网络出版时间：2018-11-15 17:58:08
网络出版地址：http://kns.cnki.net/kcms/detail/11.1895.X.20181115.1757.019.html

支持向量机回归在臭氧预报中的应用

苏筱倩[1]，安俊琳[1]*，张玉欣[2]，梁静舒[3]，刘静达[1]，王鑫[1]

（1. 南京信息工程大学，气象灾害教育部重点实验室，气候与环境变化国际合作联合实验室，气象灾害预报预警与评估协同创新中心，南京 210044；2. 青海省人工影响天气办公室，西宁 810001；3. 中国气象局气象探测中心，北京 100081）

摘要：采用南京工业区 2016 年 5 月 20 日~8 月 15 日这一高臭氧（O_3）期的 O_3、O_3 前体物和常规气象资料数据，利用支持向量机回归（SVMr）方法分别预报 O_3 的小时值、日最大值和最大 8h 滑动平均值。结果表明，O_3 小时值预报的相关系数（R^2）为 0.84，平均绝对误差（MAE）和平均绝对百分误差（MAPE）分别为 3.44×10^{-9} 和 24.48，O_3 前期浓度、紫外 B 波段辐射（UVB）和 NO_2 浓度是关键因子。O_3 日最大值预报的主要因子是 NO_x 在 07:00 的浓度和 UVB。预报 $O_3$8h 时 UVB 和气温起重要作用。加入前体物项能够使 O_3 的预报精度提升 10%~28%。与多元线性回归方法相比，SVMr 对 O_3 浓度的预报有明显优势。

关键词：支持向量机回归；臭氧预报；臭氧小时值；臭氧日最大值；臭氧日最大 8h 滑动平均

中图分类号：X515 文献标识码：A 文章编号：0250-3301（2019）

DOI:10.13227/j.hjkx.201809134

Application of Support Vector Machine Regression in Ozone Forecast

SU Xiao-qian[1], AN Jun-lin[1]*, ZHANG Yu-xin[2], LIANG Jing-shu[3], LIU Jing-da[1], WANG Xin[1]

(1. Key Laboratory of Meteorological Disaster, Ministry of Education, Joint International Research Laboratory of Climate and Environment Change, Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Weather Modification Office of Qinghai Province, Xining 810001, China; 3. Meteorological Observation Centre of China Meteorological Administration, Beijing 100081, China)

Abstract: The support vector machine regression (SVMr) was proposed to forecast hourly ozone (O_3) concentrations, daily maximum O_3 concentrations and maximum 8 h moving average O_3 concentrations (O_3 8 h) by employing the observations of meteorological variables, O_3 and its precursors during the high O_3 periods from May 20 to August 15, 2016 at an industrial area of Nanjing. The squared correlation coefficient (R^2) of the hourly O_3 concentrations forecast was 0.84. The mean absolute error (MAE) and mean absolute percentage error (MAPE) were 3.44×10^{-9} and 24.48, respectively. The key factors of hourly O_3 forecast were O_3 pre-concentrations, ultraviolet radiation B (UVB) and NO_2 concentrations. The main factors for O_3 daily maximum forecast were NO_x concentrations at 07:00 and UVB. Temperature and UVB played an important role in predicting O_3 8 h. In general, precursors could increase the accuracy of O_3 prediction by 10%~28%. Concerning the O_3 concentrations forecast, SVMr gave significantly better predictions than multiple linear regression methods.

Key words: support vector machine regression; O_3 prediction; hourly O_3 concentrations; daily maximum O_3 concentrations; maximum 8 h moving average O_3 concentrations

近年来，随着中国工业化、城镇化进程的加快和汽车保有量的增加，光化学烟雾、雾-霾等复合型大气污染问题正严重影响着生态环境和公共健康[1~3]。研究大气污染物的预报方法，建立有效的大气污染物预警机制，对改善城市空气质量，政府制定控制策略有重大的应

收稿日期：2018-09-17 ；**修订日期**：2018-11-12

基金项目：国家自然科学基金项目（91544229）；国家重点研发计划专项（2016YFC0202400）；江苏省高校“青蓝工程”项目

作者简介：苏筱倩（1994~），女，硕士研究生，主要研究方向为大气环境，E-mail：xiaoqiansue@163.com

* 通信作者，E-mail：junlinan@nuist.edu.cn

用价值。空气污染预报方法主要分为数值预报和统计预报。主流的数值预报模型有三维欧拉型模式（comprehensive air quality model with extensions，CAMx）、社区多尺度空气质量模式（community multiscale air quality，CMAQ）、化学天气数值模式（weather research and forecasting model coupled to chemistry，WRF-chem）、MM5-chem 大气化学模式（fifth-generation Penn state/NCAR mesoscale model coupled to chemistry，MM5-chem）和嵌套网格空气质量预报模式（nested air quality prediction modeling system，NAQPMS）等，数值预报方法能够模拟污染物的转化、迁移和扩散，反映污染物的变化规律，但是其建立在获取大量的气象数据、污染物排放源数据和空气监测数据的基础上，需要掌握污染变化的机制，计算耗时长。统计预报方法如回归模型[4~6]，有计算简单，资料要求较低和准确度高的优势，在业务预报中应用广泛，但其大多以线性回归理论为基础，难以应用到非线性系统。近年来，随着计算机技术的发展，神经网络[7~9]、决策树[10,11]和支持向量机（support vector machine，SVM）等基于统计理论的机器学习方法，在解决非线性问题时表现出优异的性能。

SVM 遵循结构风险最小化原则，善于解决小样本、非线性和高维模式识别问题[12]。与遵循经验风险最小化原则的人工神经网络等传统机器学习方法不同，SVM 避免了过拟合、局部最优或局部优化能力差、调参困难和收敛慢等问题[13,14]。近年来，支持向量机回归（support vector machine regression，SVMr）不仅用于预报太阳辐射[15]、云量[16,17]和能见度[18,19]，还广泛应用于预报 O_3 等大气污染物浓度。有研究者将 SVMr 对污染物的预报结果与线性回归模型[20]、多层感知机（MLP）[21]、向量自回归模型（VARMA）和自回归积分滑动平均模型（ARIMA）[22]进行比较，发现 SVMr 预报效果更优。Yeganeh[23]考虑了 SVMr 中 4 种核函数的差异，发现径向基核函数（RBF）最适合。有研究者将 SVM 与小波分解[24]、相空间重构理论[25]、遗传算法优化的 BP 神经网络（GA-BPNN）[26]等相结合来预报污染物，发现结合预报的精度大于仅使用 SVM 或人工神经网络。Xu 等[27]基于 SVM 开发了太原、哈尔滨和重庆这 3 个城市的空气质量预警系统，经实验比较，此系统的准确性和有效性均高于其现有的空气质量预警结果，具有应用价值。

之前的学者多致力于 SVMr 与其他预报方法的比较，对 O_3 的预报较少，预报 O_3 的时次单一且多考虑气象因素。本研究在考虑气象因素的基础上，还加入了 NO、NO_2、氮氧化物（nitrogen oxides，NO_x）、挥发性有机物（volatile organic compounds，VOCs）和一氧化碳（carbon monoxide，CO）这 3 种前体物，同时预报了 O_3 的小时平均值、日最大值和最大 8h 滑动平均这 3 种 O_3 监测、预报常用的国家空气质量标准中指标，以期为 O_3 公众预警预报业务提供一种新的思路和方法。

1 材料与方法

1.1 观测站点

本研究观测站点位于江苏省南京市浦口区南京信息工程大学气象楼楼顶(32°12'N，118°42'E，海拔高度 62 m)。站点东边 500m 处为主干道宁六路、高架快速路和地铁 S8 号线；站点东北 5km 处为包括石油化工、钢铁厂、化工厂和热电厂等在内的工业区；其西南 900m 处为南京龙王山风景区。常规气象资料数据来源于距观测站点约 1.5km 的中国气象局综合观测实习基地。站点具体位置见图 1。

图 1 观测点的位置和附近环境

Fig. 1 Observation location of the site and its surroundings

1.2 仪器及监测方法

O_3、NO、NO_2、NO_x 和 CO 的观测均采用美国赛默飞世尔科技公司生产的大气污染环境监测分析仪，包括 49i 紫外发光 O_3 分析仪，42i 化学发光 NO-NO_2-NO_x 分析仪和 48i 红外吸收 CO 分析仪。详细仪器参数及校准方法可参见文献[28]。

大气中 VOCs 观测采用由德国 AMA 公司生产的 GC5000 自动在线气相色谱氢火焰离子监测系统（gas chromatography-flame ionization detector，GC-FID）进行连续监测。详细仪器参数及校准方法可参见文献[29]。

1.3 支持向量机回归（SVMr）模型

SVM 是 Vapnik 于 1995 年首先提出的机器学习方法，此方法建立在统计学习中 VC 维理论和结构风险最小原理的基础上，其主要思想是将低维空间中的 x 用非线性函数φ映射到一个高维特征空间$\varphi(x)$，在高维空间中寻求线性回归超平面从而解决低维空间中的非线性问题。高维特征空间中的线性函数可以构造为：

$$y = \langle w\varphi(x)\rangle + b \tag{1}$$

式中，y 为输出，$\langle w\varphi(x)\rangle$表示特征空间的内积，权重向量 $\boldsymbol{w}$ 和偏置常数 b 可以通过最小化风险函数[式（2）]得到。

$$Q = \frac{\|w\|^2}{2} + C\sum_{i=1}^{N} L_\varepsilon(x_i,\ y_i,\ f) \tag{2}$$

$$\text{式中，}\ L_\varepsilon(x_i,\ y_i,\ f) = \begin{cases} |y_i - f(x_i)| - \varepsilon,\ |y_i - f(x_i)| \geq \varepsilon \\ 0,\ \text{otherwise} \end{cases} \tag{3}$$

式中，C 是预先设定的惩罚系数，惩罚大于 ε 的误差。ε 是训练集和实际观测值之间的偏差。引入松弛变量ξ_i和ξ_i^*，则式（2）的求解可以转变为式（4）：

$$\min\ \left(\frac{\|w\|^2}{2} + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)\right) \tag{4}$$

约束条件：

$$\begin{gathered} y_i - \big((w \times x_i) + b\big) \leq \varepsilon + \xi_i \\ \big((w \times x_i) + b\big) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i^*,\ \xi_i \geq 0 \end{gathered} \tag{5}$$

引入拉格朗日乘子α_i和α_i^*，建立拉格朗日函数进而求解原问题的对偶问题。最终得到最优超平面的回归函数，如式（6）：

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_i,\ x) + b \tag{6}$$

式中，$K(x_i,\ x)$成为核函数，其表达式如下：

$$K(x_i,\ x) = \varphi\langle x_i,\ x\rangle = \varphi(x_i)^T\varphi(x) \tag{7}$$

本研究使用 LIBSVM3.22 软件包（http://www.csie.ntu.edu.tw/~cjlin/libsvm/），选择的是对非线性系统拟合好且在大气污染物浓度预报中应用最多的 RBF 核函数，其表达式为 $K(x_i,\ x) = \exp\left[-\|x_i - x\|^2/(2\sigma^2)\right]$，其中$\sigma$为可调参数。建模时将数据分为训练集和预报集，采用 k 折交叉验证（k=3）和 MATLAB 编程自动寻找使均方误差（mean square error，MSE）最小的参数 c 和 g，再将其应用到预报集得出结果。

1.4 实验数据

本研究选取 2016 年 5 月 20 日~8 月 15 日这一 O_3 高值时段进行预报，缺失的数据都被剔除，期间有效的小时平均数据为 1436 组，日平均数据为 71 组，样本数达到了 SVMr 建模所需的特征量[22,26,30]。为提高模型的泛化能力，选取各月约 70%的数据作为训练集，剩余时

段的数据作为预报集，所用的数据通过 MATLAB 中的 mapminmax 函数归一化到（0，1）以消除不同数据间量纲的影响。O_3 及其前体物的单位均为体积分数（$\times10^{-9}$，浓度），观测的 56 种 VOCs 分为烷烃、芳香烃、烯烃和炔烃这 4 类进行讨论，具体划分可参见文献[31]。为方便表达，文中模式输入方案的变量名称均用缩写表示（见表 1）。

表 1 变量缩写及其含义

Table 1 Abbreviation of variables and their descriptions

模式中缩写名称	表示意义	单位
[O_3_1]	预报时刻前 1h 的 O_3 体积分数	$\times10^{-9}$
[O_3_2]	预报时刻前 2h 的 O_3 体积分数	$\times10^{-9}$
[O_3_24]	预报时刻前 24h 的 O_3 体积分数	$\times10^{-9}$
T	预报时刻的气温	℃
T_{max}	预报日的最高气温	℃
T_a	预报日的平均气温	℃
UVB_h	预报时刻的紫外 B 波段辐射	W ㎡$^{-2}$
UVB_t	预报日的紫外 B 波段辐射累积值	W ㎡$^{-2}$
SH	预报时刻的日照时数	h
RH	预报时刻的相对湿度	%
RH_a	预报日的相对湿度日平均值	%
R	预报日的累积降水量	mm
[NO]、[NO_2]、[NO_x]	预报时刻的 NO、NO_2、NO_x 体积分数	$\times10^{-9}$
[NO7]、[$NO_2$7]、[NO_x7]	预报日 07:00 的 NO、NO_2、NO_x 体积分数	$\times10^{-9}$
[烷]、[烯]、[芳]、[炔]	预报时刻的烷烃、烯烃、芳香烃、炔烃体积分数	$\times10^{-9}$
[烷 $_a$]、[烯 $_a$]、[芳 $_a$]、[炔 $_a$]	预报日的烷烃、烯烃、芳香烃、炔烃体积分数日平均值	$\times10^{-9}$
[CO]	预报时刻的 CO 体积分数	$\times10^{-9}$
[CO_a]	预报日的 CO 体积分数日平均	$\times10^{-9}$

1.5 评价标准

为比较 SVMr 预报结果与观测值之间的差别，采用以下统计参量对模型进行评价和选择。

均方误差（mean square error，MSE）：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(O_i-P_i)^2 \tag{8}$$

决定系数（squared correlation coefficient，R^2）：

$$R^2=1-\frac{\mathrm{SS_{res}}}{\mathrm{SS_{tot}}}=1-\frac{\sum（O_i-P_i）^2}{\sum（P_i-O_\mathrm{m}）^2} \tag{9}$$

平均绝对差值（mean absolute error，MAE）：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|O_i-P_i| \tag{10}$$

平均绝对百分误差（mean absolute percentage error，MAPE）：

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\frac{|P_i-O_i|\cdot100}{O_i} \tag{11}$$

式中，O_i 表示在 i 时的观测值，P_i 表示在 i 时的预报值，n 表示样本总数目，$\mathrm{SS_{res}}$ 表示残差平方和，$\mathrm{SS_{tot}}$ 表示总离差平方和，O_m 表示观测值的平均数。

2 结果与讨论

2.1 O_3 小时值预报

首先，对气温、风速、风向、相对湿度和太阳辐射等常规气象变量的小时平均值与 O_3 小时平均值做相关性分析，将相关系数绝对值大于 0.40（均通过显著性水平 α=0.05 的显著性检验）的气象变量作为模型的输入。由此得到的气象输入变量为气温、相对湿度、紫外 B 波段辐射（ultraviolet radiation B，UVB）和日照时数，相关系数分别为 0.47、−0.74、0.50 和 0.53。预报时刻前期的 O_3 浓度能够提高模型预报的准确性[32]，经模型测试，选取预报时刻前 1h 的 O_3 浓度作为输入变量。由于模型预报时间不超过 15s，将上述气象变量拟作为最优组合全部带入模型，每次只去除一项输入因子，以检验各变量在模型中的表现。结果显示（表 2），去除 UVB 项后，模型的 MAE 和 MAPE 均增加 8%，去除日照时数项，MAE 和 MAPE 分别增加了 2%和 12%，其它因素的变化相对不明显。这说明在气象因素中，太阳辐射作为大气中光化学反应的主要能源，对 O_3 小时值的预报尤为重要。

表 2 气象变量对 SVMr 预报精度的影响

Table 2 Effect of meteorological variables on prediction accuracy of SVMr

输入方案	R^2	MAE×10^{-9}	MAPE
[O_3_1]、UVB_h、SH、*T*、RH	0.81	3.83	27.92
去 *T*	0.81	3.82	28.61
去 RH	0.81	3.69	28.21
去 UVB_h	0.79	4.14	30.15
去 SH	0.80	3.91	31.32

由于 O_3 与其前体物有着非线性，强耦合的关系，这里在气象变量确定的基础上，直接将各前体物分别带入模型进行筛选，确定拟最优方案。结果如表 3 所示。可以发现前体物中，加入 NO_2 后模型的 MAE 和 MAPE 减小最多，均达到 10%，其次为 CO，其 MAE 和 MAPE 分别减少 9%和 6%。宁六路等交通干道和附近的综合工业区是 NO_2 和 CO 重要的人为源地，日出后，NO_2 等前体物经过复杂的光化学反应生成 O_3。另外，4 种 VOCs 中，芳香烃的表现最优，其次为烯烃。张玉欣等[33]利用箱模式计算了南京工业区夏季 VOCs 的相对增量反应性，发现烯烃和芳香烃是控制 O_3 浓度最有效的两类物种，这与模型的预报结果相吻合。根据模型结果，将 NO_2、CO 和芳香烃的小时平均值加入拟最优方案。

表 3 前体物对 SVMr 预报精度的影响

Table 3 Effect of precursors on prediction accuracy of SVMr

输入方案	R^2	MAE×10^{-9}	MAPE
[O_3_1]、UVB_h、SH、*T*、RH	0.81	3.83	27.92
加[NO]	0.81	3.76	32.78
加[NO_2]	0.84	3.43	25.08
加[NO_x]	0.82	3.66	32.55
加[CO]	0.82	3.79	26.29
加[芳]	0.81	3.77	27.87
加[烷]	0.80	3.89	29.59
加[烯]	0.80	3.86	28.34
加[炔]	0.79	4.20	33.31

O_3 小时值的预报结果如图 2 所示，小时预报值与观测值的趋势吻合得很好，两者相关系数达到 0.84，模型对峰值、谷值都有较为准确的捕捉，MAE、MAPE 分别为 3.44×10^{-9} 和 24.48。表 4 展示了对拟最优方案进行敏感性分析后的预报结果。可以看出，无论去除哪一项变量，预报的准确度都有不同程度地降低，这说明选取的变量对模型预报都有优化作用。综合来看，在气象因子中，UVB 和气温的作用最明显，从模型中去除它们后，MAE 分别增

加了 12%和 3%，MAPE 分别增加 11%和 16%。在前体物中，NO_2 的作用最显著。对比表 2 与表 4 可以发现，在气象变量的基础上加入前体物后，模型的 MAE 和 MAPE 分别降低了 10%和 12%，R^2 由 0.81 提高到 0.84，这说明考虑前体物为预报因子能够有效提高 SVMr 预报 O_3 小时浓度的精度。

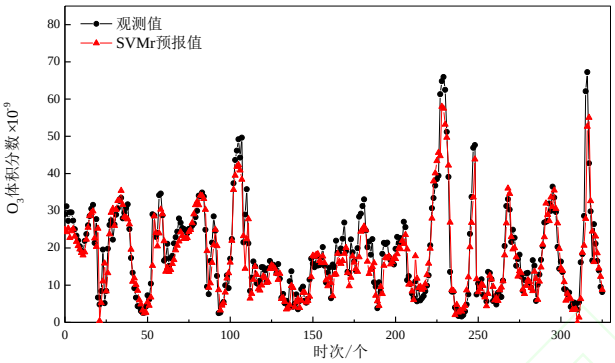


图 2 O_3 小时浓度的观测值和预报值对比

Fig. 2 Observed Vs Predicted hourly O_3 concentration

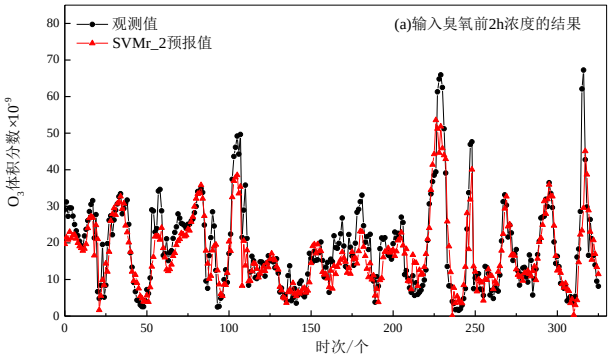
多元线性回归方法作为经典统计预报方法之一，在当今业务预报中仍常使用，本研究利用 SPSS 软件，输入与 SVMr 最优方案相同的预报因子，构建多元线性回归方程，结果如表 4 所示，SVMr 各项统计量均优于多元线性回归方法，在南京工业区夏季 O_3 小时浓度的预报中显示出优势。

表 4 O_3 小时浓度的预报结果

Table 4 Forecast results of hourly O_3 concentration

输入方案	R^2	MAE×10^{-9}	MAPE
[O_3_1]、UVB_h、SH、T、RH、[NO_2]、[CO]、[芳]	0.84	3.44	24.48
去 T	0.84	3.54	28.33
去 UVB_h	0.83	3.84	27.11
去 RH	0.83	3.53	27.08
去[NO_2]	0.82	3.79	26.36
去[CO]	0.84	3.51	26.10
去 SH	0.84	3.53	26.07
去[芳]	0.84	3.45	24.51
多元线性回归	0.81	3.90	26.58

用 O_3 前 2h 浓度替代前 1h 浓度带入模型，其 R^2 也能达到 0.75，如图 3（a）预报值和观测值的趋势基本一致，但峰值有较大差异，MAE 和 MAPE 分别增加到 4.43×10^{-9} 和 31.21。若将 O_3 前 24h 浓度带入模型，R^2 降低到 0.45。这说明在 O_3 小时值预报里，O_3 的前期浓度是影响最大的因素，前体物各变量与其相比作用不明显。



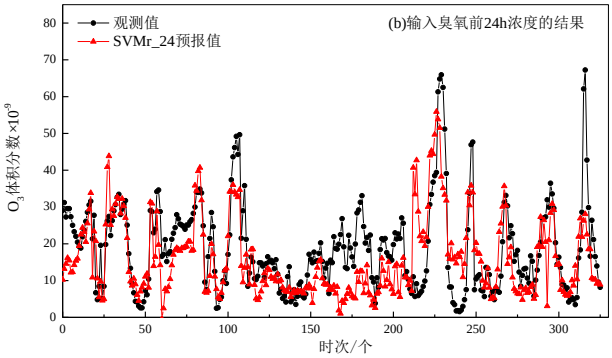


图 3 O_3 不同前期浓度对 SVMr 预报的影响

Fig. 3 Effect of different pre-concentrations of O_3 on prediction using SVMr

2.2 O_3 日最大值预报

按照前文所述，通过统计相关性分析，选取日最高气温、日平均相对湿度、日累积降水量和日累积 UVB 作为输入的气象变量，其相关系数分别为 0.50、−0.61、−0.46 和 0.60。在模型测试中，相对湿度的优化作用不明显，为避免信息冗余并保证模型的可解释性，将其去除。

图 4 描述了各前体物的日变化。由于夜间的积累及早高峰车辆的排放，各前体物在早上呈现上升的趋势，随着日出后边界层高度抬升，太阳辐射加强近地面的湍流混合作用，同时光化学反应也开始进行，各前体物浓度在早上 07:00~08:00 达到最大值后逐渐下降。由此，选取 NO、NO_2 和 NO_x 在 07:00 的浓度、VOCs 和 CO 在 08:00 的浓度作为输入变量，它们分别代表了各前体物在光化学反应前的初始浓度。同时，选取各前体物浓度的日平均值带入模型，它们在一定程度上代表了前体物在一天中的平均状况。

在气象变量的基础上分别输入各前体物项，经比较，最终选取 NO_x 在 07:00 的浓度值、烯烃的日平均值作为输入变量，MAE 和 MAPE 分别降低了 28%和 23%，说明前体物项显著提升了 O_3 日最大值的预报精度。在测试时发现，虽然 NO、NO_2 和 NO_x 在 07:00 的浓度值都可以使模型得到优化，但若将三者同时放入模型，其预报效果反而会降低，这可能是由于 NO、NO_2、NO_x 之间存在相关关系，在没有 O_3 前期浓度参与预报时，这种相关关系产生的影响更加凸显。因此，在 NO、NO_2 和 NO_x 中，只挑选更主要的因子，即 NO_x 在 07:00 的浓度值。另外，从加入 VOCs 的预报效果来看，其日平均值要略优于 08:00 的值。这可能是因为 O_3 日最大值一般出现于下午 14:00~16:00，而 VOCs 的化学反应速率很快，其早上 08:00 的高浓度不足以影响到 O_3 下午的最大值。与 O_3 小时浓度预报相似，4 种 VOCs 中烯烃的表现最优，邵平等[34]采用丙烯等量体积分数比较夏季南京北郊 VOCs 的反应活性，也发现烯烃所占的比例最高。

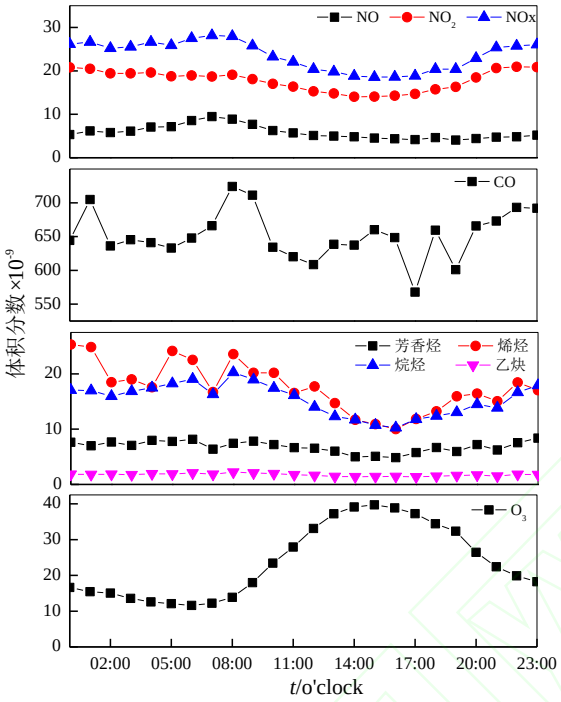


图 4 O_3 及前体物的日变化曲线

Fig. 4 Diurnal variations of O_3 and its precursors

将选取的上述变量带入模型，结果显示，O_3 日最大值的预报值和观测值间的 MAE 和 MAPE 分别为 5.48×10^{-9} 和 17.26，模型预报结果比较理想。表 5 给出了不同输入变量对模型预报效果的影响。可以看出各输入变量的影响差异不大，NO_x 在 07:00 的浓度最重要，去除它后，MAE 和 MAPE 分别提高了 28%和 20%，其次是 UVB，说明早高峰时期排放的 NO_x 通过光化学反应容易积累生成高浓度的 O_3。

表 5 O_3 日最大浓度的预报结果

Table 5 Forecast results of daily maximum O_3 concentration

输入方案	MAE（$\times10^{-9}$）	MAPE
T_{max}、R、UVB_t、[烯 $_a$]、[NO_x7]	5.48	17.26
去 R	5.85	18.32
去 T_{max}	6.35	19.32
去 UVB_t	6.64	19.72
去[NO_x7]	7.01	20.71
去[烯 $_a$]	5.89	19.14
多元线性回归	7.68	25.09

由图 5（a）可以直观地看到，SVMr 对 15 天模拟的差异不大，但高估了 7 月 6 日的 O_3 日最大浓度，低估了 5 月 25 日和 7 月 7 日的浓度，这几日在不同时段都有不同程度地降水，反映出模型对晴天或无降水日的模拟效果优于降水日，将天气分类研究可能会提高模型预报 O_3 浓度的日最大值的准确性。用相同因子建立多元线性回归方程，其 MAE 和 MAPE 较 SVMr 的结果分别提高了 40%和 45%，SVMr 表现出明显的优势。

2.3 O_3 最大 8h 滑动平均预报

经统计，本文研究时期中 O_3 最大 8h 滑动平均（O_3 8h）出现最多的时段是 12:00~19:00，占总天数的 30%，其次为 13:00~20:00 和 11:00~18:00，分别占总天数的 20%和 18%。这表明 O_3 8h 污染常出现在白天晴热高温、太阳直射紫外线较强的时段。

与前文的方法相同，选取出日平均相对湿度、日累积降水量、日累积 UVB、日平均气温和 CO 的日平均浓度作为输入变量。在气象因子的基础上加入前体物项后，模型的 MAE 和 MAPE 降低了 25%和 22%。预报结果如图 5（b）所示，SVMr 模拟效果很好，但依然存在降水日的预报偏差较大的现象。

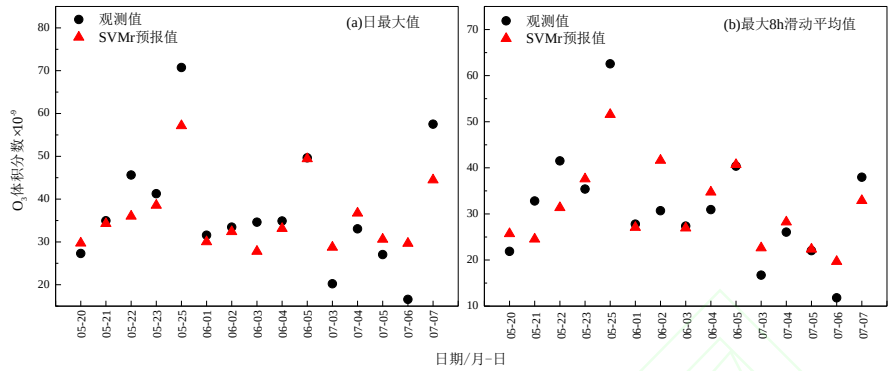


图 5 O_3 日最大浓度和最大 8h 滑动平均的观测值与预报值对比

Fig. 5 Observed Vs Predicted daily maximum O_3 concentration and maximum 8 h moving average O_3 concentration

从表 6 可以看到，去除日均气温后，模型的准确度最低，MAE 和 MAPE 分别增加了 47%和 63%，其次为 UVB 的日累积值和相对湿度。在预报 $O_3$8h 时，气象要素的重要性大于前体物，特别是气温、UVB 这类直接影响白天光化学反应的因子。相对湿度可以表征大气中的水汽含量，在近地面 O_3 的化学过程中，水汽含量的增加有利于反应 $O(^1D)+H_2O\rightarrow 2OH$ 的进行从而间接促进 O_3 的增加，另一方面，过高的水汽含量会直接对 O_3 产生湿清除。另外，水汽对 O_3 前体物之间光化学反应的非线性作用也会间接影响 O_3 的生成。在前体物中，VOCs、NO、NO_2 和 NO_x 相对于 CO 更加活泼，在预报 $O_3$8h 这种较长时间段的平均浓度时难有优势。

表 6 O_3 最大 8h 滑动平均的预报结果

Table 6 Forecast results of maximum 8 h moving average O_3 concentration

输入方案	MAE×10^{-9}	MAPE
R、RH_a、T_a、UVB_t、$[CO_a]$	4.87	17.94
去$[CO_a]$	6.51	23.12
去 R	6.48	25.14
去 RH_a	6.55	26.10
去 T_a	7.14	29.26
去 UVB_t	6.86	24.02
多元线性回归	8.58	31.98

3 结论

（1）SVMr 对南京工业区夏季的 O_3 浓度有准确预报。预报 O_3 小时值时的 R^2 达到 0.84，MAE 和 MAPE 分别为 3.44×10^{-9} 和 24.48。O_3 前期浓度是提高小时值预报精度的关键因子，其次为 UVB 和 NO_2。预报 O_3 日最大值的主要因素是 NO_x 在早高峰的浓度和 UVB，而日均气温和 UVB 等热力学因子在预报 $O_3$8h 时更加重要。O_3 8h 和日最大值的预报结果均表现出“晴天优于降水日”的特征，在研究时将天气分型可能降低预报误差。

（2）相较于仅考虑气象因素，方案中加入前体物项后 SVMr 预报的准确性有了 10%~28%的提升，前体物项有效地优化了模型。4 种 VOCs 中，烯烃和芳香烃等对南京工业区 O_3 生成贡献大的高活性物种在 O_3 浓度预报中值得关注。

（3）与多元线性回归方法相比，SVMr 的 MAE 和 MAPE 分别提高了 14%~78%和 9%~76%，SVMr 在预报 O_3 浓度时具有明显优势。

参考文献:

[1] Karlsson P E, Klingberg J, Engardt M, *et al*. Past, present and future concentrations of ground-level ozone and potential impacts on ecosystems and human health in northern Europe [J]. Science of the Total Environment, 2017, **576**: 22-35.

[2] Wang T, Xue L K, Brimblecombe P, *et al*. Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects [J]. Science of the Total Environment, 2016, **575**: 1582-1596.

[3] Yang C X, Yang H B, Guo S, *et al*. Alternative ozone metrics and daily mortality in Suzhou: The China air pollution and health effects study (CAPES) [J]. Science of the Total Environment, 2012, **426**(2): 83-89.

[4] Zhai L, Li S, Zou B, *et al*. An improved geographically weighted regression model for $PM_{2.5}$ concentration estimation in large areas [J]. Atmospheric Environment, 2018, **181**: 145-154.

[5] 安俊琳, 王跃思, 朱彬. 主成分和回归分析方法在大气臭氧预报的应用——以北京夏季为例[J]. 环境科学学报, 2010, **30**(6): 1286-1294.

An J L, Wang Y S, Zhu B. Principal component and multiple regression analysis predicting ozone concentrations: Case study in summer in Beijing [J]. Acta Scientiae Circumstantiae, 2010, **30**(6): 1286-1294.

[6] Barrero M A, Grimalt J O, Cantón L. Prediction of daily ozone concentration maxima in the urban atmosphere [J]. Chemometrics and Intelligent Laboratory Systems, 2006, **80**(1): 67-76.

[7] Gao M, Yin L T, Ning J C. Artificial neural network model for ozone concentration estimation and Monte Carlo analysis [J]. Atmospheric Environment, 2018, **184**: 129-139.

[8] Mao X, Shen T, Feng X. Prediction of hourly ground-level $PM_{2.5}$ concentrations 3 days in advance using neural networks with satellite data in eastern China [J]. Atmospheric Pollution Research, 2017, **8**(6): 1005-1015.

[9] Gómez-Sanchis J, Martín-Guerrero J D, Soria-Olivas E, *et al*. Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration [J]. Atmospheric Environment, 2006, **40**(32): 6173-6180.

[10] 丁愫, 陈报章, 王瑾, 等. 基于决策树的统计预报模型在臭氧浓度时空分布预测中的应用研究[J]. 环境科学学报, 2018, **38**(8): 3229-3242.

Ding S, Chen B Z, Wang J, *et al*. An applied research of decision-tree based statistical model in forecasting the spatial-temporal distribution of O_3 [J]. Acta Scientiae Circumstantiae, 2018, **38**(8): 3229-3242.

[11] 靳小兵, 徐会明, 卜俊伟, 等. 决策树方法在一次历史异常雷电活动中的预报能力检验[J]. 高原山地气象研究, 2013, **33**(4): 56-60.

Jin X B, Xu H M, Bu J W, *et al*. The forecast ability test of decision tree method in an abnormal lightning activity [J]. Plateau and Mountain Meteorology Research, 2013, **33**(4): 56-60.

[12] Vapnik V N. The Nature of Statistic Learning Theory [M]. New York: Springer, 2000. 17-34.

[13] García Nieto P J, Combarro E F, Montañés E, *et al*. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study [J]. Applied Mathematics and Computation, 2013, **219**(17): 8923-8937.

[14] Lu W Z, Wang W J. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends [J]. Chemosphere, 2005, **59**(5): 693-701.

[15] Quej V H, Almorox J, Arnaldo J A, *et al*. ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment [J]. Journal of Atmospheric and Solar-Terrestrial Physics, 2017, **155**: 62-70.

[16] 赵文婧, 赵中军, 汪结华, 等. 基于支持向量机的云量精细化预报研究[J]. 干旱气象, 2016, **34**(3): 568-574.

Zhao W J, Zhao Z J, Wang J H, *et al*. A study on refined forecast of cloud cover based on support vector machine [J]. Arid Meteorology, 2016, **34**(3): 568-574.

[17] 熊秋芬, 顾永刚, 王丽. 支持向量机分类方法在天空云量预报中的应用[J]. 气象, 2007, **33**(5): 20-26.

Xiong Q F, Gu Y G, Wang L. Application of SVM method to cloud amount forecast [J]. Meteorological Monthly, 2007, **33**(5): 20-26.

[18] 吴波, 胡邦辉, 王学忠, 等. 基于近似支持向量机的能见度释用预报研究[J]. 热带气象学报, 2017, **33**(1): 104-110.

Wu B, Hu B H, Wang X Z, *et al*. Visibility forecast based on proximal support vector machine [J]. Journal of Tropical

Meteorology, 2017, **33**(1): 104-110.

[19] 郑朝霞, 周梅, 季致建, 等. SVM 方法在霾识别和能见度预报中的应用[J]. 气象科技进展, 2016, **6**(6): 30-34.
Zheng Z X, Zhou M, Ji Z J, *et al*. Application of SVM method to identification of haze and prediction of visibility [J]. Advances in Meteorological Science and Technology, 2016, **6**(6): 30-34.

[20] Canu S, Rakotomamonjy A. Ozone peak and pollution forecasting using support vectors [R]. Yokohama: IFAC Workshop on Environmental Modeling, 2001.

[21] Garcá Nieto P J, Garcá-Gonzalo E, Sánchez A B, *et al*. Air Quality Modeling Using the PSO-SVM-Based Approach, MLP Neural Network, and M5 Model Tree in the Metropolitan Area of Oviedo (Northern Spain) [J]. Environmental Modeling & Assessment, 2017, **23**(3): 229-247.

[22] Garcá Nieto P J, Sánchez Lasheras F, Garcá-Gonzalo E, *et al*. PM_{10} concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study [J]. Science of the Total Environment, 2018, **621**: 753-761.

[23] Yeganeh B, Motlagh M S P, Rashidi Y, *et al*. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model [J]. Atmospheric Environment, 2012, **55**(3): 357-365.

[24] Zhu J, Wang Z H, Jin T L, *et al*. Combination of wavelet decomposition and least square support vector machine to forecast atmospheric ozone content time series [J]. Climatic & Environmental Research, 2010, **15**(3): 295-302.

[25] Niu M F, Gan K, Sun S L, *et al*. Application of decomposition-ensemble learning paradigm with phase space reconstruction for day-ahead $PM_{2.5}$ concentration forecasting [J]. Journal of Environmental Management, 2017, **196**: 110-118.

[26] Feng Y, Zhang W F, Sun D Z, *et al*. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification [J]. Atmospheric Environment, 2011, **45**(11): 1979-1985.

[27] Xu Y Z, Yang W D, Wang J Z. Air quality early-warning system for cities in China [J]. Atmospheric Environment, 2017, **148**: 239-257.

[28] 王俊秀, 安俊琳, 邵平, 等. 南京北郊大气臭氧周末效应特征分析[J]. 环境科学, 2017, **38**(6): 2256-2263.
Wang J X, An J L, Shao P, *et al*. Characteristic study on the "weekend effect" of atmospheric O_3 in northern Suburb of Nanjing [J]. Environmental Science, 2017, **38**(6): 2256-2263.

[29] An J L, Zhu B, Wang H L, *et al*. Characteristics and source apportionment of VOCs measured in an industrial area of Nanjing, Yangtze River Delta, China [J]. Atmospheric Environment, 2014, **97**: 206-214.

[30] Luna A S, Paredes M L L, Oliveira G C G D, *et al*. Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil [J]. Atmospheric Environment, 2014, **98**: 98-104.

[31] 安俊琳, 朱彬, 李用宇. 南京北郊大气 VOCs 体积分数变化特征[J]. 环境科学, 2013, **34**(12): 4504-4512.
An J L, Zhu B, Li Y Y. Variation characteristics of ambient volatile organic compounds (VOCs) in Nanjing northern suburb, China [J]. Environmental Science, 2013, **34**(12): 4504-4512.

[32] Chelani A B. Prediction of daily maximum ground ozone concentration using support vector machine [J]. Environmental Monitoring and Assessment, 2010, **162**(1-4): 169-176.

[33] 张玉欣, 安俊琳, 王俊秀, 等. 南京工业区挥发性有机物来源解析及其对臭氧贡献评估[J]. 环境科学, 2018, **39**(2): 502-510.
Zhang Y X, An J L, Wang J X, *et al*. Source analysis of volatile organic compounds in the Nanjing industrial area and evaluation of their contribution to ozone [J]. Environmental Science, 2018, **39**(2): 502-510.

[34] 邵平, 安俊琳, 杨辉, 等. 南京北郊夏季近地层臭氧及其前体物体积分数变化特征[J]. 环境科学, 2014, **35**(11): 4031-4043.
Shao P, An J L, Yang H, *et al*. Variation characteristics of surface ozone and its precursors during summertime in Nanjing northern suburb [J]. Environmental Science, 2014, **35**(11): 4031-4043.