**基于误差质量控制图的空气污染指数预测模型研究[[1]](#footnote-1)**

 王强 冀文彦 胡雅芬

1. 北京城市学院首都城市环境建设研究基地 北京 100083；

2.北京城市学院首都城市环境建设研究基地 北京 100083；

3.北京城市学院城市执法研究所 北京 100083)

摘要：空气质量预测模型的预测效果往往在事实发生之后才能判断，如何在预测前判断模型的预测效果是保证预测精度的前提。本研究将全面质量管理中质量控制图的思想用于提前判断空气质量预测模型的预测效果。研究首先构建了空气质量神经网络预测模型；其次，基于神经网络模型的历史数据构建误差质量控制图；最后采用北京市空气污染指数的实际数据对模型进行验证和讨论。本研究设计的误差质量控制图可以有效判断预测模型在下一个预测点的预测效果，在此基础上对模型进一步调整可以大大提高模型的预测精度和可靠性。

关键词：误差质量控制图；神经网络；空气污染指数预测

中图分类号：G922 文献标志码: A 文章编号：

**一、引言**

由于城市空气的质量水平越来越受到人们的关注，我国开始采用空气污染指数向社会公众发布空气质量信息。空气污染指数(API)一般利用空气污染程度分级表征空气质量状况和发展变化趋势。空气污染指数的非线性特征非常明显，所以基于神经网络的预测方法效果比较好。目前用于空气污染预报的神经网络模型以MLP 模型为主。但是, 这种算法不能考察样本选择对系统学习的影响，而且收敛速度很慢、缺乏良好的网络范化能力, 所以这种算法预报效果精度较差，因此很多学者希望改进神经网络算法。文献[1]运用BP 神经网络算法原理,建立了城市环境空气质量预测模型,并对该模型的泛化能力进行了误差评价。文献[2]进一步提出了MLP 模型的改进方法，并且分析了API 观测数据误差、气象预报数据等因素对污染指数的预报都的影响。文献[3]将自适应调节动量值算法引入传统神经网络模型, 其预测精度进一步提高。针对观测数据少的情况，文献[4]将灰色系统同神经网络结合，提出有机灰色神经网络空气污染预报模型。文献[5]对这类模型进一步改进，通过遗传算法寻优来确定神经网络层次结构，改进的模型在沈阳市冬季空气中二氧化硫日均含量预报的中取得了较好的精度。文献[6]将数值逻辑方法和模糊逻辑方法进行结合, 构建模糊神经网络模型并应用于NO2 浓度预报中。文献[7、8、9] 将主成分分析技术应用于RBF 网络进行氮化物浓度预测。文献[10]将BP模型中原来的性能函数调整，大大减小了对应新输入预测数据的误差。文献[11]开发了一种适应性的ARBF模型, 该模型解决了网络训练过程中隐层节点确定问题，可以根据实际情况调节隐层节点个数。

上述预测模型虽然在各自的环境中可以实现准确预测，但是考虑现实世界问题的复杂性，没有任何一种空气污染指数预测模型在任何情况下都是最好的[12]。因此，在实际预测过程中，人们总是选择历史预测精度最好的模型进行预测，但是这些模型在未来的预测精度仍然无法判断[13，14，15]。因此如何判断预测模型预测结果的精度变化趋势变得非常重要。全面质量管理中质量控制图可以有效判断质量的变化趋势[16]，如果将二者结合，则可能用它来评估预测模型在未来的表现。

综上所述，本研究利用空气污染指数历史数据建立空气污染指数神经网络预测模型，将该模型作为一台“生产机器”考虑，其输出结果的误差就是这台机器的“产品”。在此基础上， 我们利用模型训练和预测中的误差数据建立误差质量控制图，进一步检测预测误差的变化趋势，并以此判断该模型在未来的预测中是否受控。

**二、基于误差质量控制图的预测原理**

依照Kolmgorol定理，利用一个三层的人工神经网络能够以任何的误差精度模拟一条任意的非线性曲线。因此在本研究中我们选择一个三层的人工神经网络来建立空气污染指数神经网络预测模型。

对于一个人工神经网络预测模型，其每一个预测结果都有一个相对误差。研究过程中，我们把这样的相对误差看作是预测模型的产品，然后模仿生产控制的方法，利用误差质量控制图来调整控制预测过程。本文所采用的相对误差计算方法如下:

 1

其中是预测数据，是真实数据。

根据上面提到的Kolmgorol定理，本文中的空气污染指数神经网络预测模型的相对误差符合均值为0的近似正态分布，这也是本文构建误差质量控制图的依据。具体的控制调整步骤如下：首先，创建空气污染指数神经网络预测模型；然后，构建误差控制图；最后，依据控制图判定模型是否受控。如果预测模型精度变化趋势在可控范围，它将直接用于下一阶段的预测点，如果控制图显示预测精度有变坏的趋势，则对模型做相应的调整，然后再用于预测。本预测模型的流程图如下图1：

空气污染指数API历史数据

ANN预测模型

误差控制图

预测

如果 ANN模型不受控

如果 ANN模型受控

图1 预测模型工作流程图

**三、基于误差质量控制图的空气污染指数预测模型**

本研究利用北京市空气污染指数的实际数据对预测方法进行验证。2015年1月1日到2015年6月30的空气污染指数逐日收集。其它输入数据，如实际气候信息从当地的环保气象站收集。

建立和训练神经网络预测模型的数据来自于 2015年1月1日到2015年5月31日之间，共计3672个数据点，这些样本用来估计预测模型的参数。

2015年6月的数据用于模型的测试和分析，每一天会有 24点，共720点。为了定量地确定准确性最好的模型，使用以下的MAPE对所有这些模型进行比较：

|  |  |
| --- | --- |
|   | 2 |

其中是预测数据，是真实数据，N是预测的点数。

**（一）空气污染指数预测模型**

空气污染指数神经网络预测模型的输入数据多是气象观测数据和空气污染数据，其输出是空气污染指数。考虑到模型的训练和精度调整, 只有充足的数据样本才能保证模型预测精度。

本研究选择包含三层的人工神经网络[17]作为基本的预测模型。对于输入层，输入由两部分组成。第一部分是3个预测点的天气数据：10m高处的风速、方向和温度，第二部分是预测点之前的n个空气污染指数API的历史数据。为了确定最佳的输入神经元的个数，第二部分输入的神经元数量从1至6逐渐增加。对于中间层（隐含层）中神经元的数量目，也采用逐一试验的方法，由3个神经元开始，逐渐增加至8，以确定最佳的隐含层神经元的数目。该模型输出层只有提前一小时空气污染指数API一个神经元。本研究总共提出11个神经网络模型作为建模的基础，表1中列举了这些模型，这些模型的输入层和隐含层中神经元个数不同。

模型的其它参数如下：Momentum rate 0.18;Study rated 0.1;Training epochs 1200。网络训练算法是Levenberg-Marquardt算法，当验证误差（相对误差）开始增长时，说明模型精度达到最优，结束训练过程。

这11个基础预测模型可被看作11个生产相对误差的机器，这些误差的变化趋势可以用相应的“质量控制图”来表示。如果误差变化趋势受控，则他们可以用来预测下一阶段的空气污染指数；否则，模型将进行调整。

表1 11个神经网络模型

|  |  |  |
| --- | --- | --- |
| 输入神经元数量不同的模型 |  | 隐含层神经元数量不同的模型 |
| 模型 |  | 输入层数量 | 隐含层数量 | 模型s | 输入层数量 | 隐含层数量 |
| ANN -1 |  | 3+1 | 5 | ANN -7 | 3+4 | 3 |
| ANN -2 |  | 3+2 | 5 | ANN -8 | 3+4 | 4 |
| ANN -3 |  | 3+3 | 5 | ANN -4 | 3+4 | 5 |
| ANN -4 |  | 3+4 | 5 | ANN -9 | 3+4 | 6 |
| ANN -5 |  | 3+5 | 5 | ANN -10 | 3+4 | 7 |
|  |  |  |  |  |  |  |
| ANN -6 |  | 3+6 | 5 | ANN -11 | 3+4 | 8 |

备注: 3+n 表示3个气象指标和n个之前的历史数据。

**（二）误差控制图**

在空气污染指数API人工神经网络预测模型的训练过程中，每一组历史数据都将对应一个预测误差。本研究“初始化”设置将在2015年1月1日到2015年5月31日之间的历史数据中随机抽取25天，其中每一天的历史数据都将被视为一个子样本集。对于其中的任何一个子样本集，随机抽样5个预测时间点（例如2:00，7:00，12:00，17:00，22:00）作为这一子样本集的样本。研究总共有25个子样本集，125个预测时间点。

绘制质量控制图所需的上限和下限的计算方法如下。其主要参数包括各组的极差和各组平均。

|  |  |
| --- | --- |
|   | 3 |
|   | 4 |
|   | 5 |
|   | 6 |

其中是子样本集的的大小，是第个子样本集的均值，是子样本集合数量，是第个子样本集合的第个样本，是第个子样本集合的极差，是第个子样本集合最大的样本，是第个子样本集合最小的样本。

质量控制图中控制线的计算按照下表2所列的公式进行。

表2 控制限计算

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **-Chart**  |  | **-Chart**  |
| 中心线 |  |  |  |  |
|  |  |  |  |  |
| 控制上限 |  |  |  |  |
|  |  |  |  |  |
| 控制下限 |  |  |  |  |

注：，三个系数是与子样本集合容量相关。根据国标GB/T4091-2001, 当子样本集合的容量为5时时，，。

**（三）X-R控制图控制过程**

对于任意一个需要预测的时间点，我们将在它之前24小时内的预测结果中随机抽取5个组成一个子集，这个子集中相对误差所对应的平均值和极差将用来绘制误差质量控制图中的误差趋势线，并用于X-R误差控制图的检查。

本文根据国标GB/T4091-2001确定预测误差的变化是否在正常范围内。只有当这些预测误差同时满足以下两个条件时，才能保证预测误差在受控的状态。第一个条件是“没有超出控制限的点有连续25个；或是在35个连续的点中仅有1个点超出控制限；或是在连续的100个点中，超过控制线的点少于2。落在控制线上的点被视为超出控制线”。第二个条件是要保证控制图没有出现如下异常分布：

1.连续分布在中心线同一侧的点大于等于有7个;

2.保持同一趋势的点不能连续超过7个，增长或是下降。

3.在11个连续点中，在中心线同一侧的等于或超过10个;

4.在14个连续点中，在中心线同一侧的等于或超过12个;

5.在17个连续点中，在中心线的同一侧等于或超过14个;

6.在20个连续点中，在中心线的同一侧等于或超过16个;

7.点有规律性或是周期性波动。

根据上述原则，如果预测模型的误差是在可控范围内，则该模型在下一个时点的预测误差将比较小，预测模型可用。反之，该模型将采用新样本（删除部分历史数据或是添加新数据）重新进行神经网络模型的训练，调整预测模型的精度。

**四、预测结果和分析**

表3所示为传统人工神经网络模型和误差控制图预测模型的相对误差。（在2015年6月的30天中，每天24个观测点，共计24X30点）结果如下。

表3 ANN 模型和误差控制图预测模型的MAPE

|  |  |
| --- | --- |
| ANN组模型 | 误差控制图预测组模型 |
| 模型代号 |  | 相对误差 |  | 模型代号 |  | 相对误差 |  | 调整次数 |
| ANN -1 |  | 0.2184 |  | XR-ANN -1 |  | 0.1829 |  | 6 |
| ANN -2 |  | 0.2087 |  | XR-ANN -2 |  | 0.1917 |  | 8 |
| ANN -3 |  | 0.1826 |  | XR-ANN -3 |  | 0.1719 |  | 4 |
| ANN -4 |  | 0.1795 |  | XR-ANN -4 |  | 0.1671 |  | 6 |
| ANN -5 |  | 0.1984 |  | XR-ANN -5 |  | 0.1752 |  | 5 |
| ANN -6 |  | 0.2184 |  | XR-ANN -6 |  | 0.1811 |  | 6 |
| ANN -7 |  | 0.2063 |  | XR-ANN -7 |  | 0.1869 |  | 8 |
| ANN -8 |  | 0.1823 |  | XR-ANN -8 |  | 0.1706 |  | 3 |
| ANN -9 |  | 0.1752 |  | XR-ANN -9 |  | 0.1548 |  | 6 |
| ANN -10 |  | 0.1952 |  | XR-ANN -10 |  | 0.1768 |  | 5 |
| ANN -11 |  | 0.2004 |  | XR-ANN -11 |  | 0.1868 |  | 7 |

表3中所有的误差控制图预测模型都比普通的人工神经网络模型的相对误差小，这说明误差控制图预测模型的预测精度高于传统的网络预测模型，采用误差控制图预测模型在可靠性和准确性上有很大提高。本研究中相对误差最小的模型是XR-ANN-9。在整个预测过程中，该模型总计调整6次，部分预测结果如下：

模型训练完成后，XR-ANN-9模型历史数据上的相对误差分布如图2：



图2 X-R-ANN-9 相对误差分布

它是一个均值为0、均方差为0.61的近似正态分布。它所对应的误差质量控制图如表4和图3所示：

表4 X-R-ANN-4模型控制限

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **-Chart**  |  | **-Chart**  |
| 中心线 |  |  |  |  |
| 控制上限 |  |  |  |  |
| 控制下限 |  |  |  |  |

XR-ANN-9模型预测过程中共6个调整点，1）6月8日中午12点；2）6月17日上午03时；3）6月23日下午6时； 4）6月26日下午6时，日；5）6月28日上午12时；6）6月30日上午12点。从表5中的结果可以得看出，利用误差控制图对预测模型进行调整后，预测模型的预测精度显着增加。

表5 XR-ANN-9模型调整前后误差对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 预测点 | 1 | 2 | 3 | 4 | 5 | 6 |
| 原有模型 | 0.179 | -0.190 | -0.246 | 0.199 | 0.256 | 0.274 |
| 调整后的模型 | 0.135 | -0.116 | 0.114 | -0.098 | -0.115 | -0.121 |

第二个调整时点是6月17日中午12点，在改点预测之前，其误差控制图的状态如图3 (a)和(b) 所示。

依据国标GB/T4091-2001，该误差质量控制图中的控制点具有以下异常分布：在中心线的同一侧连续分布的点有7点，因此该预测模型的预测误差没有在受控状态，其精度变化趋势趋坏，需要在进行预测之前进行调整。因此我们相应减少了从2015年1月1日至10日的旧数据，并重新训练神经网络预测模型。预测模型调整后的相对误差为13.6%，远远小于原来模型的17．8%。这说明该误差质量控制图可以有效的控制误差，提高模型的预测精度。



(a) X-Charts



(b) R-Chart

图3 预测质量控制图

从图4得知，当输入层神经元的个数为7时（3个气象数据和４个先前的空气污染指数API），传统的人工神经网络模型和误差控制图预测模型的预测精度都达到最优值。继续增加或是减少输入层神经元的数量会降低预测精度。



图4 调整次数和输入神经元之间的关系

图5说明当隐含层神经元数量为6的时候，传统神经网络模型和误差控制图预测模型的预测精度都达到最优值。继续增加或是减少输入层神经元的数量都会降低预测精度。



图5 预测精度和隐含层神经元个数间的关系



图6 模型MAPE和调整次数间的关系

图6说明预测模型的调整次数跟随模型的相对误差进行变化。在开始阶段，模型的调整次数随着相对误差的增加而减少，当调整次数达到最小时，模型的调整次数却随着相对误差的增加而增加。这说明相对误差越小，模型的精度越高，但是模型的鲁棒性降低。但是当相对误差很大时，预测模型的预测精度和鲁棒性都降低，预测过程中模型的调整次数会增加。

**五、结论**

本研究将全面质量管理中质量控制图的方法应用于人工神经网络空气污染指数API预测模型中，提出了一种误差控制图空气污染指数预测模型。通过分析监测预测模型误差在质量控制图中的分布模式，判断模型的预测误差变化趋势是否在可控的范围内。实际数据结果表明误差质量控制图可以用来判断预测模型预测精度的变化趋势，并作为模型调整的依据，从而大大提高模型的可靠性和准确性。

神经网络模型中输入层和隐藏层中神经元数目对模型的精度和调整次数都有影响。以北京市空气污染指数API的历史数据为例，当输入层神经元系数为7，隐含层的神经元为6时，模型的预测精度最高。增加或减少人工神经网络模型的神经元个数，都会使模型的精度降低。

误差控制图预测模型调整次数跟随模型的相对误差进行变化。模型的调整次数随着相对误差的增加先是减少，当调整次数达到最小时，开始增大，因此需要在调整次数和相对误差之间找到一个最佳平衡点。

**参考文献：**

[1]张鹏达.基于BP 神经网络的城市环境空气质量预测模型[J].自动化技术与应用，2014,33 （1）：9-19.

[2]J IANG D H, ZHANG Y, HU X, et al. Progres s in developing an ANN model for air pollution index forecast[J ]. Atmospheric Environment, 2004(38): 7055- 7064.

[3]吴小红, 康海燕, 任德官. 基于神经网络中小城市空气污染指数预估器的设计[J]. 数学的实践与认识, 2005, 35(2): 87- 91.

[4]郎君, 苏小红, 周秀杰. 基于有机灰色神经网络模型的空气污染指数预测[J]. 哈尔滨工业大学学报, 2004, 26 (12): 1598 - 1601.

[5]王俭, 胡筱敏, 郑龙熙, 等. BP 模型的改进及其在大气污染预报中的应用[J]. 城市环境与城市生态, 2002, 15(5): 17- 19.

[6]NEAGU C - D, AVOURIS N , KALAPANIDASE, et al. Neural and neuro- fuzzy integration in a knowledge - based sys tem for air quality prediction[J ]. Applied Intelligence,2002(17): 141- 169.

[7] LU W Z, WANG W J , WANG X K, et al.Us ing improved neural network model to analyze RSP, NOX and NO2 levels in urban air in Mong KOK,Hong Kong [J ]. Environmental Monitoring and As ses sment, 2003 (87): 235- 254.

[8]吴有训, 彭慕平, 刘勇. 基于径向基函数网络的宣城市空气质量预测[J].安徽师范大学学报( 自然科学版) ，2 0 1 1，34 （4）.

[9]李翔. 基于ＧＡＢ和模糊ＢＰ神经网络的空气质量预测[J].华中科技大学学报（自然科学版）, 2013，41（增刊Ⅰ）.

[10]黄世芹. 改进BP 神经网络在城市环境大气污染分季节预报中的应用[J ]. 贵州气象,2005, 29(3):6- 8.

[11] WANG W J , LU W Z, WANG X K, et al. Prediction of maximum daily ozone levelus ing combined neural network and s tatis -tical characteris tics [J ]. Environment International,2003(29): 555- 562.

[12] Chatfield, C.; Yar, M. Holt-Winters Forecasting - Some Practical Issues. Statistician 1988, 37, 129-140.

[13] Bracale, A.; Carpinelli, G.; Proto, D.; Russo, A.; Varilone, P. New Approaches for Very Short-term Steady-State Analysis of An Electrical Distribution System with Wind Farms. Energies 2010, 3, 650-670.

[14]Shirland, L.E. Statistical quality control with microcomputer applications; Wiley: 1993; p.^pp.

[15] Hart, M.K.; Hart, R.F. Quantitative methods for quality and productivity improvement; ASQC Quality Press Milwaukee, Wisconsin: 1989; p.^pp.

[16]Kariniotakis, G.; Stavrakakis, G.; Nogaret, E. Wind power forecasting using advanced neural networks models. Energy conversion, ieee transactions on 1996, 11, 762-767.

[17] Baldi, P.; Hornik, K. Neural Networks and Principal Component Analysis - Learning from Examples without Local Minima. Neural Networks 1989, 2, 53-58.

**Air Pollution Index Forecasting ModelBased on Error QualityControl Charts**

Wang Qiang, Ji Wenyan HU Yafen

(1．capital city environment construction in Beijing city university research base Beijing 100083;2．capital city in Beijing city university and the research of environmental construction Beijing 100083;3．of the city of Beijing city law enforcement institute Beijing 100083)

**Abstract:** Usually the performance of air pollution index (API) forecasting model can only beevaluatedafter it is happened. Therefore, it is necessary to judge the performance offorecasting models for a certain point in the future in terms of accuracy. The quality control charts in total quality management is used to to judge the performance of the forecasting model in this paper. Firstly, an artificial neuron network (ANN) forecasting model is designed; then, the error quality chart is created on the basis of history error data; finally, The data of Beijing cityis used to test the performance of the model. It shows that the model can adjustitself according to its error quality charts before each time APIforecasting. This makesthe model excellent in terms of accuracy and reliability.

**Keywords:**Error Quality Control Charts; Artificial Neuron Network ; Air Pollution Index Forecasting;

责任编辑:郝峻弘

1. 收稿日期：2016年08月01日

作者简介：王强（1973-），男（汉），河北保定市人，讲师，博士，北京城市学院首都城市环境建设研究基地研究员，研究方向为城市管理。冀文彦（1982年—），女（汉），内蒙古包头人，讲师，硕士，北京城市学院科技与产业发展部主任助理，首都城市环境建设研究基地副主任，研究方向为城市管理，环境治理。胡雅芬（1983-），女（汉），江西新余市人，副教授，博士，北京城市学院城市执法研究所副所长，研究方向为城市经济。

基金项目：首都城市环境建设研究基地“北京市城乡结合部环境问题综合治理研究 ”项目资助，项目编号：16JDGLB008。 [↑](#footnote-ref-1)