doi:10. 16055/j. issn. 1672-058X. 2022. 0002. 009

基于 LSTM-HFTS-EC的 $PM_{2.5}$区间多尺度组合预测研究

罗 瑞[1], 刘金培[1], 陈华友[2], 陶志富[2]
(1. 安徽大学 商学院,合肥 230601; 2. 安徽大学 数学科学学院,合肥 230601)

摘 要:针对 $PM_{2.5}$ 传统点值预测会损失浓度值的波动信息,进而无法充分表示和估计其波动和变化的区间范围,提出了一种基于长短期记忆模型(LSTM)-混合模糊时间序列(HFTS)-误差修正(EC)的 $PM_{2.5}$ 区间多尺度组合预测方法;在结合深度学习和区间多尺度分解方法的基础上,进一步考虑预测误差中隐含的有效信息,建立区间时间序列组合预测模型;该模型能够从随机性较大的时间序列中提取复杂数据特征,解决传统预测方法存在的滞后性以及对误差信息利用不充分等问题;最后,通过实证分析说明该方法适用于具有较大波动的 $PM_{2.5}$ 区间预测,与已有方法相比具有较高的精确度和良好的适用性。

关键词:区间组合预测;$PM_{2.5}$;长短期记忆神经网络;误差修正
中图分类号:O212 **文献标志码**:A **文章编号**:1672-058X(2022)02-0059-09

0 引 言

$PM_{2.5}$ 指大气中空气动力学当量直径小于等于 2.5 μm 的颗粒物,它长时间悬浮于空气中,会侵蚀人体免疫力,引发心血管和呼吸道等疾病[1]。$PM_{2.5}$ 的浓度值是连续变化的,每日最低值和最高值差异大,具有一定的时间跨度和区间关联度。因此,$PM_{2.5}$ 的日均值时间序列无法有效反映其浓度值的真实变化,而以 $PM_{2.5}$ 日最高和最低浓度值分别作为上下限的区间型数据则包含了更多的真实信息,不但可以有效反映其日变化趋势和范围,并且具有更高的稳定性和更强的泛化能力[2]。因此,$PM_{2.5}$ 区间时间序列预测具有更重要的理论和现实意义。

$PM_{2.5}$ 浓度值预测方法主要分为 3 类:数值模拟法[3]、统计预测法[4-6]和深度学习方法[7-10]。数值模拟法在气象学原理的基础上,通过数学方程模拟 $PM_{2.5}$ 的扩散、转化以及消散的过程[2]。此类模型中的参数存在一定的不确定性,导致其预测结果也会存在一定的偏差。统计预测法则是通过回归或机器学习等建立 $PM_{2.5}$ 与影响因素之间的关系[4],实现对 $PM_{2.5}$ 的预测,主要包括多元线性回归(MLR)[5]、支持向量机(SVM)[6]、人工神经网络(ANN)[4]等预测方法。统计类方法虽然能够对 $PM_{2.5}$ 的变化规律以及相关影响因素之间的潜在关系进行拟合,但需要从大量的样本数据中提取时间序列特征,对于波动程度

收稿日期:2021-03-29;**修回日期**:2021-04-27.
基金项目:国家自然科学基金(72071001, 71871001, 71901001);教育部人文社会科学规划项目(20YJAZH066);安徽省自然科学基金项目(2008085MG226, 2008085QG333);安徽省高校人文社会科学重点研究项目(SK2019A0013).
作者简介:罗瑞(1996—),女,湖北京山人,硕士研究生,从事预测与决策研究.
通讯作者:刘金培(1984—),男,山东滨州人,教授,博士,从事预测与决策研究. Email: liujinpei2012@163.com.

较大的时间序列,存在拟合效果不稳定的缺点。

深度学习方法通过学习时间序列数据的内在规律和表现层次,能更好地拟合时间序列,提高预测精度[7-10],主要包括循环神经网络(RNN)、卷积神经网络(CNN)和长短期记忆模型(LSTM)等方法。其中,CNN 的优势在于可以学习和有效提取数据的空间特征;RNN 则存在梯度消失的问题,不适用于长期时间序列的预测;LSTM 是对 RNN 的改进,它解决了 RNN 存在的问题,能够有效提取数据的时间尺度特征。因此,与 RNN 和 CNN 相比,LSTM 更适用于时间序列的预测[7]。Ong 等[8]提出基于 RNN 的深度学习预测框架,对城市的 $PM_{2.5}$ 浓度值进行预测。Wu 等[9]提出了基于 LSTM 的 $PM_{2.5}$ 预测方法,对武汉市 $PM_{2.5}$ 浓度值进行预测。曲悦等[10]分别利用 BP 神经网络、CNN 与 LSTM 对 $PM_{2.5}$ 等空气污染物进行预测。上述基于深度学习的预测模型均能较好提取单一尺度的数据特征,相对于数值模拟法和统计预测法具有更好的预测效果。但是,对于多尺度的复杂时间序列,存在有效信息提取不完全问题。最新研究表明,针对非线性、非平稳性、波动性强的时间序列,先对其进行多尺度分解,使得各子序列具有更好的波动规律性,再对分解后的各层序列分别进行预测分析,可有效提高预测精度[11-12]。为了简化复杂数据,使各子序列平稳化和规律化,从而有效减少预测误差,本文将深度学习与多尺度分解相结合,先选取区间分解方法将 $PM_{2.5}$ 区间序列分解为不同波动频率的子序列,进而采用深度学习方法对高频波动的子序列进行预测。

综上所述,现有研究存在以下 3 个方面的问题:已有研究主要关注 $PM_{2.5}$ 的日均点值时间序列预测,而针对 $PM_{2.5}$ 区间预测的研究较少;如何结合深度学习模型,建立复杂区间时间序列的多尺度分解预测新方法,仍然需要进一步探讨;已有区间时间序列预测方法大多仅关注提高原始序列的预测性能,而没有充分利用预测误差序列中隐含的有效信息。

针对以上问题,提出一种新的基于 LSTM-HFTS-EC 的 $PM_{2.5}$ 区间多尺度组合预测新方法。首先,提出区间时间序列经验模态分解(IEMD)方法,将 $PM_{2.5}$ 区间时间序列依次分解为区间趋势序列、低频波动序列和高频波动序列;然后,分别利用 Holt-Winters 模型、混合模糊时间序列模型(HFTS)和 LSTM 模型对区间趋势序列、低频波动序列和高频波动序列进行预测,并将预测结果集成为 $PM_{2.5}$ 的区间预测值;为了进一步提高区间预测的精确度,再利用 LSTM 模型对 $PM_{2.5}$ 区间预测值进行误差修正,即得到 $PM_{2.5}$ 的最终区间预测结果。最后,将本文的预测方法进行实证预测分析,通过对比来检验本文所提出的组合预测方法的准确性和适用性。

1 $PM_{2.5}$ 区间时间序列

定义 1[2] 记 $\tilde{x}=[x_l,x_u]=\{x \mid x_l \leqslant x \leqslant x_u\}$, x_l, $x_u \in R$,则称 $\tilde{x}$ 为一个区间数。这里,x_l 和 x_u 分别为区间数 $\tilde{x}$ 的下界和上界。此外,区间数 $\tilde{x}=[x_l,x_u]$ 也可以表示成中心和半径的形式,$\tilde{x}=\{c_{\tilde{x}},r_{\tilde{x}}\}$,其中,$c_{\tilde{x}}=(x_l+x_u)/2$ 为区间数 $\tilde{x}$ 的中心,$r_{\tilde{x}}=(x_u-x_l)/2$ 为区间数 $\tilde{x}$ 的半径。在此基础上,令 $\tilde{x}(t)=[x_l(t),x_u(t)]=\{c_{\tilde{x}}(t),r_{\tilde{x}}(t)\}$, $t=1,2,\cdots,n$, 则称 $\{\tilde{x}(t)\}$ 为区间时间序列。

$PM_{2.5}$ 浓度值的日度区间数如图 1 所示,区间下界为每日 $PM_{2.5}$ 的最低浓度值,区间上界为每日 $PM_{2.5}$ 能达到的最高浓度值,区间半径则代表了 $PM_{2.5}$ 浓度值的日变化范围。

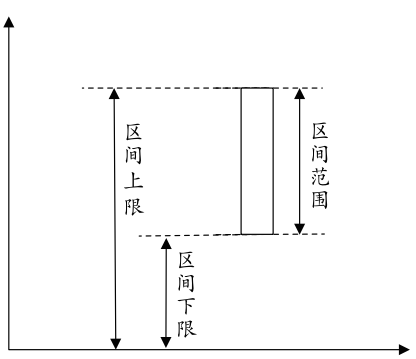


图 1 $PM_{2.5}$ 浓度值日度区间数

Fig. 1 Interval number of daily range of $PM_{2.5}$ concentration

以合肥市为例,合肥市 2018-07-08 至 2018-07-16 共 9 d 的 $PM_{2.5}$ 浓度的区间序列如图 2 所示,可见 $PM_{2.5}$ 浓度值处于连续变化中,日度区间序列的上下界差异大,具有较大的变化范围和较强的波动性。可见,传统的 $PM_{2.5}$ 日均值时间序列预测的代表性弱,无法充分反映其波动规律。

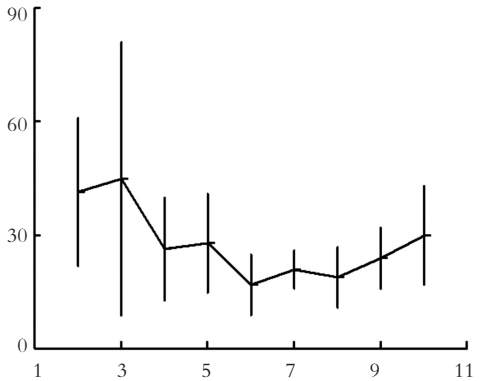


图 2　合肥市 $PM_{2.5}$ 日度区间序列

Fig. 2　Daily interval sequence of $PM_{2.5}$ in Hefei

2　方法与原理

2.1　IEMD

在经验模态分解的基础上,提出 IEMD 方法。具体步骤如下:

步骤 1　针对区间时间序列 $\tilde{x}(t)=\{c_{\tilde{x}}(t),r_{\tilde{x}}(t)\},t=1,2,\cdots,n$。确定中心 $c_{\tilde{x}}(t)$ 和半径 $r_{\tilde{x}}(t)$ 的所有极值点,并用 3 次样条插值的方法拟合得到中心时间序列的上下包络线 $e_{c\max}(t)$ 和 $e_{c\min}(t)$ 以及半径时间序列的上下包络线 $e_{r\max}(t)$ 和 $e_{r\min}(t)$。

步骤 2　分别计算中心和半径时序上下包络线的均值 $m_c(t)$ 和 $m_r(t)$,其中 $m_c(t)=(e_{c\max}(t)+e_{c\min}(t))/2$,$m_r(t)=(e_{r\max}(t)+e_{r\min}(t))/2$。再分别计算中心和半径时序与其平均值之间的差值,记为 $d_c(t)$ 和 $d_r(t)$,这里 $d_c(t)=c(t)-m_c(t)$,$d_r(t)=r(t)-m_r(t)$。

步骤 3　分别判断序列 $d_c(t)$ 和 $d_r(t)$ 是否满足本征模态函数(IMF)的条件[13]。若满足,则记为 1 个 IMF,记 $f_{cm}(t)=d_c(t)$,$f_{rn}(t)=d_r(t)$,$m=1,2,\cdots,M$,$n=1,2,\cdots,N$。将剩余项 $r_c(t)=c(t)-f_{cm}(t)$ 和 $r_r(t)=r(t)-f_{rn}(t)$ 分别作为新的 $c(t)$ 和 $r(t)$。若不满足,则记 $c(t)=d_c(t)$,$r(t)=d_r(t)$ 重复步骤 1~步骤 2 的过程。

步骤 4　重复步骤 1~步骤 3,直到无法再从 $c(t)$ 和 $r(t)$ 中分解出新的 IMF 为止。此时,则 $c(t)$ 和 $r(t)$ 分别分解为多个 IMF 和一个趋势项,即

$$c(t)=\sum_{m=1}^{M}f_{cm}(t)+r_c(t)$$

$$r(t)=\sum_{n=1}^{N}f_{rn}(t)+r_r(t)$$

步骤 5　分别对 $c(t)$ 和 $r(t)$ 的 IMF 进行重构。令 $S_{cu}=\sum_{m=1}^{u}f_{cm}$,$S_{rv}=\sum_{n=1}^{v}f_{rn}$,根据 t 检验判断 S_{cu} 和 S_{rv} 显著不为零时所对应的 u 和 v 的取值。将中心序列分解出的 f_{c1} 至 $f_{c,u-1}$ 合并为高频序列,剩余的 f_{cm} 合并为低频序列,分别记为 $h_c(t)$ 和 $l_c(t)$。同理,可得到半径序列的高频和低频序列,分别记为 $h_r(t)$ 和 $l_r(t)$。

步骤 6　计算得到区间趋势序列 $\tilde{r}(t)=[r_l(t),r_u(t)]=[r_c(t)-r_r(t),r_c(t)+r_r(t)]$,低频波动序列 $l_l(t)=l_c(t)-l_r(t)$ 与 $l_u(t)=l_c(t)+l_r(t)$,高频波动序列 $h_l(t)=h_c(t)-h_r(t)$ 与 $h_u(t)=h_c(t)+h_r(t)$。此时,原始区间时间序列被分解为区间趋势序列、低频和高频波动序列。

本文提出的 IEMD 分解方法能够将区间时间序列分解为区间趋势序列、低频波动序列和高频波动序列,它不仅能够充分提取区间时间序列中所包含的不同尺度的特征信息,而且避免了序列的过度分解,可以进一步简化预测过程的复杂性。

2.2　单项预测方法

2.2.1　LSTM

LSTM 为 RNN 的一种改进,成功解决了 RNN 存在的梯度爆炸和梯度消失问题[14]。通过设置门限控制信息的取舍,解决了长期依赖问题,实现了神经网络的遗忘和记忆功能。LSTM 有 3 个门限,分别为遗忘门(Forget Gate)、输入门(Input Gate)和输出门(Output Gate)。LSTM 的算法结构如图 3 所示。

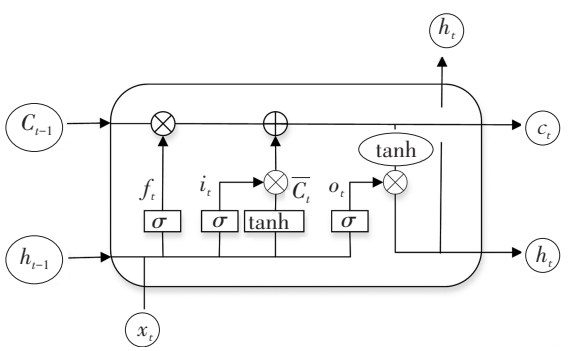


图 3　长短期记忆神经网络的算法结构图

Fig. 3　Algorithm structure diagram of long-short-term memory neural network

在图 3 中,方框内上方的水平线即为单元状态,控制信息传递给下一时刻。LSTM 的前馈计算过程分为 3 步。

第一步决定历史信息是否可以通过单元状态,即

遗忘掉不重要的历史信息,这一步由遗忘门来控制,上一时刻的输出信息 h_{t-1} 和当前时刻的输入信息 x_t 经过 sigmoid 激活函数 σ 得到函数值 $f_t \in [0,1]$,决定历史信息 C_{t-1} 通过单元状态的程度。

$$f_t=\sigma(W_f \cdot [h_{t-1},x_t]+b_f)$$

第二步产生需要更新的信息,即决定当前时刻的输入信息流入单元状态的程度[14]。这一步分为两部分,第1部分由输入门控制[15],根据式(1)产生一个输入信息可流入单元状态的程度值 i_t;第2部分由 tanh 函数决定,在上一时刻的输出信息 h_{t-1} 和当前时刻的输入信息 x_t 的基础上,根据式(2)得到新的候选值 $\bar{C}_t$。进而,由式(3)得到单元状态中需要更新的新的信息 C_t。

$$i_t=\sigma(W_i \cdot [h_{t-1},x_t]+b_i) \tag{1}$$

$$\bar{C}_t=\tanh(W_c \cdot [h_{t-1},x_t]+b_c) \tag{2}$$

$$C_t=f_t \cdot C_{t-1}+i_t \cdot \bar{C}_t \tag{3}$$

第三步决定单元状态中有多少信息需要在当前时刻输出,这一步由输出门决定。根据式(4)得到输出门的值 o_t,由式(5)计算 LSTM 当前时刻的输出值 h_t。其中,$\boldsymbol{W}$、$\boldsymbol{b}$ 分别表示各"门限"的权重矩阵和偏置向量。

$$o_t=\sigma(W_o \cdot [h_{t-1},x_t]+b_o) \tag{4}$$

$$h_t=o_t * \tanh(C_t) \tag{5}$$

2.2.2 HFTS 预测

HFTS 预测模型是传统模糊时间序列(FTS)预测[16]的改进。模糊集、模糊时间序列和模糊关系的定义如下:

定义 2[17] 令 U 为给定论域,将论域划分为 n 个子区间,则 $U=\{u_1,u_2,\cdots,u_n\}$,定义 A 为论域 U 上的模糊集,并记 $A=f_A(u_1)/u_1+f_A(u_2)/u_2+\cdots+f_A(u_n)/u_n$。其中,$f_A(\cdot)$是定义在模糊集 A 上的隶属函数,$f_A(\cdot):U\to[0,1]$;$f_A(u_i)$ 表示 u_i 在模糊集 A 上的隶属度,$i=1,2,\cdots,n$。

定义 3[17] 令 R 中一子集 $Y(t),(t=1,2,\cdots)$ 为给定论域,$f_i(t)$ 为定义在其上的模糊集($i=1,2,\cdots$),且 $F(t)=\{f_1(t),f_2(t),\cdots\}$,则称 $F(t)$ 为定义在 $Y(t)$ 上的模糊时间序列。

定义 4[17] 假设 $F(t)$ 由 $F(t-1)$ 所引起,即 $F(t-1)\to F(t)$,此关系可表示为 $F(t)=F(t-1)\circ R(t,t-1)$,则称 $F(t)$ 为一阶模糊,$R(t,t-1)$ 为 $F(t-1)$ 与 $F(t)$ 之间的模糊关系。其中,符号"$\circ$"表示合成运算。

HFTS 预测方法的算法步骤如下:

步骤 1 利用模糊 C 均值聚类(FCM)[20]将论域 U 划分为 n 个区间,并且确定训练数据属于各区间的隶属度。

步骤 2 结合训练数据属于各模糊集的隶属度,将时间序列转化为 FTS。

步骤 3 利用 BP 神经网络定义模糊关系,将前 m 期数据属于各模糊集的隶属度作为 BP 神经网络的输入值,将后一期数据的实际值作为 BP 神经网络的输出值。

步骤 4 确定 BP 神经网络隐含层的神经元个数、激活函数以及相关参数,构建网络结构。

步骤 5 利用训练集数据对网络进行训练。

步骤 6 利用训练好的网络进行预测。

HFTS 能有效地解决 FTS 存在的 3 个问题:在模糊化阶段,HFTS 利用系统的 FCM 方法将数据集模糊化,以此得到各数据更加客观的隶属度,从而解决了 FTS 中隶属度存在极大主观性的问题;同时,在建立模糊关系阶段,HFTS 通过 BP 神经网络消除了模糊关系的结构性选择问题,并且避免了复杂的模糊关系矩阵计算[18];另外,HFTS 通过将时间序列的实际值作为目标值,利用 BP 神经网络进行预测,避免了去模糊化阶段中可能出现的预测误差,从而提高了预测性能。

2.2.3 Holt-Winters 模型

Holt-Winters 模型适用于对含有趋势变动、季节变动和周期变动的时间序列进行预测。本文应用的乘法 Holt-Winters 预测模型[19],如下:

$$\begin{cases} l_t=\alpha(x_t/s_{t-m})+(1-\alpha)(l_{t-1}+b_{t-1}) \\ b_t=\beta(l_t-l_{t-1})+(1-\beta)b_{t-1} \\ s_t=\gamma(x_t/(l_{t-1}+b_{t-1}))+(1-\gamma)s_{t-m} \\ \hat{x}_{t+h}=(l_t+hb_t)s_{t-m+h} \end{cases} \tag{6}$$

其中,l_t 为 t 时刻的周期项,表示去除季节变化影响后的时间序列的平均数,m 为时间周期;b_t 为 t 时刻趋势项,表示时间序列趋势的线性变动值;s_t 为 t 时刻季节项,表示季节因子的指数平滑平均数;α、β、γ 为平滑系数,取值区间为 $[0,1]$;$\hat{x}_{t+h}$ 为预测值,h 表示需要预测期数[14]。

3　预测模型

针对 $PM_{2.5}$ 区间时间序列数据非线性、非平稳性和波动幅度较大等特点,提出一种新的基于 LSTM-HFTS-EC 的 $PM_{2.5}$ 区间多尺度组合预测方法,结构框架如图 4 所示,具体步骤如下:

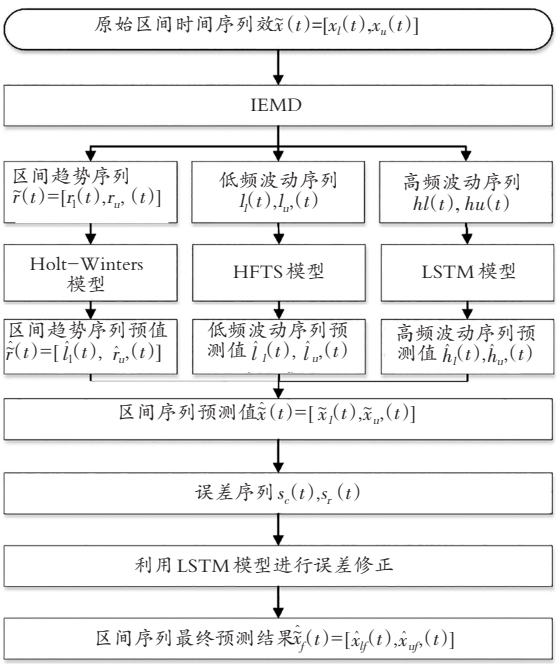


图 4　组合预测模型结构框架

Fig. 4　Combinatorial prediction model framework

步骤 1　区间时间序列分解。基于 IEMD 方法,将 $PM_{2.5}$ 区间时间序列 $\tilde{x}(t)=[x_l(t),x_u(t)]$ 依次分解为区间趋势序列 $\tilde{r}(t)=[r_l(t),r_u(t)]$、低频波动序列 $l_l(t)$ 和 $l_u(t)$ 以及高频波动序列 $h_l(t)$ 和 $h_u(t)$。

步骤 2　组合预测。利用 Holt-Winters 模型、HFTS 模型和 LSTM 模型对区间趋势序列、低频波动序列和高频波动序列分别进行预测,并得到预测值 $\hat{\tilde{r}}(t)=[\hat{r}_l(t),\hat{r}_u(t)]$、$\hat{l}_l(t)$、$\hat{l}_u(t)$、$\hat{h}_l(t)$ 和 $\hat{h}_u(t)$。然后将其预测结果进行集成,得到 $PM_{2.5}$ 区间浓度预测值 $\hat{\tilde{x}}(t)=[\hat{x}_l(t),\hat{x}_u(t)]$。其中,$\hat{x}_l(t)=\hat{r}_l(t)+\hat{l}_l(t)+\hat{h}_l(t)$,$\hat{x}_u(t)=\hat{r}_u(t)+\hat{l}_u(t)+\hat{h}_u(t)$。

步骤 3　计算误差。将 $PM_{2.5}$ 区间预测值 $\hat{\tilde{x}}(t)$ 用中心和半径表示 $\hat{c}(t)$,$\hat{r}(t)$,计算中心预测误差序列 $s_c(t)$ 和半径预测误差序列 $s_r(t)$。其中,$s_c(t)=c(t)-\hat{c}(t)$,$s_r(t)=r(t)-\hat{r}(t)$。

步骤 4　误差修正。利用 LSTM 分别对中心和半径预测误差序列 $s_c(t)$ 和 $s_r(t)$ 进行预测,得到中心和半径的误差预测值 $\hat{s}_c(t)$ 和 $\hat{s}_r(t)$。根据中心和半径的误差预测值 $\hat{s}_c(t)$ 和 $\hat{s}_r(t)$,对 $PM_{2.5}$ 区间预测值进行误差修正,得到 $PM_{2.5}$ 区间最终预测结果 $\hat{\tilde{x}}_f(t)=[\hat{x}_{lf}(t),\hat{x}_{uf}(t)]$。其中,$\hat{x}_{lf}(t)=\hat{x}_l(t)+\hat{s}_c(t)-\hat{s}_r(t)$,$\hat{x}_{uf}(t)=\hat{x}_u(t)+\hat{s}_c(t)+\hat{s}_r(t)$。

步骤 5　模型检验。利用区间平均相对误差(IARV)、区间平均绝对误差(IMAE)、区间平均绝对百分比误差(IMAPE)和区间均方根误差(IRMSE)4 种预测误差评价指标对本模型以及其他预测模型进行对比分析,以此检验本文提出模型的预测效果。

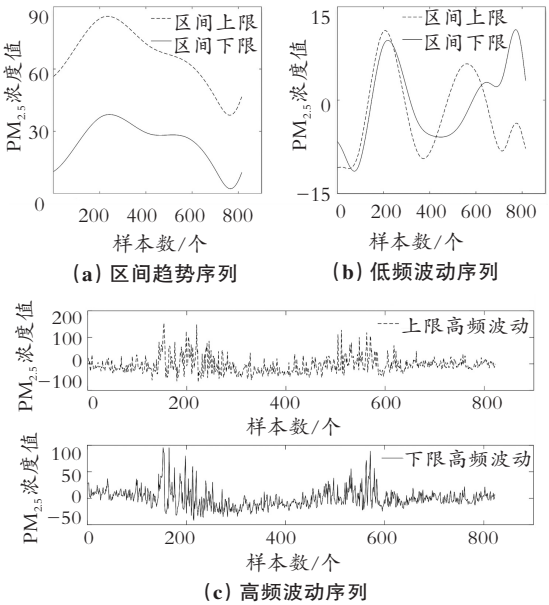
4　实证研究

4.1　数据来源和预处理

选取合肥市 2018-07-01 至 2020-12-30 共计 914 d 的 $PM_{2.5}$ 数据进行预测实证分析,数据来源于安徽省生态环境厅(http://sthjt.ah.gov.cn)。对收集的 $PM_{2.5}$ 实时数据进行预处理,将日实时数据的最高值和最低值分别作为区间的上下限,得到 $PM_{2.5}$ 区间时间序列。其中,选取前 822 个区间数据作为训练集,后 92 个区间数据作为测试集。

4.2　区间数据的多尺度分解

根据提出的组合预测模型流程,利用 IEMD 方法将合肥市 $PM_{2.5}$ 原始区间时间序列依次分解为区间趋势序列、低频波动序列和高频波动序列,结果如图 5 所示。其中,IEMD 是一种自适应分解方法,勿需提前设置分解函数和分解层数,该方法可以自行分解成对应的层数。图 5(a)为分解得到的区间趋势序列,它体现了合肥市 $PM_{2.5}$ 浓度值的整体趋势,可见受新冠疫情等因素影响,合肥市 $PM_{2.5}$ 浓度值在一段时期内处于下降趋势,而随着工业生产的全面复工以及出行的常态化,其浓度值也逐渐上升。图 5(b)为低频波动序列,反映了合肥市 $PM_{2.5}$ 浓度值的短期变化规律,从图 5 中可以看出其浓度值具有明显的季节性和周期性,说明冬季合肥市 $PM_{2.5}$ 浓度值会达到最高峰,空气污染比较严重,在夏季合肥市 $PM_{2.5}$ 浓度值会处于最低水平,空气质量相对较好。图 5(c)表示高频波动序列,它体现了在众多因素影响下合肥市 $PM_{2.5}$ 浓度值的具体波动细节,可以发现合肥市 $PM_{2.5}$ 浓度值的波动幅度较大。



(a) 区间趋势序列 (b) 低频波动序列

(c) 高频波动序列

图 5 $PM_{2.5}$ 原始区间数据分解图

Fig. 5 $PM_{2.5}$ original interval data decomposition

4.3 组合预测

在上一阶段的基础上,首先采用 Holt-Winters 对区间趋势序列进行预测,平滑系数分别设置为 $\alpha=0.3$、$\beta=0.3$ 和 $\gamma=0.4$。然后基于 HFTS 对低频波动序列进行预测,其中,FCM 的类别数 $n=5$,BP 神经网络的输入值 $m=5$,隐含层的神经元个数设置为 16,激活函数为 tansig,迭代次数为 1 000。最后选取 LSTM 对高频波动序列进行单步预测,其中,本文设置 1 个输入层、1 个隐含层和 1 个输出层,隐含层的神经元个数为 128,激活函数设置为 tanh,时间步长设置为 1,批处理大小取值为 50,迭代次数设置为 500。同时,将各单项预测方法的预测结果进行相加集成,得到合肥市 $PM_{2.5}$ 组合预测值如图 6 所示。可以看出,虽然组合预测值整体可以反映 $PM_{2.5}$ 实际值的变化趋势,但是 $PM_{2.5}$ 区间浓度值上限和下限的预测结果与实际值之间在波动细节上仍存在一定的差异。

说明高频波动序列中隐含的一部分复杂数据特征没有被有效利用,因此,预测结果仍然有改进的空间。

4.4 误差修正

基于组合预测的结果,进一步采用误差修正的方法,从预测误差中间接提取隐含的有效信息,进一步提高预测精度。首先,计算得到原始数据与 $PM_{2.5}$ 组合预测结果之间的差值作为误差序列。然后,采用 LSTM 对误差序列进行预测,在相关参数设置不改变的情况下,将前一期数据作为输入值,后一期数据作为输出值,得到误差序列的预测结果。最后,利用误差预测值对组合预测的结果进行修正,将两者相加集成,得到 $PM_{2.5}$ 区间浓度值最终预测结果,如图 6 所示。由此可见,进行误差修正后的预测结果更加接近于实际值,预测效果得到了较大地提升。说明误差修正能够从预测误差序列中进一步提取有效信息,提高模型的预测精度。

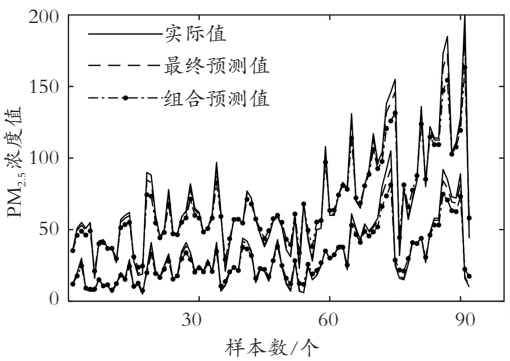


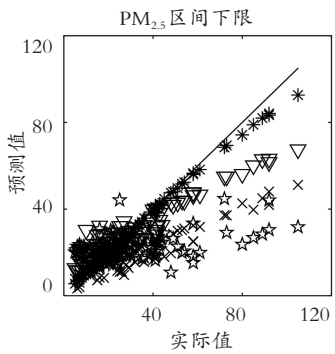
图 6 预测结果对比图

Fig. 6 Comparison of forecast results

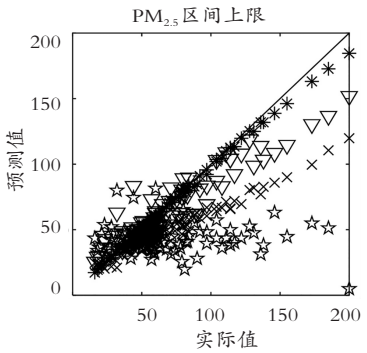
4.5 多种方法预测效果对比分析

为了验证所提出模型的有效性,进行了以下 3 个方面的对比,并用散点图表示。各数据点越趋于对角线表明预测结果越接近实际值,反之,效果越差。对比分析结果如下:

(1) 不同分解方法的对比。在本文方法的基础上,对分解方法进行变换,采用变分模态分解(VMD)、集成经验模态分解(EEMD)、奇异谱分解(SSA)对 $PM_{2.5}$ 原始区间数据进行分解处理,并将其记为 VMD-CF-ECM、EEMD-CF-ECM 和 SSA-CF-ECM。对比结果如图 7 所示(✱、☆、▽、×分别代表 IEMD-CF-ECM、VMD-CF-ECM、EEMD-CF-ECM、SSA-CF-ECM),本文模型(IEMD-CF-ECM)的预测效果明显优于这 3 种模型,由此可见 IEMD 非常适用于分析波动幅度不规律、非线性和非平稳性的区间时间序列,它能根据数据固有的波动尺度特征来进行时间序列分解,具有客观性和自适应性。



(a) 下限预测效果对比

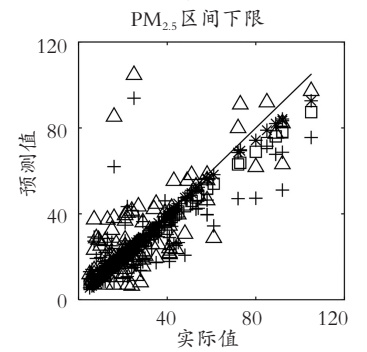


(b) 上限预测效果对比

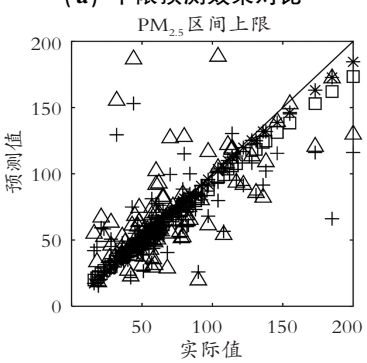
图 7　本模型与采用不同的分解方法比较图

Fig. 7　Comparison of proposed model with the model using different decomposition methods

(2) 不同单项预测方法的对比。在本模型基础上,对分解后的不同频率序列分别采用同一单项方法进行预测,如 LSTM、HFTS 和 Holt-Winters,并依次将其设定为 IEMD-LSTM-ECM、IEMD-HFTS-ECM 和 IEMD-HW-ECM。由图 8 可知(✳、□、+、▽分别代表 IEMD-CF-ECM、IEMD-LSTM-ECM、IEMD-HFTS-ECM、IEMD-HW-ECM),采用单项预测方法的预测精度低于本文提出的组合预测方法,这说明利用 IEMD 将 $PM_{2.5}$ 区间序列分解为不同尺度的子序列,然后根据不同时序的数据特征,选择最合适的模型,利用组合预测方法对其进行预测,能够极大程度地提高预测性能,从而达到最优预测效果。



(a) 下限预测效果对比

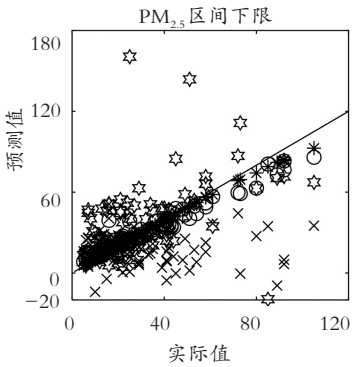


(b) 上限预测效果对比

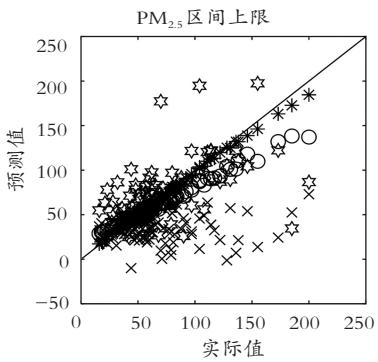
图 8　本模型与采用单一预测方法比较图

Fig. 8　Comparison of proposed model with the model using a single prediction method

(3) 与已有预测模型的横向对比。将本文模型与其他 $PM_{2.5}$ 预测模型进行对比,如 LSTM 模型[10]、ANN 模型[21]和 Holt-Winters 模型[21]。对比结果如图 9 所示(✳、○、☆、×分别代表 IEMD-CF-ECM、LSTM、ANN、HW),可见对比模型的预测效果远低于本文提出模型,由此可见,本模型具有更高的预测精度。



(a) 下限预测效果对比



(b) 上限预测效果对比

图 9　本模型与其他预测方法比较图

Fig. 9　Comparison of the proposed model with the model using other forecasting methods

另外,本文模型与其他比较模型的预测误差评价指标如表 1 所示。可以看出,本模型预测误差的各评价指标都低于其他比较模型,进一步反映了本文提出模型的预测精度要高于其他比较模型,具有较好的预测效果和较强的适应性。

表 1　10 种预测方法预测误差评价指标对比

Table 1　Comparison of prediction error evaluation indexes of ten forecasting methods

预测模型	IARV	IMAE	IMAPE	IRMSE
VMD-CF-ECM	1.482 0	24.788 3	0.522 7	3.640 7
EEMD-CF-ECM	0.204 7	9.840 1	0.272 7	1.408 0
SSA-CF-ECM	0.605 2	17.360 5	0.314 6	2.421 3
IEMD-LSTM-ECM	0.034 0	3.588 3	0.074 5	0.576 9
IEMD-HFTS-ECM	0.600 3	15.546 1	0.418 5	2.361 4
IEMD-HW-ECM	0.702 7	16.416 7	0.484 1	2.540 4
LSTM	0.158 5	7.820 7	0.180 8	1.167 0
ANN	0.995 4	19.595 5	0.598 9	3.154 2
HW	1.939 9	29.579 4	0.599 5	4.245 9
IEMD-CF-ECM	0.008 7	1.644 2	0.033 8	0.292 9

综上所述,本研究具有以下3方面的优势:第一,本文提出的区间时间序列组合预测模型可以对 $PM_{2.5}$ 浓度值变化趋势和范围进行更好预测,克服了传统点值时间序列预测波动信息损失的缺点;第二,本文将深度模型LSTM与区间多尺度分解等方法相结合,提出了一种区间时间序列组合预测框架,能够从大量复杂的时间序列数据中提取关键性的数据特征,克服了传统预测模型存在的滞后性问题;第三,本文通过提取组合预测预测误差提供的有效信息,进行误差修正,使预测精度得到了进一步地提升。

5 结束语

$PM_{2.5}$ 浓度值是一个连续变化、随机性强、波动频率不规律的时间序列,传统的日均值分解预测模型很难准确地获取高频序列中的随机性特征,也无法完全体现 $PM_{2.5}$ 的区间变化规律。因此,本文提出了一种新的基于LSTM-HFTS-EC的 $PM_{2.5}$ 区间多尺度组合预测模型方法,首先利用IEMD将 $PM_{2.5}$ 区间时间序列进行分解,再基于Holt-Winters模型、HFTS和LSTM模型分别对分解出的区间趋势序列、低频波动序列和高频波动序列进行预测,并将预测结果集成为 $PM_{2.5}$ 的区间预测值。进而利用LSTM模型对 $PM_{2.5}$ 区间预测值进行误差修正,得到 $PM_{2.5}$ 的最终区间预测结果。最后通过实证预测分析,说明本文的方法适用于具有较大波动的 $PM_{2.5}$ 区间预测,与已有方法相比具有更高的精确度和良好的适用性。此外,本文的研究也为预测其他具有连续变化和波动范围大特征的实际问题提供了一种新的思路。

参考文献(References):

[1] 翁克瑞,刘淼,刘钱. TPE-XGBOOST与LassoLars组合下 $PM_{2.5}$ 浓度分解集成预测模型研究[J]. 系统工程理论与实践, 2020, 40(3): 748—760.
WENG Ke-rui, LIU Miao, LIU Qian. An integrated prediction model of $PM_{2.5}$ concentration based on TPE-XGBOOST and LassoLars[J]. Systems Engineering - Theory & Practice, 2020, 40(3): 748—760.

[2] 丁勤祥,陶志富,葛璐璐等. 基于L1范数的IOWGA算子的区间组合预测模型[J]. 统计与决策, 2019, 35(22): 20—23.
DING Qin-xiang, TAO Zhi-fu, GE Lu-lu. Interval combination forecasting model of IOWGA operators based on L1 norm[J]. Statistics and Decision, 2019, 35(22): 20—23.

[3] SHANG Z, DENG T, HE J. A novel model for hourly $PM_{2.5}$ concentration prediction based on CART and EELM[J]. Science of the Total Environment, 2019, 651(2): 3043—3052.

[4] VOUKANTSIS D, KARATZAS K, KUKKONEN J. Intercomparison of air quality data using principal component analysis and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki[J]. Science of the Total Environment, 2011, 409(7): 1266—1276.

[5] AMANOLLAHI J, AUSATI S. $PM_{2.5}$ concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran, Iran[J]. Air Quality Atmosphere and Health, 2020, 13(2): 161—171.

[6] WANG P, ZHANG H, QIN Z. A novel hybrid-garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting[J]. Atmospheric Pollution Research, 2017, 8(5): 850—860.

[7] LI X, PENG L. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation[J]. Environmental Pollution. 2017, 231(1): 997—1004.

[8] ONG B T, SUGIURA K, ZETTSU K. Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$[J]. IEEE Access, 2016, 27(6): 1553—1566.

[9] WU X L, WANG Y, HE S Y. $PM_{2.5}/PM_{10}$ ratio prediction based on a long short-term memory neural network in Wuhan, China[J]. Geoscientific Model Development, 2020, 13(3): 1499—1511.

[10] 曲悦,钱旭,宋洪庆. 基于机器学习的北京市 $PM_{2.5}$ 浓度预测模型及模拟分析[J]. 工程科学学报, 2019, 41(3): 401—407.
QU Yue, QIAN Xu, SONG Hong-qing. Machine-learning-based model and simulation analysis of $PM_{2.5}$ concentration prediction in Beijing[J]. Chinese Journal of Engineering, 2019, 41(3): 401—407.

[11] 刘金培,汪漂,黄燕燕. 基于区间时间序列小波多尺度分解的组合预测方法[J]. 统计与决策, 2020, 36(19): 5—9.
LIU Jin-pei, WANG Piao, HUANG Yan-yan. A combined prediction method based on interval time series wavelet multi-scale decomposition[J]. Statistics and Decision, 2020, 36(19): 5—9.

[12] GAN K, SUN S L, WANG S Y. A Secondary-

decomposition-ensemble learning paradigm for forecasting $PM_{2.5}$ concentration[J]. Energy Conversion and Management, 2018, 9(6): 989—999.

[13] 潘和平,张承钊. FEPA-金融时间序列自适应组合预测模型[J]. 中国管理科学, 2018, 26(6): 26—38.
PAN He-ping, ZHANG Cheng-zhao. FEPA: an adaptive integrated prediction model of financial time series[J]. Chinese Journal of Management Science, 2018, 26(6): 26—38.

[14] GREFF K, SRIVASTAVA R K, KOUTNIK J. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222—2232.

[15] SENG D W, ZHANG Q Y, ZHANG X F. Spatiotemporal prediction of air quality based on LSTM neural network [J]. Alexandria Engineering Journal, 2021, 60(2): 2021—2032.

[16] CHEN C, JHONG Y, WU W. Fuzzy time series for real-time flood forecasting[J]. Stochastic Environmental Research and Risk Assessment, 2019, 33(3): 645—656.

[17] CHEN S, CHEN C. Taiex forecasting based on fuzzy time series and fuzzy variation groups[J]. IEEE Transactions on Fuzzy Systems, 2011, 19(1): 1—12.

[18] BALLA-ARABE S, GAO X, WANG B. A fast and robust level set method for image segmentation using fuzzy clustering and lattice Boltzmann method[J]. IEEE Transactions on Cybernetics, 2013, 43(3): 910—920.

[19] BAPTISTA VENTURA L M, PINTO F D O, SOARES L M. Forecast of daily $PM_{2.5}$ concentrations applying artificial neural networks and holt-winters models[J]. Air Quality, Atmosphere & Health, 2019, 12(3): 317—325.

Research on $PM_{2.5}$ Interval Multi-scale Combination Prediction Based on LSTM-HFTS-EC

LUO Rui[1], LIU Jin-pei[1], CHEN Hua-you[2], TAO Zhi-fu[2]

(1. Business School, Anhui University, Hefei 230601, China; 2. School of Mathematical Science, Anhui University, Hefei 230601, China)

Abstract: Traditional $PM_{2.5}$ point value prediction would lose the fluctuation information of concentration value, and thus could not adequately represent and estimate the range of its fluctuation and change. A multi-scale combination prediction method for $PM_{2.5}$ range is proposed based on long short-term memory (LSTM), hybrid fuzzy time series (HFTS) and error correction (EC). Based on deep learning and interval multi-scale decomposition method, the combined prediction model of interval time series is established by further considering the effective information hidden in the prediction error. This model can extract complex data features from time series with large randomness, and solve the problems of lag existing in traditional forecasting methods and insufficient use of error information. Finally, the empirical analysis shows that this method is suitable for the prediction of $PM_{2.5}$ range with large fluctuation, and has higher accuracy and good applicability by comparing with the existing methods.

Key words: interval combination forecast; $PM_{2.5}$; long short-term memory; error correction

责任编辑:罗姗姗

引用本文/Cite this paper:

罗瑞,刘金培,陈华友,等. 基于 LSTM-HFTS-EC 的 $PM_{2.5}$ 区间多尺度组合预测研究[J]. 重庆工商大学学报(自然科学版),2022,39(2):59—67.

LUO Rui, LIU Jin-pei, CHEN Hua-you, et al. Research on $PM_{2.5}$ interval multi-scale combination prediction based on LSTM-HFTS-EC[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(2):59—67.