文章编号：1671-6612（2020）06-676-06

基于自回归 LSTM 神经网络的地铁车站环境参数概率预测

曲洪权[1]　李　博[1]　庞丽萍[2]　梁思远[1]

（1.北方工业大学信息学院北京　100144；

2.北京航空航天大学航空科学与工程学院北京　100191）

【摘　要】　在地铁车站这种人员密集的场景中，预测未来一段时间内的环境参数变化对于列车正常运行和乘客安全有重要意义。和传统的点预测方法相比，提出了一种基于自回归循环 LSTM 网络的概率预测方法。不同于点预测，概率预测给出预测变量在下一时刻的概率密度函数，考虑了地铁车站环境参数预测的不确定性，对于站内提前、可靠对紧急情况做出反应有重要的意义。提出的预测模型，将预测变量的历史数据和车站外部环境参数作为输入变量来预测的环境参数的下一时刻的概率密度函数，进而得到下一时刻该环境参数的变化范围和分位数等信息。为了验证所提方法的准确性，收集了 3 个地铁站的环境参数数据，并使用概率预测方法进行了预测。结果显示，提出的方法不仅可以预测最可能的环境参数预测结果，而且可以预测极端情况的发生概率，对于预防紧急事故和决策有重要意义。

【关键词】　地铁车站；环境参数；概率预测；LSTM 网络

中图分类号　X820　　文献标识码　A

Probabilistic Forecasting of Metro Station Environmental Parameters Based on Autoregressive LSTM Network

Qu Hongquan[1]　Li Bo[1]　Pang Liping[2]　Liang Siyuan[2]

(1. School of Information Science and Technology, North China University of Technology, Beijing, 100144;

2. School of Aeronautic Science and Engineering, Beijing, 100191)

【**Abstract**】　In the metro station which is crowded with people, it is very important to predict the changes of environmental parameters in the future for the normal operation of trains and the safety of passengers. Compared with the traditional point prediction method, this paper proposes a probabilistic forecasting method based on autoregressive LSTM network. Different from the point prediction, the probabilistic forecasting model gives the probability density function of the prediction variable at the next moment, and calculates the uncertainty of the environmental parameter's prediction, which is of great significance for the station to respond to the emergency in advance. The proposed model in this paper uses the historical data of the prediction variables and the external environmental parameters as the input variables to predict the probability density function of the environmental parameters

基金项目：国家重点研发计划（2017YFB1201100）资助项目；辽宁省“兴辽英才计划”项目资助（XLYC1802092）
作者简介：曲洪权（1973-），男，博士，教授，研究方向为数据科学，E-mail：qhqphd@ncut.edu.cn
通信作者：庞丽萍（1973-），女，博士，教授，研究方向为人机与环境工程，E-mail：pangliping@buaa.edu.cn
收稿日期：2020-03-19

at the next time, so as to obtain the information of the changing range and quantiles. In order to verify the accuracy of the method, we collected the environmental parameter data of 3 subway stations, and used the probabilistic forecasting method to predict. The results show that the method can not only predict the most likely environmental parameters prediction results, but also predict the probability of extreme cases, which is of great significance for the prevention of emergency accidents and decision-making.

【Keywords】 metro station; environmental parameters; probabilistic forecasting; LSTM network

0　引言

在城市地区，地铁是解决交通拥堵问题的最有效的公共交通方式之一，同时，乘客数量随着地铁的发展不断增加[1-3]。由于地铁车站大部分处于地下空间，通风条件差，污染物容易在站内沉淀，对旅客健康造成不良影响，例如 NH_3，VOCS，NO_2 等。因此，有必要分析地铁站的环境趋势，并建立一个相对准确的模型来预测地铁站的环境参数[4]。

近年来，神经网络因其强大的非线性拟合能力的优势被广泛应用在环境参数预测研究中，这种方法不需要大多数关于过程机制的基本知识，通过实验数据即可建模[5-8]。许多研究者们使用神经网络建模预测环境参数，并取得了很好的效果[9-12]。Kamal 等人证明人工神经网络(ANN)可以简化和加快环境空气质量的计算[13]。Bodri 等人使用 ANN 模型预测一天的地面气温（SAT），模型预测结果与实测数据非常接近[14]。Kim 等人的研究结果表明，在室内空气质量预测中，与其他数据驱动的预测模型相比，递归神经网络（RNN）模型可以提供更好的建模性能和更高的可解释性，并证明了关键变量选择的重要作用[15]。Lim 等人提出了一种新的关键变量选择方法，并进一步揭示了关键变量对预测的重要性[16]。Qu 等人开发了一种基于滑动时间窗的随机矢量功能链接神经网络（RVFLNN）的建模方法，并解决了大数据计算速度慢的问题[17]。

上述方法的预测目标是在每个时间步骤中预测一个准确的值。但是，实践中的结果可能会受到许多因素的影响，预测环境参数的概率分布可能更合理[18-20]。Aznarte 等人研究了利用分位数回归法预测 NO_2 极端浓度，并改进了概率预测方法[21]。对于地铁站内环境而言，环境参数的概率预测也具有重要作用。通常，预测地铁环境参数的目的是判断未来一段时间内的污染浓度是否超过最大允许浓度，更加关注于污染物超标的发生概率。因此，本文提出了一种基于自回归 LSTM 网络的概率预测方法，输入外部参数和预测参数的历史数据，预测地铁站内环境参数的概率分布，为地铁车站的环境控制提供重要支持。

1　地铁车站现场数据采集与处理

1.1　设备简介

地铁车站现场数据采集使用的实验设备为 CPR-KA 空气质量监测仪，如图 1 所示。该设备采用泵吸式采样方式，采样方式为 300ml/min，数据记录间隔为 2 分钟/次，共监测 6 种参数，测量范围和分辨率如表 1 所示。

图 1　一体化综合监测设备

Fig.1　Integrated monitoring equipment

表 1　设备参数

Table 1　Device parameters

参数名称	测量范围	分辨率
SO_2	0～2000 ppb	1 ppb
NO_2	0～2000 ppb	1 ppb
VOC	0～10 ppm	1 ppb
PM_{10}	0～0.5 mg/m^3	0.001 mg/m^3
Temperature	−50～80℃	0.1℃
RH	0～100% RH	0.8% RH
CO	0～50 ppm	0.01 ppm
NH_3	0～30 ppm	0.1 ppm

CO_2	0～5% vol	0.01% vol

1.2　地铁测试车站类型及测试时间

（1）实验选定 3 个车站作为测试站点，如表 2 所示，分别为 Station1、Station2、Station3，并且为了保证测试结果的多样性，我们选定的车站包含普通站和换乘站，全高屏蔽门和半高屏蔽门，每个车站的类型均不相同。测试期间，每个车站测试时间为 1 天，一体化综合监测设备均放置在站台中间，距离地面高度为 1.2m，如图 2 所示。

表 2　测试站点

Table 2　Measured metro station

车站	测试时间	车站类型
Station1	2019-6-13	普通站，半高屏蔽门
Station2	2019-6-14	换乘站，全高屏蔽门
Station3	2019-6-20	普通站，全高屏蔽门

（2）为了分析车站环境参数的变化以及其影响因素，我们收集了测试当天的车站客流量和列车发车频率（由地铁运营公司提供）以及大气气象数据（室外大气温度，室外大气相对湿度，来自中国气象数据网 http://data.cma.cn/）。大气环境数据（室外 PM10，室外 CO，室外 NO_2，室外 SO_2，来自空气质量历史数据 http://beijingair.sinaapp.com/），总计 8 种外部变量。

图 2　设备放置地点

Fig.2　Measured position in the platform

图 2 中，左图红色点标记了设备的摆放位置，右图为车站现场照片。经过以上所有测试，我们获得 1260 组观测值，时间间隔 2 分钟，包括 9 种站内环境参数，8 种外部影响参数，并对所有数据进行预处理操作，包括补充缺失值，异常值处理，归一化以及去噪，保证数据的可靠性。

2　概率预测模型

2.1　网络模型

将站内环境参数未来时刻的预测视作构建一个条件分布，本文提出的模型可以用如下公式（1）表示：

$$P(Y_{t_0+1:t_0+\tau}\,|\,Y_{1:t_0}, X_{1:t_0+\tau};\Phi) \tag{1}$$

式（1）式中，t_0 是分割过去时刻和未来的时间点；τ 是预测范围的长度；$Y_{t_0+1:t_0+\tau}$ 和 $Y_{1:t_0}$ 分别属于 $[t_0+1:t_0+\tau]$ 和 $[1:t_0]$ 时间范围内的环境参数值；$X_{1:t_0+\tau}$ 是 $[1:t_0+\tau]$ 范围内的外部变量值；Φ 表示模型的参数。

在公式（1）中，整个时间序列 $[1:t_0+\tau]$ 被时间点 t_0 分为两部分，分别是 $[1:t_0]$ 和 $[t_0+1:t_0+\tau]$。$[1:t_0]$ 为条件区间，包含过去的信息，$[t_0+1:t_0+\tau]$ 称为预测区间。概率预测模型利用预测变量和外部变量过去的信息来预测未来值。

针对于每一个时间点的预测而言，模型可以写成如下公式（2）的形式：

$$\begin{aligned} &P(Y_{t_0+1:t_0+\tau}\,|\,Y_{1:t_0}, X_{1:t_0+\tau};\Phi) \\ &=\prod_{t=t_0+1}^{t_0+\tau} P(Y_t\,|\,Y_{t-1}, X_t;\Phi) \\ &=\prod_{t=t_0+1}^{t_0+\tau} l(Y_t\,|\,\theta(\mathbf{h}_t,\Phi)) \end{aligned} \tag{2}$$

$$\mathbf{h}_t = h(\mathbf{h}_{t-1}, Y_{t-1}, X_t, \Phi) \tag{3}$$

式（3）式中，$\mathbf{h}_t$ 是自回归 LSTM 网络的输出；h 代表 LSTM 网络；Y_t 是环境参数 Y 在时刻 t 的取值；$l(\cdot)$ 是用来拟合预测变量分布的似然函数；$\theta(\cdot)$ 是计算似然函数参数的函数。

由于模型是自回归结构，网络的前一时刻输出 $\mathbf{h}_{t-1}$ 与上一时刻预测变量的观测值 Y_{t-1} 作为下一时刻的输入。似然函数 $l(Y_t\,|\,\theta(\mathbf{h}_t,\Phi))$ 为一个固定分布，参数由函数 $\theta(\mathbf{h}_t,\Phi)$ 以及网络输出 $\mathbf{h}_t$ 决定。本文中将似然函数的分布确定为高斯分布，如公式（4）所示，参数 $\theta=(\mu,\sigma)$ 分别为 t 时刻的高斯分布的均值和标准差，其中均值是由网络输出 $\mathbf{h}_t$ 经过一个线性变换得到，标准差先经过线性变换之后进行非线性变换得到，确保 $\sigma>0$，由公式（5）和（6）得到：

$$\begin{aligned} &l_G(Y\,|\,\theta(\mathbf{h},\Phi)) = l_G(Y\,|\,\mu,\sigma) = \\ &(2\pi\sigma^2)^{-\frac{1}{2}}\exp(-(Y-\mu)^2/(2\sigma^2)) \end{aligned} \tag{4}$$

均值的计算：

$$\mu(\mathbf{h}_t) = \mathbf{w}_\mu^T\mathbf{h}_t + b_\mu \tag{5}$$

标准差的计算：

$$\sigma(\mathbf{h}_t) = \log(1+\exp(\mathbf{w}_\sigma^T\mathbf{h}_t + b_\sigma)) \tag{6}$$

式（4）、（5）、（6）中，μ 和 σ 分别为似然函数的均值和标准差，w 和 b 分别是线性变换的权

重和偏置。

对于训练和预测过程，它们的网络结构是相同的。对于训练过程，Y 的值是已知的，但在预测过程中 Y 是未知的。为了继续预测，需要从最后一个时间步长的分布中得到一个采样值，作为下一步预测的输入数据。关于训练和预测的详细内容将在 2.2 节和 2.3 节中分别进行描述和讨论。

2.2　训练

在训练自回归 LSTM 网络时，输入变量为 X_t 和 Y_{t-1}。所有训练数据都在条件区间 $[1:t_0]$ 内。自回归 LSTM 网络依据时间展开，进行连续的训练过程。在每一个时间步骤 t，它们的输入是 (Y_{t-1}, X_t) 和上一时刻的网络输出 $\mathbf{h}_{t-1}$，并且 $t \in [1:t_0]$。网络输出 $\mathbf{h}_t = h(\mathbf{h}_{t-1}, Y_{t-1}, X_t, \Phi)$ 被用来计算 t 时刻的似然函数的参数 $\theta_t = \theta(\mathbf{h}_t, \Theta)$。最后，使用公式（7）优化模型参数。

$$L = \sum_{t=t_0+1}^{t_0+\tau} \log l(Y_t | \theta(\mathbf{h}_t)) \tag{7}$$

式（7）式中，$\mathbf{h}_t$ 是网络的输出；Y_t 是预测变量的真实值。

最终，通过最大化对数似然函数 L 作为损失函数来优化学习网络的参数 $h(\cdot)$ 和高斯分布的参数 $\theta(\cdot)$，使用随机梯度下降(SGD)来对模型进行优化，从而得到整个预测模型的权重参数 Θ。

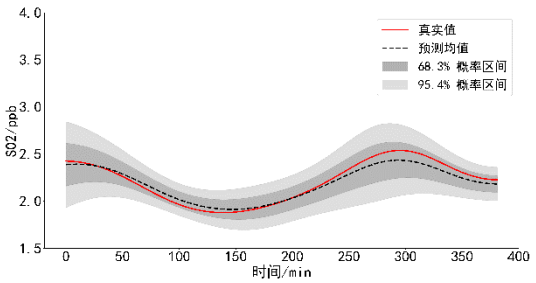
2.3　预测

预测过程的网络结构与参数和训练过程相同。但是，预测网络的输入与训练网络的输入不同，预测变量的实际值在区间 $[t_0+1:t_0+\tau]$ 内是未知的。因此，通过从预测分布中抽样获得抽样 $\hat{Y}_{t_0+1:t_0+\tau}$：$P_\Phi(Y_{t_0+1:t_0+\tau} | Y_{1:t_0}, X_{1:t_0+\tau})$，并作为下一时间步骤的输入变量。

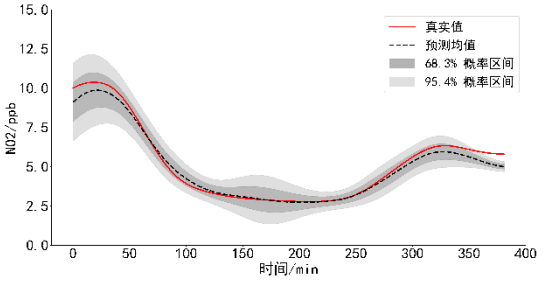
通过滚动窗口预测，可以给出 $[t_0+1:t_0+\tau]$ 范围内所有预测时刻的概率密度函数。整个的预测步骤如下：首先，$\mathbf{h}_{t_0}$ 在训练过程结束时获得；然后利用公式(3)计算 $\mathbf{h}_{t_0+1}$。在得到网络输出 $\mathbf{h}_{t_0+1}$ 后，建立高斯似然函数 $l(Y_{t_0+1} | \theta_{t_0+1})$。最后，抽样获得 $\hat{Y}_{t_0+1}$：$l(Y_{t_0+1} | \theta_{t_0+1})$，并作为下一时刻的输入数据。重复此预测过程，直到 $[t_0+1:t_0+\tau]$ 区间中的点全部预测完毕。

3　预测结果分析

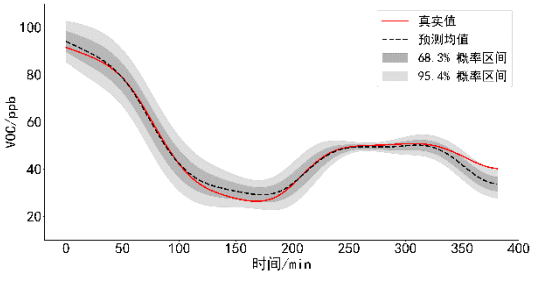
为了验证本文预测方法的准确性，实验使用地铁车站采集的 1260 组数据进行预测。训练和测试数据按 7:3 的比例进行划分。时间步长设置为 120s，训练迭代次数设置为 1000。实验环境为 python3.7（处理器：Intel(R) Core(TM) i5-8400 CPU @2.8GHz；内存：8.00 GB），模型的评价是基于正态分布的 3σ 准则。3σ 准则指出，对于许多合理对称的单峰分布，几乎所有的数据都分布在在平均值附近的三个标准差内[23]。对于标准正态分布，68.3%的观测值在范围 $[\mu-\sigma, \mu+\sigma]$ 内，95.4%在范围 $[\mu-2\sigma, \mu+2\sigma]$ 内，99.7%在范围 $[\mu-3\sigma, \mu+3\sigma]$ 内。本文实验在 3σ 准则的基础上，删除第三个区间，并根据预测得到的高斯分布的均值和方差，定义了 2 个区间，分别为 $[\mu-\sigma, \mu+\sigma]$ 和 $[\mu-2\sigma, \mu+2\sigma]$，预测结果如图 3 所示。表 3 表明预测值的分布在不同区间的比例。



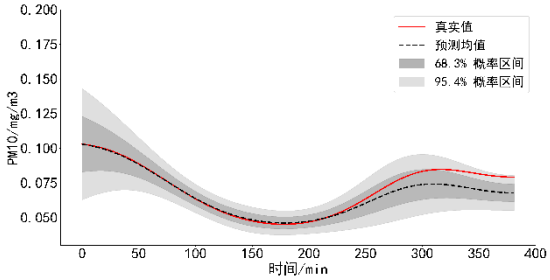
（a）SO_2



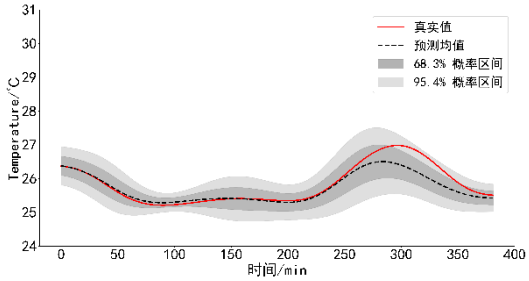
（b）NO_2



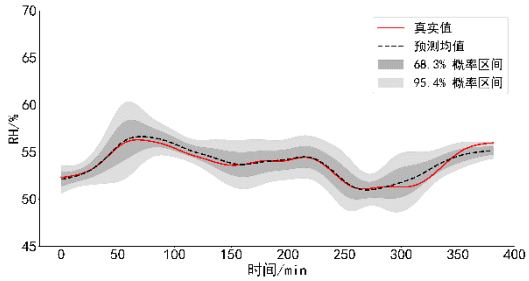
（c）VOC



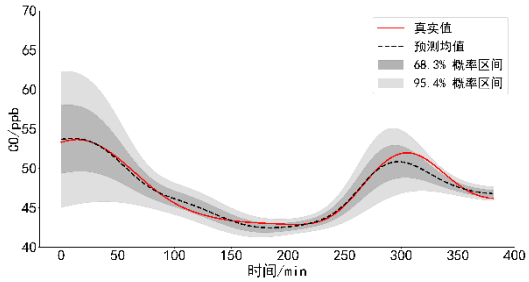
（d）PM_{10}



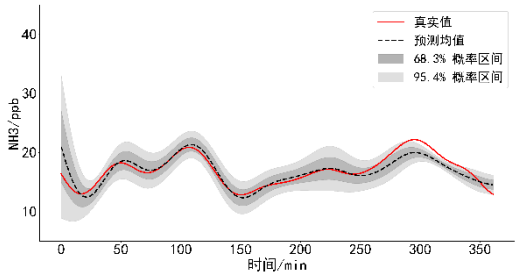
（e）Temperature



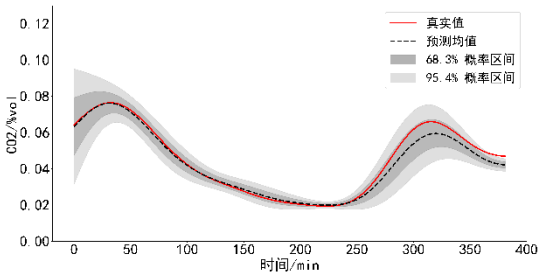
（f）RH



（g）CO



（h）NH_3



（i）CO_2

图 3　环境参数预测曲线

Fig.3　Prediction curves of environmental parameters

Station1 站内环境参数的预测结果如图 3 所示，图中红色实线代表环境参数真实值，灰色虚线条代表概率预测的均值；阴影部分代表预测不同的概率区间。

表 3　真实值在不同预测区间的分布

Table 3　Ratio of actual values falling in two ranges

环境参数	$\mu\pm\sigma$	$\mu\pm2\sigma$
SO_2/%	72.25	93.46
NO_2/%	85.30	100.00
VOC/%	88.74	96.60
PM_{10}/%	100.00	100.00
Temperature/%	74.86	87.85
RH/%	89.01	93.72
CO/%	81.41	100.00
NH_3/%	84.03	100.00
CO_2/%	93.98	100.00
Mean/%	85.51	96.85

如表 3 所示，我们计算了 Station1 车站 9 个环境参数的真实值在 2 个区间内的分布比例，并计算出所有参数分布在不同区间的平均值，在最后一行列出。结果表明，2 个区间的真实值分布比例平均为 85.51%、96.85%，同时，如图 3 所示，环境参数的真实值大部分在概率预测模型的概率区间内。预测结果都包含在概率预测模型预测的正态分布范围内，这意味着我们的模型所预测的正态分布能够有效地覆盖预测变量的变化范围，并给出不同的概率区间。

4　结论

基于自回归 LSTM 网络，本文提出了一种地铁车站环境参数概率预测方法，利用外部变量和地铁站中预测变量的历史数据来预测站内环境参数

未来时刻的概率分布，并使用地铁车站现场采集的观测数据对模型进行了验证。最终，得到以下结论：

（1）本文提出的模型可以在过去数据和未来数据之间建立条件分布，其预测结果是一系列包含均值和标准差的高斯分布。与传统的点预测方法相比，概率预测方法还可以提供其他信息，例如，预测变量数值分布得上边界和下边界以及对应的概率。

（2）本文实验通过计算落在 2 个概率区间内的实际值的频率以证明模型的可靠性，结果表明，在 2 个概率区间内，平均分布有 85.51%，96.85% 的实际值，预测的高斯分布能够表示未来时刻的环境参数值的变化范围。

参考文献：

[1] Martins V, Moreno T, Minguillón, María Cruz, et al. Exposure to airborne particulate matter in the subway system[J]. Science of The Total Environment, 2015,511: 711-722.

[2] Kim M J, Sankararao B, Kang O Y, et al. Monitoring and prediction of indoor air quality (IAQ) in subway or metro systems using season dependent models[J]. Energy and Buildings, 2012,46(none):48-55.

[3] Liu H, Lee S C, Kim M J, et al. Multi-objective optimization of indoor air quality control and energy consumption minimization in a subway ventilation system[J]. Energy and Buildings, 2013,66:553-561.

[4] Kim M J, Braatz R D, Kim J T, et al. Indoor air quality control for improving passenger health in subway platforms using an outdoor air quality dependent ventilation system[J]. Building and Environment, 2015,92(oct.):407-417.

[5] 苏华,徐来福,田胜元.用神经网络拟合逐时气象参数[J]. 制冷与空调,2005(z1):37-40.

[6] 段淇倡,刘顺波,周光伟.补偿模糊神经网络的改进及其在通风空调故障诊断中的应用[J].制冷与空调,2013(2):121-125.

[7] 王石,易佳婷.人工神经网络-研究制冷系统的新方法[J]. 制冷与空调,2005,5(2):47-51.

[8] Bo Y, Yao C, Lin Z, et al. Beam Structure Damage Identification Based on BP Neural Network and Support Vector Machine[J]. Mathematical Problems in Engineering, 2014:1-8.

[9] Xiao-Ping B, Hong L I, Qi-Ming Z, et al. Progress of Research on Artificial Neural Network in Air Pollution Prediction[J]. Science & Technology Review, 2006, 24(12):77-81.

[10] Chen Q C Q, Shao Y S Y. The Application of Improved BP Neural Network Algorithm in Urban Air Quality Prediction: Evidence from China[C]. Workshop on Computational Intelligence & Industrial Application, IEEE Computer Society, 2008.

[11] Wang F, Cheng S Y, Li M J, et al. Optimizing BP networks by means of genetic algorithms in air pollution prediction[J]. Beijing Gongye Daxue Xuebao / Journal of Beijing University of Technology, 2009,35(9):1230-1234.

[12] Lu T, Viljanen M. Prediction of indoor temperature and relative humidity using neural network models: model comparison[J]. Neural Computing & Applications, 2009,18(4):345-357.

[13] Kamal M M, Jailani R, Shauri R L A. Prediction of Ambient Air Quality Based on Neural Network Technique[C]. Research and Development, 2006. SCOReD 2006. 4th Student Conference on. IEEE, 2006.

[14] Bodri L, V. Čermák. Prediction of Surface Air Temperatures by Neural Network, Example Based on Three-Year Temperature Monitoring at Spořilov Station[J]. Studia Geophysica et Geodaetica, 2003,47(1): 173-184.

[15] Kim M H, Kim Y S, Lim J J, et al. Data-driven prediction model of indoor air quality in an underground space[J]. Korean Journal of Chemical Engineering, 2010,27(6): 1675-1680.

[16] Lim J J, Kim Y S, Oh T S, et al. Analysis and prediction of indoor air pollutants in a subway station using a new key variable selection method[J]. Korean Journal of Chemical Engineering, 2012,29(8):994-1003.

[17] Hongquan Q, Shuo F, Liping P, et al. Rapid Temperature Prediction Method for Electronic Equipment Cabin[J]. Applied Thermal Engineering, 2018,138.

[18] Valentin Flunkert, David Salinas, Jan Gasthaus. Deepar: Probabilistic forecasting with autoregressive recurrent networks[J]. arXiv preprint arXiv:1704.04110, 2017.

[19] Buizza R. The value of probabilistic prediction[J]. Atmospheric Science Letters, 2008,9.

[20] Gneiting, Tilmann, Katzfuss, Matthias. Probabilistic Forecasting[J]. Social Science Electronic Publishing, 2014,(1):125-151.

[21] Aznarte, José L. Probabilistic forecasting for extreme NO_2, pollution episodes[J]. Environmental Pollution, 2017, 229:321-328.

[22] Pukelsheim Friedrich. The Three Sigma Rule[J]. American Statistician, 1994,48(2):88-91.