文章编号:1008-1534(2019)06-0436-06 开放科学(资源服务)标识码(OSID):

基于 SARIMA-SVR 组合模型的空气质量指数预测

郑洋洋,白艳萍,续 婷

(中北大学理学院,山西太原 030051)

摘 要:空气质量指数(AQI)在波动中既具有整体的时间序列线性特征和明显的季节性波动周期,又具有多种因素影响的不确定性,为了提高 AQI 的预测精度,基于 Ri386 3.3.3 和 Matlab R2014a 两种编程软件,提出了一种同时具有线性和非线性的复合特征的时间序列预测模型——SARIMA-SVR 组合模型。以太原市 2014 年 1 月—2019 年 7 月的 AQI 月均值数据为基础,利用 SARIMA 时间序列模型进行线性预测,利用 SVR 模型对残差进行非线性预测,加和得到组合预测模型的预测结果,分析比较 SARIMA,SVR 和 SARIMA-SVR 这 3 种模型的预测结果和平均绝对百分比误差。结果表明,组合预测模型发挥了 2 种模型各自的优势,相较于单一预测模型的预测结果而言,其预测精度更高,稳定性更好。通过此模型得到的空气质量预测结果不仅可为人们的日常生活提供指导,而且可为大气污染的防治工作提供科学依据和借鉴意义。

关键词:应用数学;SARIMA;SVR;SARIMA-SVR 组合模型;空气质量指数预测

中图分类号:O29;X84 文献标志码:A **doi**: 10.7535/hbgykj.2019yx06011

Air quality index prediction based on SARIMA-SVR combined model

ZHENG Yangyang, BAI Yanping, XU Ting

(School of Science, North University of China, Taiyuan, Shanxi 030051, China)

Abstract: Air quality index (AQI) both has volatility of time series of the whole linear features, obvious seasonal fluctuation cycle, at the same time has a variety of factors of uncertainty. In order to improve the prediction accuracy of AQI, based on Ri386 3.3.3 and Matlab R2014a programming software, this paper proposes a composite characteristics of both linear and non-linear time series prediction model, namely SARIMA-SVR combined model. Based on the monthly average data of AQI from January 2014 to July 2019 in Taiyuan, SARIMA time series model is first used for linear prediction, then SVR model is used for non-linear prediction of residual, and finally the combined prediction model is added and obtained. By analyzing and comparing the prediction results and average absolute percentage errors of SARIMA, SVR and SARIMA-SVR models, the results

收稿日期:2019-09-29;修回日期:2019-10-16;责任编辑:张 军

基金项目:国家自然科学基金(61774137);山西省回国留学人员科研项目(2016-088)

第一作者简介:郑洋洋(1992—),女,山西运城人,硕士研究生,主要从事现代优化算法理论及应用方面的研究。

通信作者:白艳萍教授。E-mail:baiyp666@163.com

郑洋洋,白艳萍,续婷.基于 SARIMA-SVR 组合模型的空气质量指数预测[J].河北工业科技,2019,36(6):436-441.

ZHENG Yangyang, BAI Yanping, XU Ting. Air quality index prediction based on SARIMA-SVR combined model[J]. Hebei Journal of Industrial Science and Technology, 2019, 36(6): 436-441.

show that the combined prediction model gives full play to the advantages of the two models, and its prediction accuracy is higher and its stability is better than that of the single prediction model. The prediction results of air quality by this model can provide reference for the prevention and control of air pollution.

Keywords: applied mathematics; SARIMA; SVR; SARIMA-SVR combination model; air quality index prediction

空气质量状况越来越受到人们的关注，国内外研究人员提出了针对空气质量预测的模型，李婷婷等[1]运用经验模态分解的方法对原始的 AQI 数据进行多尺度分解，运用灰色预测、ARIMA 模型、BP 神经网络和 SVR 等方法进行趋势序列预测，将平均相对误差较小的前 3 种单项预测方法进行组合，进而得到最终预测结果，结果表明，基于经验模态分解的空气质量指数组合预测方法具有较高的预测精度和良好的适用性。还有一些学者运用时间序列模型[2-8]和优化 SVM 的方法[9]完成了空气质量的预测。但是，由于空气质量受多种不确定性因素的影响，单一预测方法没有拟合 AQI 值变化的非线性部分，因此预测误差相对较大。针对这一问题，笔者提出：首先利用 SARIMA 时间序列模型对 AQI 值进行趋势预测，其次运用 SVR 模型对残差进行预测，也就是考虑了不确定性的干扰性因素，并将残差的预测结果与 SARIMA 模型预测的趋势数据进行加和，形成 SARIMA-SVR 组合预测模型，使之更能适应 AQI 时间序列的发展趋势，预测精度更高，为预测空气的 AQI 值提供了新的思路和方法。

1　理论介绍

1.1　SARIMA 模型原理

SARIMA$(p,d,q)\times(P,D,Q)_s$ 模型是在 ARIMA 模型中增加了季节项，称为季节性自回归滑动平均模型，ARIMA 模型主要分析和研究时间序列问题。此模型是先将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立起来的一种模型[10]。ARIMA 模型将预测指标随时间推移而形成的数据序列看作是一个随机序列，这组随机变量所具有的依存关系体现着原始数据在时间上的延续性，它既受外部因素的影响，又有其自身的变动规律。

ARIMA 模型包含自回归模型 AR(p)、移动平均模型 MA(q)和自回归移动平均模型 ARMA(p,q)，ARIMA 模型的形式如下：

$$\begin{cases}\varphi(L)\nabla^d X_t=\Theta(L)\varepsilon_t,\\ E(\varepsilon_i)=0,\mathrm{var}(\varepsilon_i)=\sigma_\varepsilon^2,\\ E(\varepsilon_s\varepsilon_i)=0,\quad s\neq t,\\ E(\varepsilon_s,\varepsilon_i)=0,\quad s<t,\end{cases}\tag{1}$$

式中：$\nabla^d=(1-L)^d$ 为差分运算；$\varphi(L)$为平稳可逆 ARIMA(p,q)模型的自回归系数多项式；$\Theta(L)$为移动平均多项式；d 表示差分次数。

SARIMA 模型与一般的ARIMA模型一样，首先对季节性因素进行 D 阶差分，其次用差分后周期为 s 的季节性时间序列建立一般的 ARIMA 模型，其形式如下：

$$\begin{cases}\varphi(L)A_P(L^s)(\nabla^d\nabla_s^D X_t)=\Theta(L)B_Q(L^s)\varepsilon_t,\\ E(\varepsilon_i)=0,\mathrm{var}(\varepsilon_i)=\sigma_\varepsilon^2,\quad s\neq t,\\ E(\varepsilon_s,\varepsilon_i)=0,\quad s<t,\end{cases}\tag{2}$$

式中：$A_P(L^s)$为 P 阶季节自回归算子；$B_Q(L^s)$为 Q 阶季节移动平均算子；∇_s^D为季节的差分运算。

1.2　SVR 模型原理

支持向量回归机(support vector regression, SVR)是支持向量对于回归问题的算法[11]。支持向量回归的基本原理是将所有数据样本进行学习训练使其都分布在两条直线之间，并且所有点到直线的总偏差最小，求此时两条线之间的最大距离，也就是支持向量回归的最优超平面。使用支持向量机做回归时，其算法过程如下。

对于给定的数据集$\{x_1,x_2,\cdots,x_n\}$和实际值$\{y_1,y_2,\cdots,y_n\}$，希望得到形如 $f(x)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(x)+b$ 的回归公式，使得 $f(x)$与实际值 y 尽可能接近，权重向量 $\boldsymbol{w}$ 与偏置 b 为待确定的模型参数，$\boldsymbol{\varphi}(x)$为非线性映射。引入松弛变量，ε-SVR 问题转化为如下二次规划问题：

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}(\xi_i+\xi_i^*),\tag{3}$$

$$\text{s.t.}\quad f(x_i)-y_i\leqslant\varepsilon+\xi_i,\tag{4}$$

$$y_i-f(x_i)\leqslant\varepsilon+\xi_i^*,\tag{5}$$

$$\xi_i\geqslant 0,\xi_i^*\geqslant 0,\quad i=1,2,\cdots,m,\tag{6}$$

式中：C 为控制惩罚参数；$\min\frac{1}{2}\|\boldsymbol{w}\|^2$ 为复杂函数的控制项；ξ_i,ξ_i^* 均为松弛变量；ε 为不敏感损失因子。通过引入拉格朗日乘子可以得到 SVM 的对偶

问题：

$$\max_{\alpha,\alpha^*}\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)y_i-\sum_{i=1}^{m}(\alpha_j^*+\alpha_j)\varepsilon-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j),\quad(7)$$

式中：α_i，α_i^*，α_j，α_j^* 为拉格朗日乘子；$K(x_i,x_j)$ 为核函数。求解上式可得回归函数：

$$f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)K(x_i,x_j)+b。\quad(8)$$

2 数据采集及分析

此次测试使用的样本数据来源于中国空气质量在线监测分析平台历史数据网(https://www.aqistudy.cn/historydata/)发布的太原市空气质量历史数据，选取的样本数据为太原市 2014 年 1 月—2019 年 7 月的月均值数据，共计 67 组数据。图 1 为太原市月度空气质量指数折线图。由图 1 可以看出，AQI 历史数据具有整体的周期性与季节性，以一年 12 个月为一个循环周期，且季节性也较为明显，可以看出在每年冬季(12 月至次年的 2 月)的 AQI 值最高，夏季(6 月至 8 月)的 AQI 值最低。

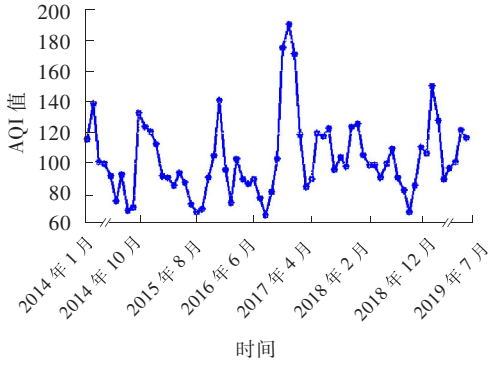

图 1 空气质量指数折线图

Fig.1 Air quality index line chart

3 SARIMA-SVR 组合预测模型原理

由前面分析可以看出，AQI 值序列整体具有时间周期性特征，同时又会有异常数据的出现，因此具有复合特征，即其数据序列包括线性时间序列分量，也包含非线性序列分量。所以单一的预测模型对 AQI 值的预测精度较低，对人类生活的指导意义相对较小。由于 SARIMA 时间序列预测模型对线性预测具有优势，且支持向量回归机(SVR)模型对少量的、非线性的数据预测具有优势，这 2 个预测模型优势互补，因此笔者将二者组合起来进行 AQI 的指数预测。

假设 AQI 数据序列 Y_t 由线性部分 L_t 和非线性部分 N_t 组成，即：

$$Y_t=L_t+N_t,\quad(9)$$

SARIMA-SVM 组合模型[12]预测步骤如下：

1)首先利用 SARIMA 模型对线性部分进行预测，预测结果为 $\hat{L}_t$，将预测结果与原始数据进行差分得到残差部分 N_t，即为原始数据序列的非线性部分；

2)对上一步得到的残差 N_t 进行重构得到 SVR 模型的样本集，并且利用 SVR 模型对残差数据进行预测得到预测结果为 $\hat{N}_t$；

3)最后将 SARIMA 模型预测的线性部分 $\hat{L}_t$ 与 SVR 模型预测的非线性部分 $\hat{N}_t$ 进行加和，得到最终的组合预测结果为 $\hat{Y}_t$，即：

$$\hat{Y}_t=\hat{L}_t+\hat{N}_t。\quad(10)$$

组合原理流程图如图 2 所示。

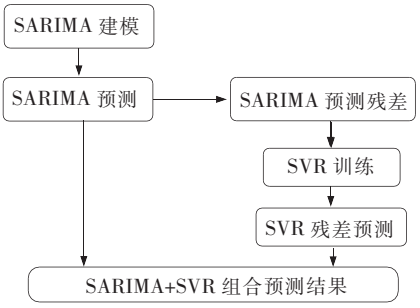

图 2 SARIMA-SVR 组合预测流程图

Fig.2 SARIMA-SVR combination forecasting flow chart

4 空气质量指数预测结果及分析

4.1 SARIMA 模型预测结果

在本节中选取前 4 年共 48 组数据(2014 年 1 月—2017 年 12 月)作为 SARIMA 预测模型的历史数据，由图 3 可知，原始的 AQI 数据有递增趋势，故该序列不平稳，对其进行一次差分后所得折线图如图 4 所示，可以看出差分后的时序图在均值为 0 附近波动，不会随着时间的发展而改变。图 5 和图 6 分别为一次差分后数据序列的自相关图和偏相关图，可以发现，自相关图显示滞后一阶自相关值基本没有超过边界，虽然 5 阶自相关值超出边界，很可能属于偶然现象，偏相关图中在 4 阶时显著不为 0，根据以上自相关和偏相关图进行阶数的一个初步判断，又根据 R 语言中的 auto.arima()函数进行自动定阶，得到最优的 SARIMA 模型为(2,0,0)(0,1,1)[12]，如图 7 为 SARIMA(2,0,0)(0,1,1)[12]模型的预测结果折线图。

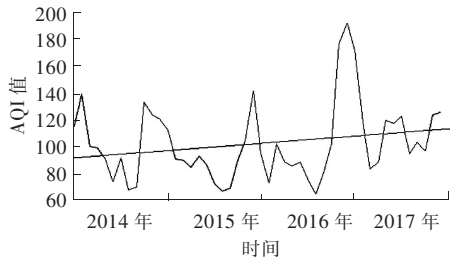

图 3　原始 AQI 时间序列折线图

Fig.3　Original AQI time series line chart

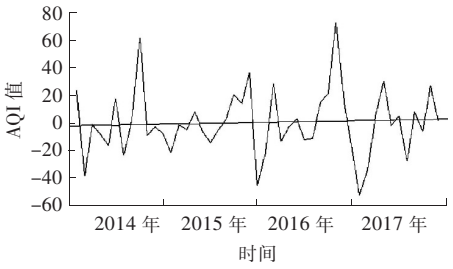

图 4　一阶差分后 AQI 时间序列折线图

Fig.4　AQI time series line chart after first-order difference

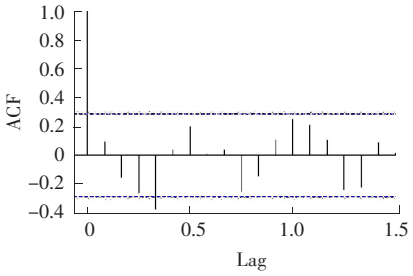

图 5　AQI 指数的 ACF 图

Fig.5　ACF map of the AQI index

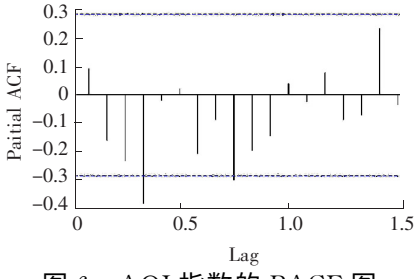

图 6　AQI 指数的 PACF 图

Fig.6　PACF chart of the AQI index

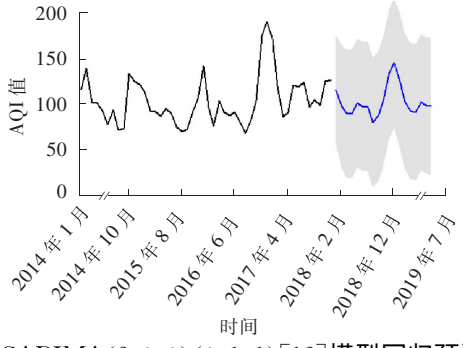

图 7　SARIMA(2,0,0)(0,1,1)[12]模型回归预测结果

Fig.7　SARIMA (2,0,0) (0,1,1)[12] model regression prediction results

4.2　SVR 模型预测结果

采用的数据为太原市 2014 年 1 月至 2019 年 7 月 AQI 月度数据，共 67 组，数据维数为一维，利用前 3 天的 AQI 值来预测后一天的 AQI 值，选取前 44 组的数据作为训练数据集，剩余的数据作为训练集，在 Matlab2014a 软件上进行多次实验，得到最终的预测结果如图 8 所示。

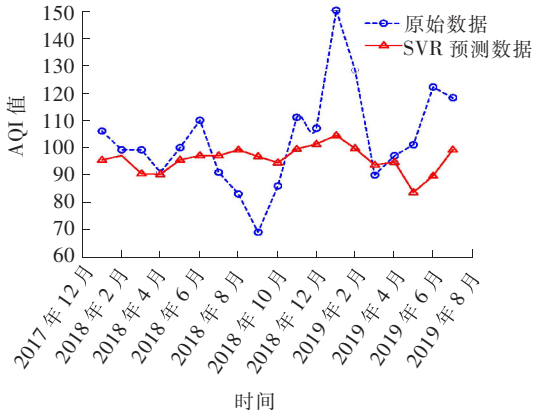

图 8　SVR 模型预测结果

Fig.8　SVR model prediction results

4.3　SARIMA-SVR 组合模型预测结果

SARIMA 模型能很好地捕捉到时间序列的周期性，如图 9 所示。

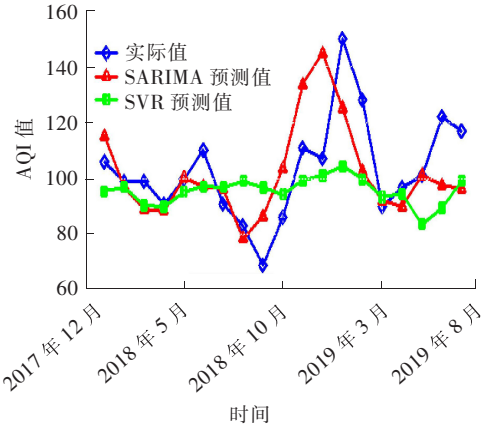

图 9　单一模型预测值与实际对比图

Fig.9　Single model predictive value and actual comparison chart

由图 9 可以看出，单一模型对峰值的预测存在较大误差，SVR 模型能较好地解决小样本、非线性、高维数和局部极小点等实际问题。考虑到 2 种预测方法各有其优点，为了进一步从原始数据序列中提取出更多信息，避免有效信息的浪费，提高预测精度，笔者将 2 种预测方法结合起来，形成 SARIMA-SVR 组合预测模型，对原始数据进行重新预测。首先，采用 SARIMA 模型对原始数据序列进行预测，然后形成预测残差，并将其作为 SVR 模型的样本

集，用前 4 组的残差来预测后一个残差，形成训练集和测试集，再利用训练好的 SVR 模型对残差进行预测，预测结果包含了数据序列的非线性规律和信息，最后，将 SARIMA 模型的预测结果与 SVR 模型的预测结果进行加和，得到 SARIMA-SVR 组合预测模型的预测结果，如表 1 和图 10 所示。

表 1 预测结果对比

Tab.1 Comparison of prediction results

时间	AQI 值	SARIMA			SVR			SARIMA+SVR		
		预测结果	残差	预测相对误差/%	预测结果	预测相对误差/%	残差预测结果	预测结果	组合预测残差	预测相对误差/%
2018-01	106	115.04	−9.04	−8.53	95.47	−9.94				
2018-02	99	97.03	1.97	1.99	96.93	−2.09				
2018-03	99	89.03	9.97	10.07	90.44	−8.65				
2018-04	91	88.47	2.53	2.78	90.02	−1.07				
2018-05	100	100.36	−0.36	−0.36	95.28	−4.72	7.08	107.44	7.44	7.44
2018-06	110	97.17	12.83	11.67	97.07	−11.76	−4.15	93.02	−16.98	−15.44
2018-07	91	96.29	−5.29	−5.81	97.01	6.60	4.64	100.93	9.93	10.91
2018-08	83	78.60	4.40	5.30	99.03	19.31	3.76	82.36	−0.64	−0.77
2018-09	69	86.34	−17.34	−25.14	96.81	40.31	−7.39	78.95	9.95	14.42
2018-10	86	103.58	−17.58	−20.45	94.35	9.71	−17.48	86.10	0.10	0.12
2018-11	111	133.47	−22.47	−20.24	99.25	−10.58	−22.37	111.10	0.10	0.09
2018-12	107	144.58	−37.58	−35.12	101.12	−5.50	−29.35	115.23	8.23	7.69
2019-01	150	125.19	24.81	16.54	104.37	−30.42	24.48	149.67	−0.33	−0.22
2019-02	128	102.44	25.56	19.97	99.70	−22.11	25.46	127.90	−0.10	−0.08
2019-03	90	91.91	−1.91	−2.12	93.44	3.83	−1.81	90.10	0.10	0.11
2019-04	97	90.00	7.00	7.21	94.50	−2.58	7.10	97.10	0.10	0.10
2019-05	101	101.17	−0.17	−0.17	83.64	−17.19	−0.07	101.10	0.10	0.10
2019-06	122	97.60	24.40	20.00	89.54	−26.61	0.43	98.03	−23.97	−19.65
2019-07	117	96.52	20.48	17.51	99.15	−15.97	20.37	116.89	−0.11	−0.09

注：SARIMA 模型预测的 MAPE 为 12.16%；SVR 模型预测的 MAPE 为 13.10%；SARIMA+SVR 模型预测的 MAPE 为 5.15%。

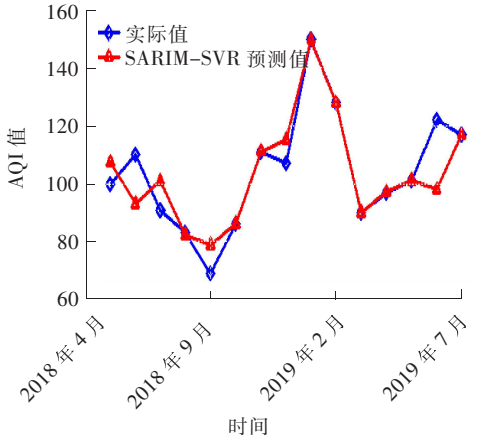

图 10 SARIMA-SVR 组合模型预测值与实际对比图

Fig.10 SARIMA-SVR combined model predictive value and actual comparison chart

由表 1 可以看出，SVR 模型对原始数据的规律捕捉和刻画能力最差，MAPE 预测的平均绝对误差为 13.10%，SARIMA 模型对数据的规律刻画能力较强，通过历史数据发现，AQI 值在不同月份的变化规律与 SARIMA-SVR 组合模型的预测结果较为相似，且在个别几个月中，SARIMA 模型的预测误差要比 SARIMA-SVR 模型的预测误差小，但是 SARIMA-SVR 组合模型的预测稳定性更好，整体的平均绝对百分误差 MAPE 在 3 种预测模型中最小，为 5.15%，因此，组合预测模型的整体预测精度要比单一预测模型的预测精度高。

最后，利用 SARIMA-SVR 组合预测模型对太原市 2019 年 8−12 月进行预测，最终的预测结果如表 2 所示。

表 2 SARIMA-SVR 模型对太原市 2019 年未来 5 个月的空气质量指数预测值

Tab.2 SARIMA-SVR model for the prediction of air quality index for the next 5 months of Taiyuan City in 2019

2019 年 8 月	2019 年 9 月	2019 年 10 月	2019 年 11 月	2019 年 12 月
88.22	105.43	134.39	157.10	125.98

由表 2 可以看出太原市在 8−9 月的月度平均

空气质量指数为良好，适宜人们出行，10 月份为轻度污染，而 11－12 月份的平均空气质量指数较高，相关部门的注意应提前做好环境保护工作，防止太原市的空气质量进一步恶化。

5 结　语

单一预测模型由于自身条件的限制，在对数据序列进行预测时，不能全面掌握数据信息而影响预测结果，因此，将 2 种单一预测模型进行组合，发挥其优势互补作用得到的组合预测模型的预测精度往往更高。

1）运用基于 SARIMA 模型和 SVR 模型相结合的 SARIMA-SVR 组合预测模型，对太原市空气质量预测的实例分析，验证了该算法的有效性。

2）在数据比较平缓的变化中出现峰值时，SVR 模型会产生比较大的误差，这说明 SVR 模型在捕捉数据规律方面存在着不足。SARIMA 模型对空气质量指数的变化规律和季节波动影响的捕捉能力较强，将 2 种单一模型组合起来得到的 SARIMA-SVR 组合模型可以综合利用 2 种单一模型所提供的信息，有效地提高了预测精度。

3）SARIMA-SVR 组合预测模型减少了预测的系统误差。实证研究表明，基于 SARIMA-SVR 组合模型对预测太原市空气质量指数是有效的，组合预测模型的精度明显优于单一预测模型的精度。

4）空气质量指数不会受到污染物浓度、气象因素、车流量、工厂排放等多种因素的影响，在实证研究过程中，笔者没有考虑这些因素，仅是对 AQI 值进行了趋势预测，因此，在将来的研究中，应加大对影响因素的研究，利用智能优化算法优化参数，进一步提高预测精度。

参考文献/References：

[1] 李婷婷，田瑞琦，汪漂.基于经验模态分解的空气质量指数组合预测方法及应用[J].价值工程，2019(16)：134-138.
LI Tingting，TIAN Ruiqi，WANG Piao.Air quality index combined prediction method based on EMD and its application[J].Value Engineering，2019(16)：134-138.

[2] 敖希琴，张怡文，陈家丽，等.基于季节性时间序列模型的合肥地区空气质量分析及预测[J].合肥学院学报(综合版)，2018，35(5)：33-39.
AO Xiqin，ZHANG Yiwen，CHEN Jiali，et al. Analysis and prediction of air quality in Hefei Area based on seasonal time series model[J].Journal of Hefei University(Comprehensive Edition)，2018，35(5)：33-39.

[3] 王坤，阮金梅，邓妮.基于 SARIMA 模型的曲靖市空气质量指数预测[J].曲靖师范学院学报，2018，37(3)：25-29.
WANG Kun，RUAN Jinmei，DENG Ni. Prediction of air quality index in Qujing based on SARIMA model[J].Journal of Qujing Normal University，2018，37(3)：25-29.

[4] VENTURA L M B，PINTO F D O，SOARES L M，et al. Forecast of daily $PM_{2.5}$ concentrations applying artificial neural networks and Holt-Winters models[J]. Air Quality，Atmosphere and Health，2019，12(3)：317-325.

[5] 孟庆云，张若晴，袁朱红，等.基于 ARIMA 模型的天津市空气质量各项指标的预测分析[J].农业灾害研究，2018，8(5)：44-45.
MENG Qingyun，ZHANG Ruoqing，YUAN Zhuhong，et al. Prediction and analysis of air quality indicators in Tianjin based on ARIMA model[J].Journal of Agricultural Catastrophology，2018，8(5)：44-45.

[6] 王涛，王凤兰，王悦婷.基于时间序列模型的 $PM_{2.5}$ 研究[J].智库时代，2018(35)：192-193.

[7] POHOATA A，LUNGU E. A complex analysis employing ARIMA model and statistical methods on air pollutants recorded in Ploiesti，Romania[J]. Revista de Chimie-Bucharest-Original Eolition，2017，68(4)：818-823.

[8] WU Lifeng，GAO Xiaohui，XIAO Yanli，et al. Using grey Holt-Winters model to predict the air quality index for cities in China[J].Natural Hazards，2017，88(2)：1003-1012.

[9] 高帅，胡红萍，李洋，等.基于 MFO-SVM 的空气质量指数预测[J].中北大学学报(自然科学版)，2018，39(4)：373-379.
GAO Shuai，HU Hongping，LI Yang，et al. Prediction of air quality index based on MFO-SVM[J]. Journal of North University of China(Natural Science Edition)，2018，39(4)：373-379.

[10] 汤银英，朱星龙，李龙.基于 SARIMA 模型的铁路月度客运量预测[J].交通运输工程与信息学报，2019，17(1)：25-32.
TANG Yinying，ZHU Xinglong，LI Long. Monthly railway passenger traffic volume forecasting based on SARIMA model[J]. Journal of Transportation Engineering and Information，2019，17(1)：25-32.

[11] 邓建球，赵建忠，陈洪，等.ABC 算法优化 SVR 的磨损故障预测模型[J].兵工自动化，2018，37(10)：60-64.
DENG Jianqiu，ZHAO Jianzhong，CHEN Hong，et al. Wear faults prediction model based on SVR optimized by ABC[J]. Ordnance Industry Automation，2018，37(10)：60-64.

[12] 程虎彪，姜大立.基于 SARIMA-SVM 组合模型的战时军用物资需求预测[J].军事运筹与系统工程，2016，30(2)：45-49.