山西大学学报(自然科学版)48(4):829—838,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024045

基于季节趋势分解的$PM_{2.5}$浓度混合预测模型

王平[1*],许濒月[1],雷卓祎[2],张贵生[3],吴青东[1]
(1.山西财经大学 资源环境学院,山西 太原 030006;
2.兰州大学 信息科学与工程学院,甘肃 兰州 730000;
3.山西大学 环境与资源学院,山西 太原 030006)

摘 要:为提高$PM_{2.5}$浓度预测精度,提出了基于季节趋势分解的时间序列混合预测模型(Hybrid-X12)。首先,使用季节趋势分解算法将$PM_{2.5}$时序分解为趋势-循环、季节和不规则子序列;然后,分别使用自回归移动平均模型(Autoregressive Integrated Moving Average Model,ARIMA)、长短期记忆网络(Long Short Term Memory,LSTM)和支持向量机(Support Vector Machine,SVM)对以上子序列进行预测;最后,集成子序列结果得到最终的预测结果。仿真实验选用华北地区主要六个城市$PM_{2.5}$月浓度数据,平均绝对误差(Mean Absolute Error,MAE)、均方根误差(Root Mean Square Error,RMSE)和一致性指数(Index of Agreement,IA)为模型评价指标。实验结果证明混合预测模型能明显提高预测精度,与传统单一模型ARIMA、LSTM和SVM相比,以北京为例,MAE分别降低了18.72%、60.14%和43.15%,验证了季节趋势分解算法有助于时序季节趋势信息挖掘,针对不同特征子序列选择合适的算法充分发挥不同模型优势,为$PM_{2.5}$浓度预测提供了新思路。

关键词:$PM_{2.5}$浓度预测;季节趋势分解;自回归移动平均模型;长短期记忆网络;支持向量机
中图分类号:P9 **文献标志码**:A **文章编号**:0253-2395(2025)04-0829-10

A Hybrid Prediction Model for $PM_{2.5}$ Concentration Based on Seasonal Trend Decomposition

WANG Ping[1*], XU Binyue[1], LEI Zhuoyi[2], ZHANG Guisheng[3], WU Qingdong[1]
(1. *College of Resources and Environment, Shanxi University of Finance and Economics, Taiyuan* 030006*, China;*
2. *School of Information Science and Engineering, Lanzhou University, Lanzhou* 730000*, China;*
3. *School of Environmental & Resource Science, Shanxi University, Taiyuan* 030006*, China*)

Abstract: To improve the accuracy of $PM_{2.5}$ concentration prediction, this paper introduces a hybrid time series forecasting model with seasonal-trend decomposition (Hybrid-X12). Firstly, the seasonal-trend decomposition algorithm performs the task of decomposing $PM_{2.5}$ time series into trend-cycle, seasonal and irregular sub-series; Then, ARIMA (Autoregressive Integrated Moving Average Model), LSTM (Long Short Term Memory), and SVM (Support Vector Machine) are applied to the above sub-series prediction tasks respectively; Finally, the final prediction result comes from the integration of the predicted results of sub-series. The simulation experiment selected $PM_{2.5}$ monthly concentration from six major cities in North China and used MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and IA (Index of Agreement) as model evaluation indicators. The experimental results demonstrated that the hybrid prediction system can significantly enhance prediction accuracy. Compared with the traditional single model ARIMA, LSTM and SVM, the MAE of the proposed model in Beijing is reduced by 18.72%, 60.14% and 43.15%, respectively. This verifies that the seasonal-trend decomposition algorithm is helpful for mining seasonal-trend information in time series. It can be concluded

收稿日期:2023-10-01;**接受日期**:2024-03-07
基金项目:国家社会科学基金(20BTJ045)
* **通信作者**:王平(1978—),女,山西晋中人,博士,副教授,研究方向为大气环境建模。E-mail:wp2004@sxu.edu.cn
引文格式:王平,许濒月,雷卓祎,等.基于季节趋势分解的$PM_{2.5}$浓度混合预测模型[J].山西大学学报(自然科学版),2025,48(4):829-838. DOI:10.13451/j.sxu.ns.2024045.

that selecting appropriate algorithms for sub-series with different characteristics ensures the full utilization of the advantages of different models, providing new ideas for $PM_{2.5}$ concentration prediction.

Key words: $PM_{2.5}$ concentration prediction; seasonal trend decomposition; autoregressive integrated moving average model; long short term memory; support vector machine

0 引言

城市化和工业化的快速发展带来了日益严重的大气污染问题,特别是北方地区冬季采暖及气象条件的叠加往往导致污染事件频发,严重影响人民的身体健康和生产的有序进行[1-2]。$PM_{2.5}$作为大气污染物监测的重要指标,是形成雾霾的重要成分,其浓度在很大程度上反映了大气环境质量,因此,准确、高效预测$PM_{2.5}$浓度对于探索大气污染物浓度变化规律,挖掘大气污染形成的内在原因,制定切实有效的大气环境管理措施有重要意义[3-5]。目前空气质量预测模型有两种。一种为机理模型,以地形和污染源时空分布信息为基础,污染物与气象的相互作用为驱动,构建复杂空气污染模型,实现多尺度和多源空气质量预测[6-8]。然而在实际应用中往往由于先验知识不足,导致机理模型建模受到很大的限制。另一种为统计模型,相比于机理模型,该类模型智能利用污染物和气象历史数据,具有更简单的建模方法和更好的性能。例如Zhao等将AIC(Akaike information criterion)、GS(Grid search)固定顺序方法和季节分解集成,建立了混合自回归移动平均(Autoregressive Integrated Moving Average Model, ARIMA)预测模型,相比于传统ARIMA模型,有效提高了北京市$PM_{2.5}$年均浓度预测精度,为中长期大气环境管理政策制定提供了可靠依据[9]。梁泽等使用遗传算法实现人工神经网络建模参数优化,完成北京日均$PM_{2.5}$浓度预测,仿真实验结果显示该方法系统变量较少并且具有较好的预测精度[10]。Mahajan等构建了一种基于聚类的混合神经网络模型,用于台湾省$PM_{2.5}$预测服务,并分析了不同聚类方案对于预测系统在精度和效率方面的影响[11]。Zhou等以多输出支持向量机(Multi-output Support Vector Machine, Multi-output SVM)和多任务学习算法(Multi-Task Learning, MTL)为基础,构建了一种新的预测框架MMSVM((MTL Multi-output SVM)),有效提高了台北市区域$PM_{2.5}$多步预报的时空稳定性和准确性[12]。Zhang等使用一种组合遗传算法支持向量机,以气溶胶光学厚度、气象数据和地形数据等为数据基础,估计了陕西省2015年$PM_{2.5}$浓度,并且分析了不同季节和年份的空间聚类模式[13]。通过分析历史空气质量数据及气象等影响因素,建立多元线性或非线性回归模型,准确预测大气污染物浓度,并且预测模型表现出较强的鲁棒性。随着大气环境监测数据容量的快速增加,大数据技术也成功应用于空气质量预测。Changhoi等提出通过递归神经网络算法与CMAQ(Community Multiscale Air Quality)相结合实现传统模型改进,预测系统输入数据根据时间排序,体现了$PM_{2.5}$时序的内在特征[14]。Li等挖掘卷积神经网络可以有效提取与空气质量相关特征和长短期记忆网络(Long Short Term Memory, LSTM)可以反映输入时序数据的历史信息的优点,构建$PM_{2.5}$浓度混合预测系统,分析了在不同输入情况下的模型泛化能力[15]。

由以上研究可以看出统计模型应用于$PM_{2.5}$浓度预测,能很好地拟合历史数据与预测目标之间的高维非线性关系,并且例如LSTM模型能刻画时序数据的在时间维度的相关性特征,这些都有助于提高预测模型的精度。$PM_{2.5}$浓度时序,特别是月均浓度数据,具有强烈的季节性、趋势性等特征,这些信息融合于统计模型,无疑将有效提高预测模型泛化能力。本文基于大气污染物浓度时序的季节趋势分解子序列特征,提出了一种针对不同特征子序列使用不同适应算法的混合预测模型,利用华北地区主要城市$PM_{2.5}$月均浓度数据进行仿真实验,并与单模型ARIMA、LSTM和SVM预测结果比较。实验结果证明提取季节趋势信息,并根据不同子序列选择适宜的模型,然后合并为最终预测结果,可以有效提高模型精度和稳定性。

1 研究数据

本文采用华北地区北京、天津、郑州、济南、石家庄和太原2014年至2021年$PM_{2.5}$月浓度数据作为实验数据,来源于网站(http://www.tianqihoubao.com/),每个站点包含96组数据。其中2014年1月至2017年12为训练集,2018年1月至2021年12月为测试集。图1为六个站点的$PM_{2.5}$浓度,可以看出所有站点的时序数据均呈现出较强的季节性,并且频率基本保持一致。此外,$PM_{2.5}$浓度随时间下降的趋势也比较明显。华北地区冬季$PM_{2.5}$浓度最高,而夏季最低,这与冬季燃煤采暖及温带季风气候相关,导致不同季节空气污染物的排放、传输和扩散等存在显著季节性变化[10]。并且随着《大气污染防治行动计划》[16]等政策的执行,华北地区空气质量逐年改善,由图也可看出所有监测站点的$PM_{2.5}$浓度呈现下降趋势。因此,预测系统融合时序季节性和趋势性特征将增加预测先验知识,有效提高模型泛化能力。2016年12月石家庄$PM_{2.5}$浓度达到276 μg/m³,为研究数据期间最高浓度,在12月石家庄发生多次空气污染事件,特别是12月19日石家庄遭遇严重雾霾天气,空气污染指数达470,由此12月均$PM_{2.5}$浓度较高。

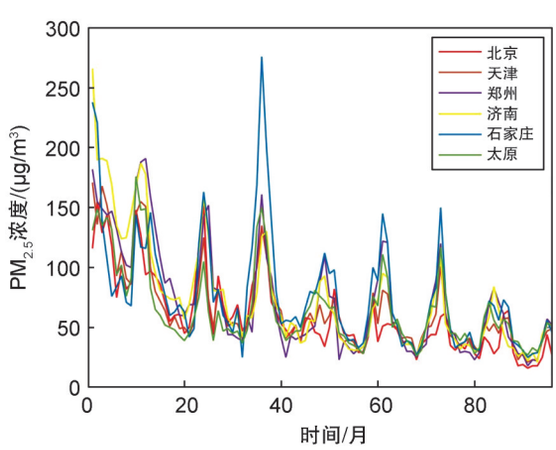


图1 2014年1月至2021年12月$PM_{2.5}$月浓度时序图

Fig. 1 $PM_{2.5}$ monthly concentration time series from January 2014 to December 2021

2 研究方法

2.1 时间序列季节趋势分解

时间序列是按时间顺序进行的一系列采样结果,可以表示为$Y_t(t=1,\cdots,n)$,X12方法通过多次迭代的移动平均方法,原时序可以分解为趋势-循环因素(TC_t)、季节因素(S_t)和不规则因素(I_t)3个子序列,如公式(1)所示,不同子序列表现出不同的时序特征。由于气候和地形差异等因素影响,时间序列的组成不能保证相互独立,并且$PM_{2.5}$时序每年同月的季节周期分量呈现出明显变化[17-18],因此本文选用乘法模型表示趋势子序列、季节子序列和不规则子序列的组合形式,即:

$$Y_t=TC_t\times S_t\times I_t。\tag{1}$$

2.2 不同子序列预测模型

2.2.1 ARIMA

ARIMA模型是一种常用的时序预测模型[19],由历史项构成的自回归(AR)部分、误差回归构成的移动平均(MA)过程和保证时序平稳性的差分组成,表示形式为ARIMA(p,d,q),其中"p"为自回归阶数,"d"表示差分阶数,"q"则代表移动平均阶数,如公式(2)所示[20]。

$$\left(1-\sum_{i=1}^{p}\varphi_iB^i\right)(1-B)^dX_t=\left(1+\sum_{i=1}^{q}\theta_iB^i\right)\varepsilon_t,\tag{2}$$

其中φ_i和θ_i分别为AR和MA参数,B是滞后算子,X_t代表时间序列,ε_t表示误差项。

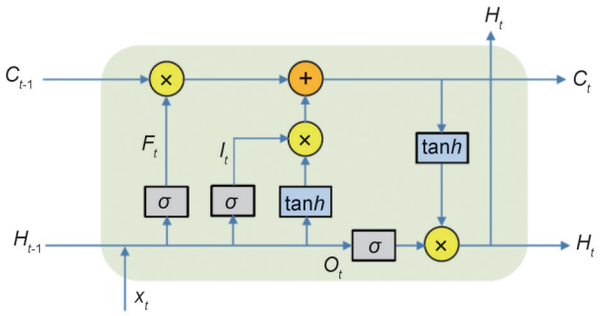
2.2.2 LSTM模型

LSTM模型相较于循环神经网络模型,克服了梯度消失和梯度爆炸问题,通过门机制维持cell单元状态,解决短期和长期的记忆依赖问题[21]。LSTM基本单元如图2所示,包含三个门单元,分别为输入门、遗忘门和输出门[18,21-22],在图中分别表示为I_t、F_t和O_t;C_t表示单元状态;σ为sigmoid函数。

2.2.3 SVM模型

SVM模型以结构化风险最小化为目标,避免了过拟合问题,并且引入核函数将非线性问题转化为在高维空间中的线性可分问题[23-24]。SVM用于回归分析可以表示为以下形式。

$$\begin{aligned}&\min\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\\&\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i),\\&s.t.\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0,0\leqslant\alpha_i,\\&\alpha_i^*\leqslant\frac{C}{l},i=1,2,\cdots,l,\end{aligned}\tag{3}$$



注:I_t、F_t和O_t分别为输入门、遗忘门和输出门;C_t为单元状态;σ为sigmoid函数。

Note: I_t, F_t and O_t are input gate, output gate and forget gate, respectively; C_t is the cell state; σ is the sigmoid function.

图2 LSTM基本单元

Fig. 2 Structure of the LSTM

其中α_i和α_i^*为拉格朗日乘子,$K(x_i,x_j)$为核函数,求解以上优化问题可得以下最优超平面回归方程:

$$f(x)=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)K(x_i,x_j)+b。 \quad (4)$$

2.2 基于季节趋势分解的混合预测模型(Hybrid-X12)

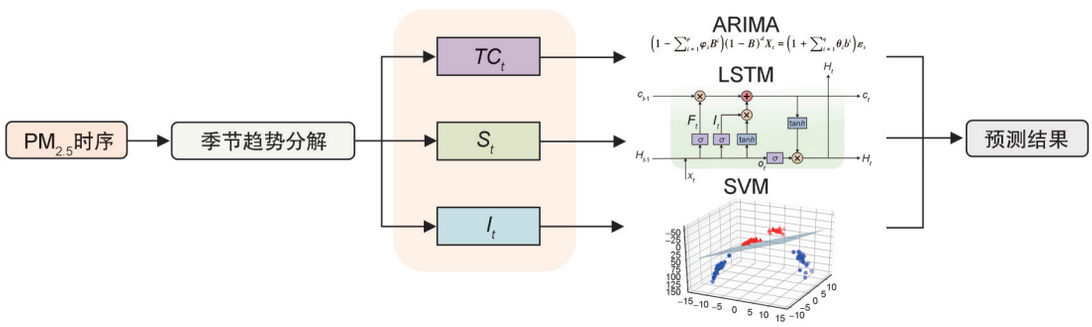
由于季节趋势分解可以提取时间序列的趋势循环、季节性和不规则因素,有助于客观分析时间序列的运动规律,并且分解得到的子序列各自具有独特的结构特征,据此可以选择适宜的分析方法,因此合理利用季节趋势分解可以提高预测模型精度。本文提出的混合预测模型的优势在于集合了不同模型的优点,利用最有效的方法处理每个子过程。首先使用Census X12对$PM_{2.5}$时序数据进行分解,得到三个子序列,分别为:趋势-循环子序列(TC_t)、季节子序列(S_t)和不规则子序列(I_t)。趋势-循环子序列(TC_t)表现出较强的线性形态,因此选择同样具有线性结构的ARIMA模型处理TC_t子序列,分析数据内在线性特征。季节子序列(S_t)具有很强的季节特性,数据在时间维度上的关联性较强,使用LSTM算法完成S_t子序列预测,能实现信息在序列中传输,充分利用采样数据的间隔和延迟较长的特点。不规则子序列(I_t)波动复杂,非线性特征明显,因此使用SVM模型表示I_t子序列的非线性特征。最后将三个子序列分别得到的预测结果相乘作为最终结果。图3给出了季节趋势混合预测模型的示意图,详细建模过程见算法1。

2.3 算法性能指标

为了验证基于季节趋势分解的混合预测模型相较于传统模型具有更好的泛化能力,本文选择三个常用的模型性能统计指标进行评估,分别为MAE、RMSE和IA,具体公式如下表示:

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i'-y_i\right|, \quad (5)$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i'-y_i)^2}, \quad (6)$$



注:TC_t:趋势-循环子序列;S_t:季节子序列;I_t:不规则子序列。

Note: TC_t: trend-cycle sub-series; S_t: seasonal sub-series; I_t: irregular sub-series.

图3 Hybrid-X12混合预测模型示意图

Fig. 3 Flowchart of hybrid-X12 model

算法1 基于季节趋势分解的$PM_{2.5}$浓度混合预测模型

输入:由$PM_{2.5}$浓度数据构成训练集$\{(X_i,Y_i)\}_{i=1}^{t}$,其中$X_i=PM_{2.5_{i-1}}$,$Y_i=PM_{2.5_i}$,及$t+1$时刻预测系统输入$X_{t+1}=PM_{2.5_t}$。

输出:$t+1$时刻$PM_{2.5}$污染物浓度预测结果$f(X_{t+1})$。

Step1:使用X12季节性调整算法将$PM_{2.5}$浓度构成的时序$\{PM_{2.5_1},\cdots,PM_{2.5_i},\cdots,PM_{2.5_{t-1}}\}$分解为趋势-循环子序列$\{TC_1,\cdots,TC_i,\cdots,TC_{t-1}\}$、季节子序列$\{S_1,\cdots,S_i,\cdots,S_{t-1}\}$和不规则子序列$\{I_1,\cdots,I_i,\cdots,I_{t-1}\}$,及$t+1$时刻$PM_{2.5_t}$的分解结果$TC_t$、$S_t$和$I_t$。

Step2:使用趋势-循环子序列$\{TC_1,\cdots,TC_i,\cdots,TC_t\}$训练ARIMA模型,根据AIC准则,确定模型参数$p$,$d$和$p$的值,建立预测模型。

Step3:季节子序列$\{S_1,\cdots,S_i,\cdots,S_t\}$构成子序列训练集$\{(X_i^S,Y_i^S)\}_{i=1}^{t}$,其中$X_i^S=S_{i-1}$,$Y_i^S=S_i$,在此训练集上训练LSTM模型,构建最优网络结构。

Step4:不规则子序列$\{I_1,\cdots,I_i,\cdots,I_t\}$构成子序列训练集$\{(X_i^I,Y_i^I)\}_{i=1}^{t}$,其中$X_i^I=I_{i-1}$,$Y_i^I=I_i$,训练SVM模型,优化核函数及相关参数。

Step5:使用Step2中训练好的ARIMA模型进行静态预测,得到$t+1$时刻预测结果TC_{t+1}。

Step6:将Step1中得到的$t+1$时刻的分解结果S_t和I_t,分别输入训练好的LSTM和SVM模型,得到$t+1$时刻的预测结果Y_{t+1}^S(令$S_{t+1}=Y_{t+1}^S$)和Y_{t+1}^I(令$I_{t+1}=Y_{t+1}^I$)。

Step7:将之前得到的各子序列预测结果代入公式$Y_{t+1}=TC_{t+1}\times S_{t+1}\times I_{t+1}$,可得$t+1$时刻最终预测结果$f(X_{t+1})$。

重复步骤Step1-Step7可以得到$t+2,\cdots,t+T$时刻的预测结果$f(X_{t+2}),\cdots,f(X_{t+T})$。

$$IA=1-\frac{\sum_{i=1}^{n}(y_i'-y_i)^2}{\sum_{i=1}^{n}(|y_i'-\bar{y}|+|y_i-\bar{y}|)^2},\quad(7)$$

其中n为样本容量,y_i为观测值,y_i'为预测系统的预测结果。以上指标中MAE和RMSE反映预测结果与真值之间的误差,指标值越小表示预测越接近真值。IA则描述了预测结果与观测值之间的相关性,指标值越接近1表示两者的相关性越高。由此可以看出以上三个模型性能指标可以较全面评价预测模型泛化能力。

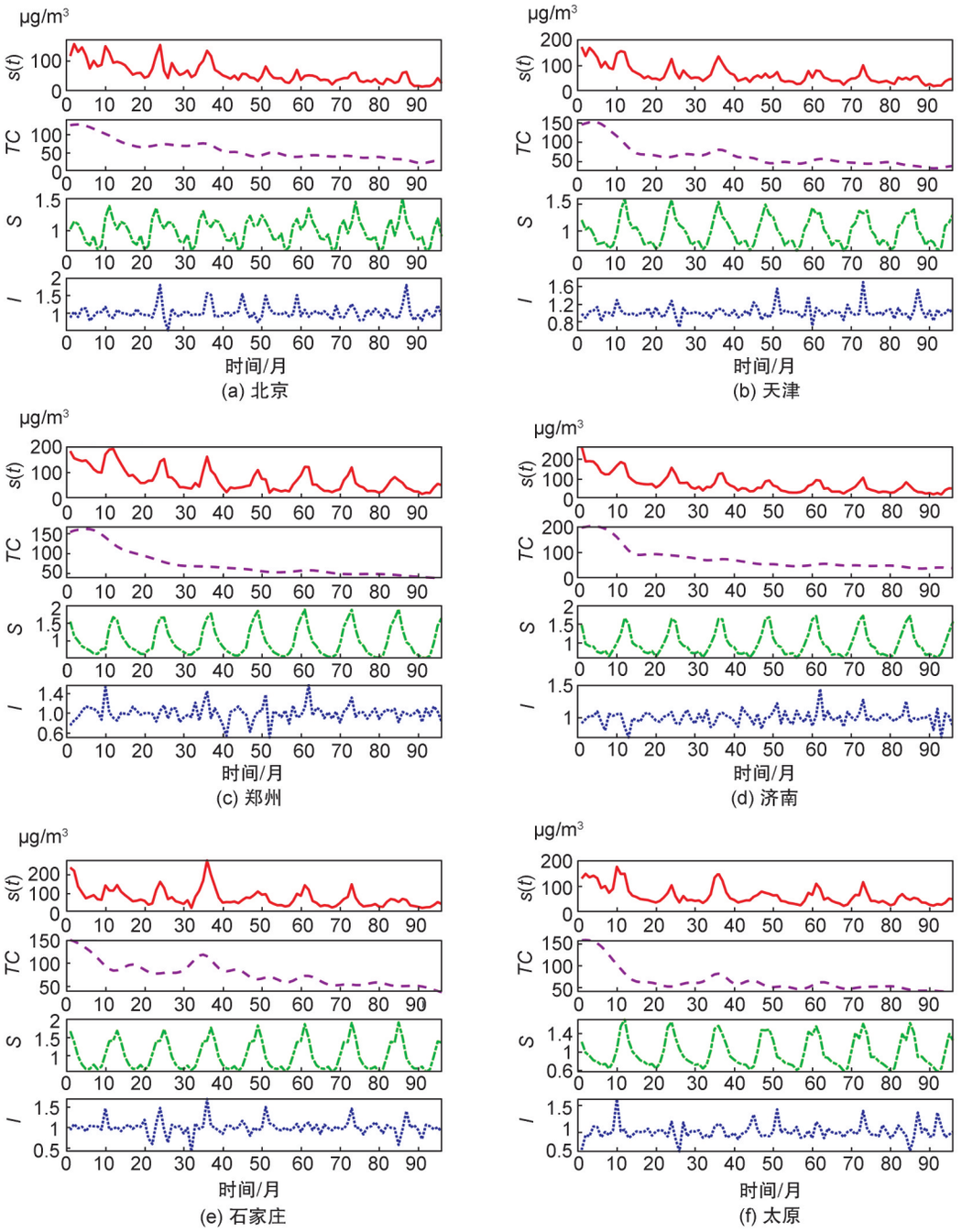
3 结果与分析

3.1 $PM_{2.5}$浓度时序季节趋势分解结果分析

通过X12季节趋势分解算法,$PM_{2.5}$浓度时序可以分解为分别表示趋势-循环、季节和不规则信息的子序列TC_t、S_t和I_t,图4为北京等六个城市2014年至2021年$PM_{2.5}$月浓度的分解结果。由图可以看出代表趋势-循环子序列的TC序列呈现显著的下降趋势,特别是天津和济南站点,分别由145.04和195.79下降至38.18和38.43,这是由于各级人民政府制定《大气污染防治条例》,将大气环境保护工作纳入国民经济和社会发展计划,并有效执行该政策,显著改善了大气环境质量。此外可以发现,六个研究站点的$PM_{2.5}$均表现出较强的季节特征,并且具有相近的频率,这也证明了华北地区$PM_{2.5}$浓度受季节因素,特别是气象条件的影响。时间序列中的不确定对时间序列的预测有着非常重要的影响,而提取出的趋势信息和季节信息可以减少这种不确定,本研究使用互信息给出定量表示。互信息可以表征一个随机变量包含的另一个随机变量的信息量,分析由季节趋势分解得到的子序列与预测变量的统计相关性。表1列出了季节趋势分解得到的各子序列与$PM_{2.5}$浓度的互信息值,趋势-循环子序列与$PM_{2.5}$相关性最强,其决定了时序的持续长期变化,季节子序列和不规则子序列也与原始时序存在相关性,因此分解得到的子序列均作为混合预测系统的输入。

3.2 子序列预测结果分析

将实验数据分为训练集和测试集,分别包含96和48组数据,由训练集完成模型建模,测试集验证模型预测精度。表2为季节趋势分解后各子序列在不同预测模型下的预测精度,由MAE指标表示。根据之前混合模型构建要求,趋势-循环子序列、季节子序列和不规则子序列分别使用ARIMA、LSTM和SVM模型建模。由表2显示的结果可以看出,ARIMA模型在训练集上的预测精度是低于测试集的,这是由于模型在训练集上建模后,应用于测试集时,测试集样本不断补充进建模数据集,随着样本量的增加,模型精度也不断提高。同样的情况也出现于用于季节子序列建模,由于LSTM模型更加适用于大数据集并且通过门函数实现时序特征传递,因此,在测试集上表现出更优的泛化能力。对于不规则子序列使用训练集优化SVM参数,然后用于测试集预测,由于训练集样本量保持不变,训



注:$s(t)$:原始时序;TC:趋势-循环子序列;S:季节子序列;I:不规则子序列。

Note: $s(t)$: $PM_{2.5}$ original time series; TC: trend-cycle sub-series; S: seasonal sub-series; I: irregular sub-series.

图4 $PM_{2.5}$浓度时序季节趋势分解结果

Fig. 4 Seasonal-trend decomposition results of $PM_{2.5}$ concentration time series

练集的拟合精度要优于测试集。依据季节趋势分解的不同子序列的特征,选择不同时序预测方法,挖掘各子序列的内在信息,尽可能准确预测子序列,从而提高原时序的预测精度。

3.3 $PM_{2.5}$浓度时序预测结果分析

根据公式(1)将各子序列的预测结果相乘就可以获得最终的时序预测结果。为了更好地说明Hybrid-X12模型相较于传统模型具有更好的泛化能力,本文选择ARIMA、LSTM和SVM为基准模型,预测结果的评价指标值显示于表3。Hybrid-X12模型的MAE和RMSE显著低于ARIMA、LSTM和SVM,而IA则得到明显提高。以北京为例Hybrid-X12模型的MAE和

表1 季节趋势分解子序列与$PM_{2.5}$浓度互信息值

Table 1 The mutual information values between seasonal-trend decomposition subseries and $PM_{2.5}$ concentration

子序列	北京	天津	郑州	济南	石家庄	太原
趋势—循环子序列(TC_t)	4.773 6	4.443 7	4.844 2	4.816 1	5.067 8	4.241 1
季节子序列(S_t)	0.637 5	0.792 0	0.862 3	0.846 5	0.792 4	0.747 8
不规则子序列(I_t)	0.694 9	0.679 7	0.637 6	0.712 2	0.809 7	0.530 8

表2 季节趋势分解子序列训练集和测试集拟合精度(MAE)

Table 2 Fitting accuracy (MAE) of training and testing sets for seasonal-trend decomposition subseries

子序列	子集	北京	天津	郑州	济南	石家庄	太原
趋势—循环子序列(TC_t)	训练集	0.598 1	0.386 0	0.225 8	0.588 6	0.551 3	0.930 8
	测试集	0.441 9	0.242 6	0.105 9	0.232 5	0.272 7	0.415 7
季节子序列(S_t)	训练集	0.125 5	0.155 1	0.167 0	0.177 2	0.199 1	0.159 8
	测试集	0.110 2	0.094 0	0.090 3	0.098 0	0.064 4	0.093 1
不规则子序列(I_t)	训练集	0.127 6	0.078 8	0.108 2	0.071 7	0.100 4	0.090 4
	测试集	0.135 8	0.106 8	0.119 0	0.084 8	0.114 0	0.095 4

表3 模型预测结果评价指标值

Table 3 Evaluation index values of model prediction results

预测模型	评价指标	北京	天津	郑州	济南	石家庄	太原
ARIMA	MAE	9.254 7	10.390 6	11.227 2	11.154 5	14.700 2	10.526 5
	RMSE	11.968 8	14.295 0	15.628 9	14.949 4	21.446 0	15.823 2
	IA	0.701 1	0.696 4	0.918 0	0.814 0	0.799 1	0.778 8
LSTM	MAE	18.870 0	21.731 8	19.970 0	19.334 7	22.473 2	16.727 8
	RMSE	24.848 0	27.079 7	28.068 7	25.482 9	26.463 2	20.420 3
	IA	0.428 1	0.507 2	0.743 9	0.711 4	0.654 3	0.583 7
SVM	MAE	13.232 4	11.278 6	15.161 3	15.055 4	16.975 1	11.344 0
	RMSE	15.845 0	14.135 7	19.452 9	17.533 6	22.157 3	15.741 2
	IA	0.505 3	0.685 8	0.829 8	0.720 4	0.760 3	0.780 7
Hybrid-X12	MAE	7.522 5	7.227 0	8.039 8	7.557 2	8.383 6	7.597 8
	RMSE	10.390 1	10.441 8	13.840 4	11.427 0	13.717 2	11.411 7
	IA	0.817 6	0.856 9	0.917 7	0.899 6	0.931 0	0.898 6

RMSE分别为7.522 5和10.390 1,相比于ARIMA、LSTM和SVM模型降低了18.72%和13.19%、60.14%和58.19%、43.15%和34.43%。ARIMA、LSTM和SVM模型的IA分别为0.701 1、0.428 1和0.505 3,Hybrid-X12提高到了0.817 6。特别的是郑州站点ARIMA模型的IA为0.918 0,略高于Hybrid-X12模型的0.917 7,但是Hybrid-X12模型的MAE和RMSE指标仍然优于ARIMA。

图5为各模型预测结果时序图,由图可以看出以上模型的预测结果比较接近观测值,均可获得较好的预测精度。四个预测模型的预测结果比较,Hybrid-X12模型的预测值最接近真值,尤其是对于极值点,混合模型能实现较精准地跟踪。ARIMA模型在建模过程中既使用了历史数据又融合了误差信息,在基准模型中表现出最佳的预测性能,然而由于ARIMA模型本质上为线性结构,因此在建模过程中无法捕捉数据的非线性特征。Hybrid-X12模型则针对线性特征明显的趋势子序列使用了ARIMA,而季节子序列和不规则子序列则分别使用LSTM和SVM模型拟合,可以挖掘季节子序列历史数据之间的相关性和不规则子序列的高维非线性特征,因此混合模型表现出优于单一模型的泛化能力,预测精度最高。LSTM模型预测结果存在比较明显地高估,特别是北京和天津站点尤为明显,LSTM模型虽然具有记忆不定时间长度的数值的优势,但是更适合应用于大数据问题,因此在仿真实验中模型泛化能力低于更加适用于小样本问题的SVM模型。图6显示了六个研究站点

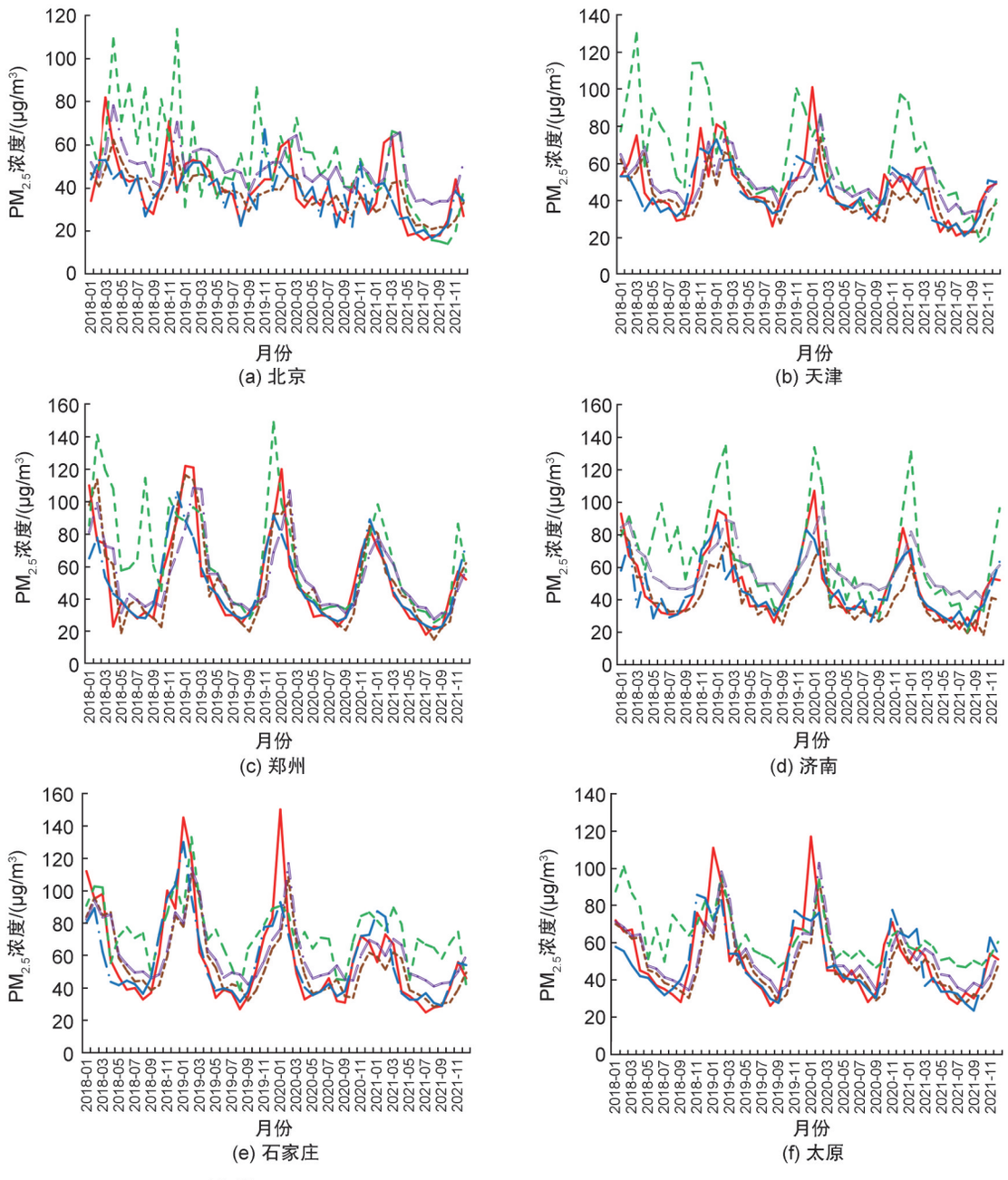


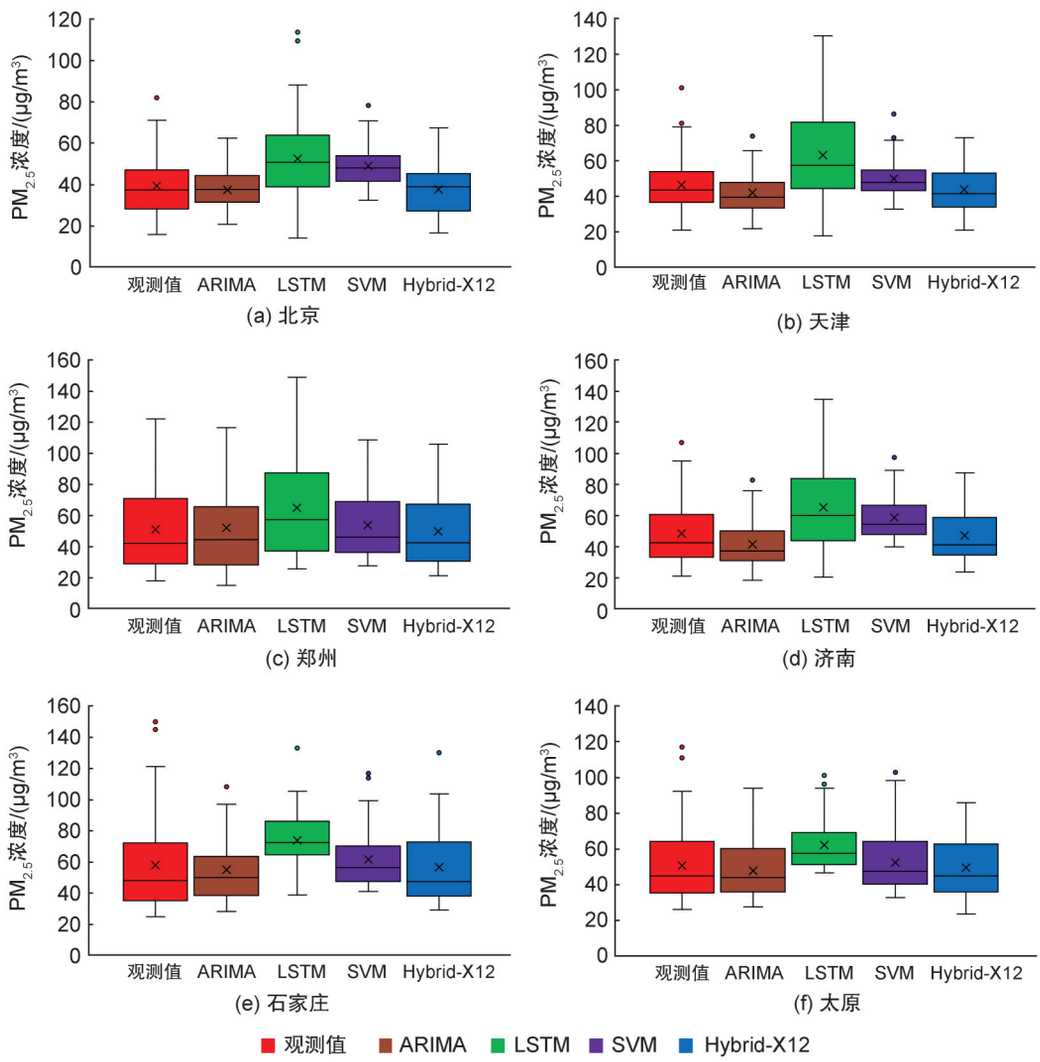
图5 $PM_{2.5}$浓度预测结果时序图

Fig. 5 Time series plots of $PM_{2.5}$ concentration prediction results

预测结果的箱型图。总体来看,Hybrid-X12模型预测值与观测值分布情况最接近,虽然北京站点,ARIMA模型的预测结果的中位数相较于Hybrid-X12模型更接近观测值,但是Hybrid-X12模型的最大值、最小值和四分位数等统计量更准确,并且针对多组实验,Hybrid-X12模型表现出更强的适应性和稳定性,这得益于混合模型充分挖掘了时序数据内在的季节趋势信息,并融合多种学习算法,互相补充,成功捕捉时序数据内在运动规律。

4 结论

$PM_{2.5}$浓度时序具有季节性和趋势性特征,不论是传统线性模型还是非线性智能模型都无法在建模过程中体现时序内在的季节趋势性特征。本文基于季节趋势分解算法建立混合预测模型,挖掘时序运动规律信息,体现各单一模型优势。仿真实验结果对比分析表明,本文提



注:·:异常值;×:均值。

图6 $PM_{2.5}$浓度预测结果箱型图

Fig. 6 Box plots of $PM_{2.5}$ concentration prediction results

出的混合模型具有较好的泛化能力,预测精度优于单一模型ARIMA、LSTM和SVM。本文虽然以$PM_{2.5}$浓度为研究数据构建混合预测模型,无疑该方法也适用于其他具有季节趋势特征的时序预测任务。

参考文献:

[1] JI X, YAO Y X, LONG X L. What Causes $PM_{2.5}$ Pollution? Cross-economy Empirical Analysis from Socioeconomic Perspective[J]. *Energy Policy*, 2018, **119**: 458−472. DOI: 10.1016/J.ENPOL.2018.04.040.

[2] ZAINI N, EAN L W, AHMED A N, *et al*. $PM_{2.5}$ Forecasting for an Urban Area Based on Deep Learning and Decomposition Method[J]. *Sci Rep*, 2022, **12**(1): 17565. DOI: 10.1038/s41598-022-21769-1.

[3] XU G Y, REN X D, XIONG K N, *et al*. Analysis of the Driving Factors of $PM_{2.5}$ Concentration in the Air: A Case Study of the Yangtze River Delta, China[J]. *Ecol Indic*, 2020, **110**: 105889. DOI: 10.1016/j.ecolind.2019.105889.

[4] ZHANG Y, SHEN L Y, SHUAI C Y, *et al*. How is the Environmental Efficiency in the Process of Dramatic Economic Development in the Chinese Cities?[J]. *Ecol Indic*, 2019, **98**: 349−362. DOI: 10.1016/j.ecolind.2018.11.006.

[5] 蒋锋, 乔雅倩. 基于样本熵和优化极限学习机的$PM_{2.5}$浓度预测[J]. 统计与决策, 2021, **37**(3): 166−171. DOI: 10.13546/j.cnki.tjyjc.2021.03.036.
JIANG F, QIAO Y Q. $PM_{2.5}$ Concentration Prediction Based on Sample Entropy and Improved Extreme Learning Machine[J]. *Stat Decis*, 2021, **37**(3): 166−171. DOI: 10.13546/j.cnki.tjyjc.2021.03.036.

[6] PEREZ P, MENARES C, RAMÍREZ C. $PM_{2.5}$ Forecast-

ing in Coyhaique, the Most Polluted City in the Americas [J]. *Urban Clim*, 2020, **32**: 100608. DOI: 10.1016/j.uclim.2020.100608.

[7] YANG X Y, ZHANG Z R. An Attention-based Domain Spatial-temporal Meta-learning (ADST-ML) Approach for $PM_{2.5}$ Concentration Dynamics Prediction[J]. *Urban Clim*, 2023, **47**: 101363. DOI: 10.1016/j.uclim.2022.101363.

[8] ZHANG Y, ZHANG X, WANG L T, *et al*. Application of WRF/Chem over East Asia: Part I. Model Evaluation and Intercomparison with MM5/CMAQ[J]. *Atmos Environ*, 2016, **124**: 285−300. DOI: 10.1016/j.atmosenv.2015.07.022.

[9] ZHAO L X, LI Z Y, QU L L. Forecasting of Beijing $PM_{2.5}$ with a Hybrid ARIMA Model Based on Integrated AIC and Improved GS Fixed-order Methods and Seasonal Decomposition[J]. *Heliyon*, 2022, **8**(12): e12239. DOI: 10.1016/j.heliyon.2022.e12239.

[10] 梁泽, 王玥瑶, 岳远紊, 等. 耦合遗传算法与RBF神经网络的$PM_{2.5}$浓度预测模型[J]. 中国环境科学, 2020, **40**(2): 523−529. DOI: 10.19674/j.cnki.issn1000-6923.2020.0107.
LIANG Z, WANG Y Y, YUE Y W, *et al*. A Coupling Model of Genetic Algorithm and RBF Neural Network for the Prediction of $PM_{2.5}$ Concentration[J]. *China Environ Sci*, 2020, **40**(2): 523−529. DOI: 10.19674/j.cnki.issn1000-6923.2020.0107.

[11] MAHAJAN S, LIU H M, TSAI T C, *et al*. Improving the Accuracy and Efficiency of $PM_{2.5}$ Forecast Service Using Cluster-based Hybrid Neural Network Model[J]. *IEEE Access*, 2018, **6**: 19193−19204. DOI: 10.1109/ACCESS.2018.2820164.

[12] ZHOU Y L, CHANG F J, CHANG L C, *et al*. Multi-output Support Vector Machine for Regional Multi-step-ahead $PM_{2.5}$ Forecasting[J]. *Sci Total Environ*, 2019, **651** (Pt 1): 230−240. DOI: 10.1016/j.scitotenv.2018.09.111.

[13] ZHANG P, MA W J, WEN F, *et al*. Estimating $PM_{2.5}$ Concentration Using the Machine Learning GA-SVM Method to Improve the Land Use Regression Model in Shaanxi, China[J]. *Ecotoxicol Environ Saf*, 2021, **225**: 112772. DOI: 10.1016/j.ecoenv.2021.112772.

[14] CHANG-HOI H, PARK I, OH H R, *et al*. Development of a $PM_{2.5}$ Prediction Model Using a Recurrent Neural Network Algorithm for the Seoul Metropolitan Area, Republic of Korea[J]. *Atmos Environ*, 2021, **245**: 118021. DOI: 10.1016/j.atmosenv.2020.118021.

[15] LI T Y, HUA M, WU X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter ($PM_{2.5}$) [J]. *IEEE Access*, 1809, **8**: 26933−26940. DOI: 10.1109/ACCESS.2020.2971348.

[16] 中国国务院. 关于印发大气污染防治行动计划的通知[EB/OL]. (2013−09)[2023−10−01]. https://www.gov.cn/zwgk/2013−09/12/content_2486773.htm. 北京:中国国务院. 2013.

[17] 刁莉, 王宁. 基于X12-LSTM模型的保费收入预测研究[J]. 计算机科学, 2020, **47**(S1): 512−516. DOI: 10.11896/jsjkx.191100077.
DIAO L, WANG N. Research on Forecast of Premium Income Based on X12-LSTM Model[J]. *Comput Sci*, 2020, **47**(S1): 512−516. DOI: 10.11896/jsjkx.191100077.

[18] 叶晓龙, 罗瑞, 刘金培, 等. 基于X11-WT-LSTM的物流货运量多尺度组合预测研究[J]. 武汉理工大学学报(信息与管理工程版), 2022, **44**(2): 263−269. DOI: 10.3963/j.issn.2095-3852.2022.02.015.
YE X L, LUO R, LIU J P, *et al*. Multi-scale Combined Forecast of Logistics Freight Volume Based on X11-WT-LSTM[J]. *J Wuhan Univ Technol Inf Manag Eng*, 2022, **44** (2): 263−269. DOI: 10.3963/j.issn.2095-3852.2022.02.015.

[19] ALADAĞ E. Forecasting of Particulate Matter with a Hybrid ARIMA Model Based on Wavelet Transformation and Seasonal Adjustment[J]. *Urban Clim*, 2021, **39**: 100930. DOI: 10.1016/j.uclim.2021.100930.

[20] BOX G E P, JENKINS G M, REINSEL C. Time Series Analysis: Forecasting and Control (Third ed) [J]. *Oakland, California, Holden-Day*, 1976, 2013. DOI: 10.1002/9781118619193.

[21] ZHOU C, CHEN X Y. Predicting China's Energy Consumption: Combining Machine Learning with Three-layer Decomposition Approach[J]. *Energy Rep*, 2021, **7**: 5086−5099. DOI: 10.1016/J.EGYR.2021.08.103.

[22] WANG J Q, DU Y, WANG J. LSTM Based Long-term Energy Consumption Prediction with Periodicity [J]. *Energy*, 2020, **197**: 117197. DOI: 10.1016/j.energy.2020.117197.

[23] WANG Q, KONG W, ZHONG J, *et al*. A Hybrid SVM and Kernel Function-based Sparse Representation Classification for Automated Epilepsy Detection in EEG Signals[J]. *Neurocomputing*, 2023, **562**: 126874. DOI: 10.1016/j.neucom.2023.126874.

[24] 李建新, 刘小生, 刘静, 等. 基于MRMR-HK-SVM模型的$PM_{2.5}$浓度预测[J]. 中国环境科学, 2019, **39**(6): 2304−2310. DOI:10.19674/j.cnki.issn1000-6923.2019.0274.
LIU X S, LIU J, *et al*. Prediction of $PM_{2.5}$ Concentration Based on MRMR-HK-SVM Model[J]. *China Environ Sci*, 2019, **39**(6): 2304−2310. DOI:10.19674/j.cnki.issn1000-6923.2019.0274.