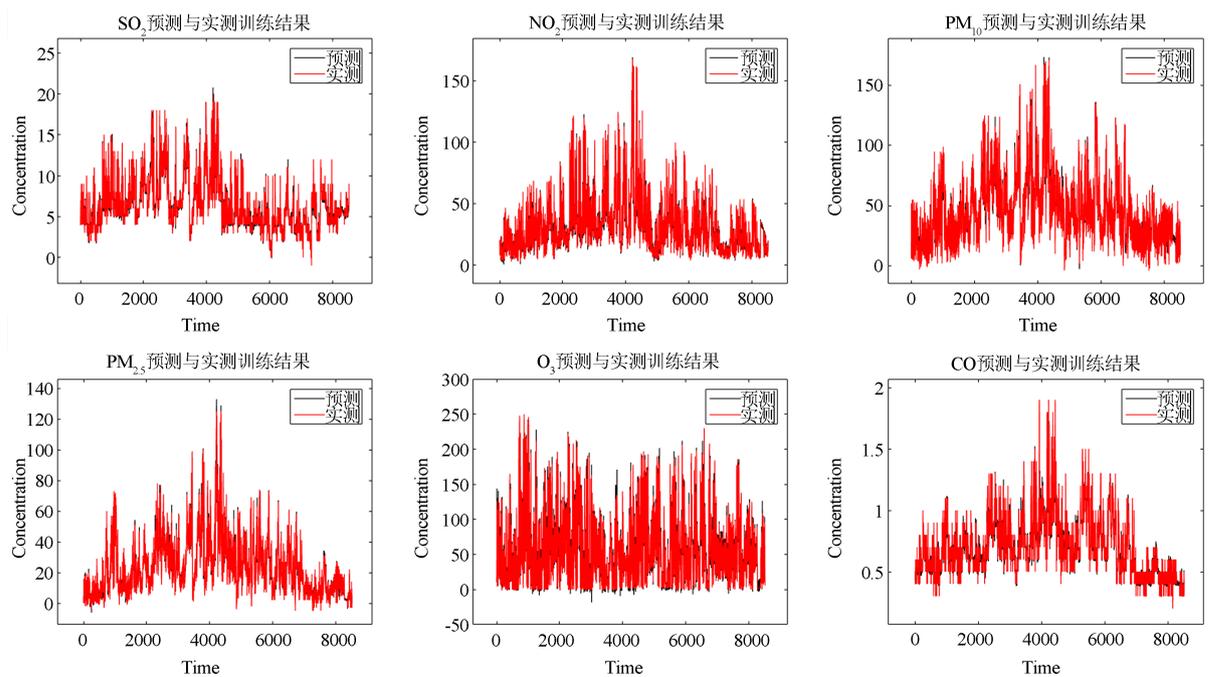


Figure 9. Training results of monitoring points
图 9. 监测点的训练结果

5.4. 基于滚动预测方法的数据处理方法

所谓滚动预测[10]是指通过添加最新的数据预测第二天的值。对于一个稳定的预测模型，不需要每天都去拟合，可以选择给他设定一个阈值，例如每周拟合一次，该期间只需通过添加最新的数据实现滚动预测即可。

我们在搭建了适应度函数，并将模型参数最优化后。考虑到实测已知数据不能直接运用到预测模型

中,为了在预测过程中得到有效合理的数据,我们运用滚动预测方法去深化解决问题,得到结果。由此可得监测点的预测结果,如图10所示。

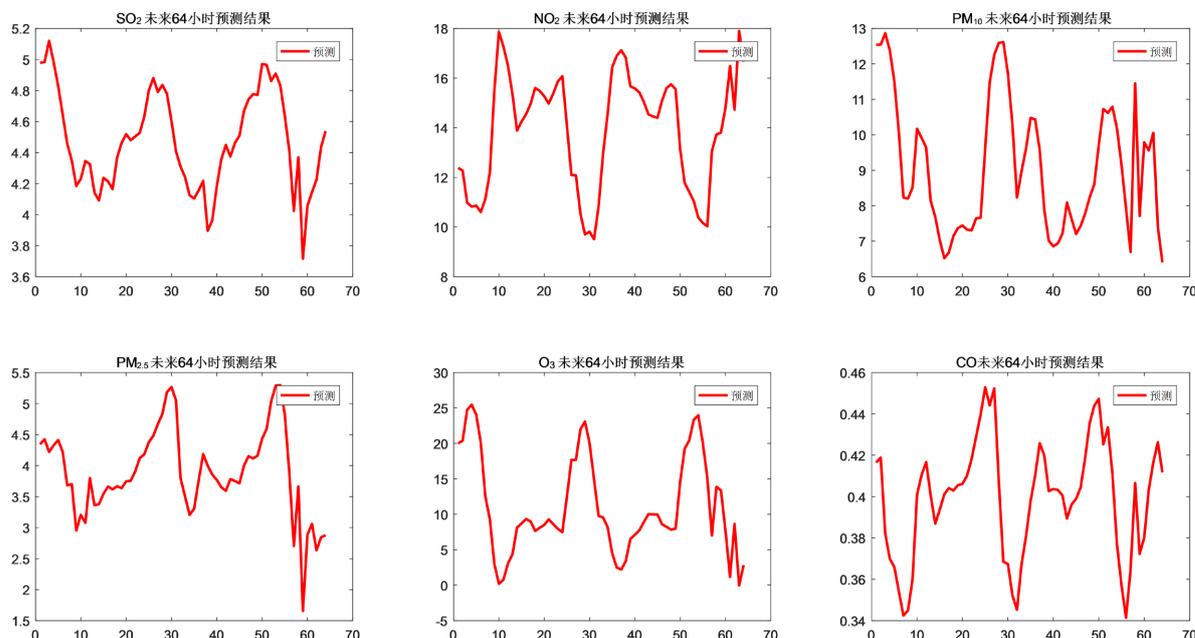


Figure 10. Prediction results of monitoring points

图 10. 监测点预测结果图

6. 小结

本文在一次预测模型的基础上建立的二次预报数学模型所到的结果比一次预测结果的准确性明显提高,同时本文对类似的相关预报研究具有明显的参考意义。本文的不足之处,在于计算量较大并且复杂,运用于其他情况时可以根据具体情况进行优化。

参考文献

- [1] 郝吉明, 马广大, 王书肖. 大气污染控制工程[M]. 北京: 高等教育出版社, 2010.
- [2] 伯鑫等. 空气质量模型(SMOKE、WRF、CMAQ等)操作指南及案例研究[M]. 北京: 中国环境出版集团, 2019.
- [3] 张天. 基于神经网络的空气质量预测预警系统的设计与实现[D]: [硕士学位论文]. 石家庄: 河北科技大学, 2019.
- [4] 唐之享. 基于BP神经网络的空气质量预测研究与实现[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2018.
- [5] 赵晓阳. 基于神经网络的空气质量预测模型构建研究[D]: [硕士学位论文]. 包头: 内蒙古科技大学, 2020.
- [6] 高淑新, 李若楠, 吴佳丽, 李一鸣, 孟微, 孟凡帅, 陈鹏心. 新民市温度幅度等差值法订正乡镇温度预报[J]. 农业灾害研究, 2017, 7(Z3): 21-22. <https://doi.org/10.19383/j.cnki.nyzhyj.2017.09-10.010>
- [7] 余胜威. 数学建模经典案例实战[M]. 北京: 清华大学出版社, 2014.
- [8] 李小冬. 核超限学习机的理论与算法及其在图像处理中的应用[D]: [博士学位论文]. 杭州: 浙江大学, 2014.
- [9] 崔珊珊. 遗传算法的一些改进及其应用[D]: [硕士学位论文]. 合肥: 中国科学技术大学, 2010.
- [10] 任守纲, 刘鑫, 顾兴健, 王浩云, 袁培森, 徐焕良. 基于R-BP神经网络的温室小气候多步滚动预测模型[J]. 中国农业气象, 2018, 39(5): 314-324.