DOI:10.3969/j.issn.1003-5060.2025.06.017

基于函数型期望分位数回归森林模型的AQI预测

陈慧琪, 凌能祥
(合肥工业大学 数学学院,安徽 合肥 230601)

摘 要:文章将函数型数据分析和期望分位数回归森林(expectile regression forest,ERF)模型相结合,分析了合肥市2015—2022年空气质量,并利用函数型ERF模型对空气质量指数(air quality index,AQI)进行预测。研究结果表明,大部分真实值均落在预测区间中,期望分位数回归森林模型表现出较好的预测结果,体现出函数型数据与随机森林模型相结合的优势。

关键词:函数型数据;期望分位数回归;随机森林;非参数回归;空气质量指数(AQI)
中图分类号:O212.7 **文献标志码**:A **文章编号**:1003-5060(2025)06-0823-05

Prediction of air quality index based on functional expectile regression forest model

CHEN Huiqi, LING Nengxiang
(School of Mathematics, Hefei University of Technology, Hefei 230601, China)

Abstract: In this paper, the air quality in Hefei City from 2015 to 2022 was analyzed by combining functional data analysis(FDA) and expectile regression forest(ERF) model, and the air quality index (AQI) was predicted based on functional ERF model. It is found that most of the actual values fall within the prediction interval, indicating that the AQI of Hefei City is well predicted by ERF model, which exhibits the advantages of FDA with random forest model.

Key words: functional data; expectile regression; random forest; non-parametric regression; air quality index(AQI)

0 引 言

随着我国经济的快速发展,能源消耗持续增加,大气污染问题日趋严重。大气污染物会对人类、生物和生态系统造成有害影响,保护大气环境刻不容缓。为此,我国建立了空气污染指数(air pollution index,API)、空气质量指数(air quality index,AQI)及各类污染物指标数据的监测发布平台用于评价空气质量。

函数型数据分析(functional data analysis,FDA)[1]的思想是将观测区间内1次观测到的数据视为整体进行分析,这些数据构成了曲线、曲面或者图像;文献[2]对一些传统统计分析方法进行改进,使之适用于函数型数据分析。函数型数据统计推断方法被广泛应用于多个领域,如文献[3]建立了函数型数据的k近邻估计并将其应用于预测$PM_{2.5}$质量浓度;文献[4]基于函数型数据分析对京津冀空气污染问题进行了研究;文献[5]建立了空间函数型期望回归的非参数估计模型,用于对中国东北地区的空气质量指标进行评价;文献[6]利用函数型数据分析和期望分位数回归方法对北京$PM_{2.5}$质量浓度进行探究;文献[7]将函数型数据与随机森林结合,并应用到心电图数据上进行分类预测。

若出现极端污染的天气,则空气质量指数波动较大。文献[8]提出了期望分位数(expectile),该方法克服了分位数回归的局限性,它不仅涉及了损失的概率,还考虑了损失的大小,对于异常值更加敏感,能够更加充分地揭示解释变量对响应变量分布的特征。环境空气质量指数受多方面因

收稿日期:2023-03-10;**修回日期**:2023-05-04
基金项目:国家自然科学基金资助项目(72071068)
作者简介:陈慧琪(1996—),女,安徽芜湖人,合肥工业大学硕士生;
凌能祥(1964—),男,安徽合肥人,合肥工业大学教授,博士生导师,通信作者,E-mail:hfut.lnx@163.com.

素的影响,因此简单的参数回归预测模型很难考虑到所有要素。近年来,机器学习算法的预测模型受到广泛关注,其预测效果相比于其他模型体现出一定的优越性。文献[9]提出了期望分位数回归梯度增强树模型(expectile regression-boost,ER-Boost);文献[10]将神经网络与期望分位数回归相结合,提出了期望回归神经网络模型;文献[11]采用深度残差网络学习框架,提出了一种期望分位数回归神经网络;文献[12]将期望分位数回归模型同支持向量机相结合;文献[13]将期望分位数和随机森林相结合建立了期望分位数回归森林(expectile regression forest, ERF)模型。

本文将函数型数据分析的思想应用到合肥市空气质量的分析中,建立期望分位数回归森林模型对合肥空气质量进行预测,并对合肥 2015—2022 年空气质量进行月度、年度的变化特征分析。

1 期望分位数

期望分位数回归作为分位数回归的推广,它考虑如下最优化问题:

$$\mu_\tau(X)=\arg\min_{\psi\in\mathbf{R}}E[\rho_\tau(Y-\psi)\mid X=x] \quad (1)$$

其中:$\rho_\tau(Y,\psi)=\tau(Y-\psi)^2I_{\{(Y-\psi)>0\}}+(1-\tau)(Y-\psi)^2I_{\{(Y-\psi)\leqslant 0\}}$,为非对称平方损失函数(least asymmetrically weighted squares, LAWS);I_A 为事件 A 的示性函数;Y 为一个随机变量;τ 为不对称参数且 $0\leqslant\tau\leqslant 1$,此处最小化式(1)所得到的 ψ 即为 τ 期望分位数,当 $\tau=0.5$ 时,期望分位数即为条件期望。虽然期望分位数回归的定义相比于分位数回归更加复杂,但其优势在于损失函数处处可导,且受分布尾部概率和尾部具体取值的双重影响,因此期望分位数回归对于分布尾部数据更加敏感。

2 函数型期望分位数回归森林模型

2.1 函数型数据分析

函数型数据的基本思想是将函数数据作为一个变量。在实际问题中,样本都是以离散的形式进行观测和记录的,假设有 n 个观测样本,每个样本有 N 对数据序列,即第 i 个样本包含数据序列 $(t_1,x_{i1}),(t_2,x_{i2}),\cdots,(t_N,x_{iN})$,首先将这些离散点对拟合成函数形式 $x_i(t)$,本文利用基函数方法来拟合数据序列,并根据曲线的各种特点采用不同的基函数系统。方法如下:

$$x_i(t)\approx\sum_{s=1}^{S}c_{is}\phi_s(t) \quad (2)$$

其中:$x_i(t)$ 为第 i 个样本曲线,$i=1,2,\cdots,N$;$\phi_s(t)$ 为第 s 个基函数;c_{is} 为对应的系数,$s=1,2,\cdots,S$。

2.2 函数型决策树

函数型决策树[7]将决策树扩展到函数型数据分析的框架中,利用基表示的协同作用获得新的特征来训练函数分类器。基于式(2)中的基函数系统可获得如下特征矩阵:

$$\boldsymbol{C}=\begin{pmatrix}c_{11}&\cdots&c_{1S}\\ \vdots& &\vdots\\ c_{N1}&\cdots&c_{NS}\end{pmatrix} \quad (3)$$

其中,矩阵中的任一元素 c_{is} 表示第 i 条曲线的第 s 个基函数的系数。将以上矩阵作为新的特征来训练决策树,由此进行分类。

2.3 期望分位数回归森林模型

期望分位数回归森林模型[13]将随机森林模型应用到期望分位数中最终获得条件期望分位数。其基本思想是:首先利用 Bagging 算法同时建立多个决策树;然后计算决策树中每个叶子节点的条件期望分位数值;最后取多个决策树的条件期望分位数的平均值作为模型预测的条件期望分位数值。此处,将函数型决策树模型和期望分位数回归森林模型相结合,考虑函数型期望分位数回归森林模型,期望分位数回归森林模型计算步骤[13]如下。

1) 使用 Bagging 算法,从总数据组中随机选取 n 组构成数据集 $D=\{C_i,Y_i\}_{i=1}^{n}$ 建立函数型决策树,其中 C_i 指式(3)中第 i 行的所有元素。重复 Z 次共建立 Z 个函数型决策树,其中第 z 个函数型决策树记为 $T(\gamma^z)$,$\gamma^z=\{L_m^z,Y_m^z\}_{m=1}^{M}$ 是第 z 个函数型决策树 $T(\gamma^z)$ 的参数,L_m^z 为第 m 个叶子节点,Y_m^z 为第 m 个叶子节点中的响应变量的值,记为:

$$Y_m^z=\{y\mid C_i\in L_m^z\},\quad m=1,2,\cdots,M;z=1,2,\cdots,Z \quad (4)$$

2) 遍历 Z 个函数型决策树并计算每个函数型决策树的条件期望分位数。即若 $w\in\mathbf{R}^p$ 是一组测试数据,它属于决策树 $T(\gamma^z)$ 第 m 个叶子节点,可记作 $w\in L_m^z$,则计算第 z 个叶子节点中 Y_m^z 的条件期望分位数,具体算法[9]如下:① Y_m^z 中样本量为 S,将 Y_m^z 按照升序排列成一组新的序列 $\{Y(s)\}_{s=1}^{S}$,且令 $y(0)=-\infty$,$y(s+1)=+\infty$;② 由于 LAWS 函数 ρ_τ 为严格凸且连续可微函

数，对于 $k=0,1,\cdots,S$，则有：

$$\frac{\partial}{\partial v}\sum_{s=1}^{S}\rho_\tau(y_{(s)},v)\Big|_{v=v_k}=\frac{\partial}{\partial \nu}\Big\{\Big[\sum_{s=1}^{S}(1-\tau)I_{(s\leqslant k)}+\tau I_{(s\geqslant k+1)}\Big](y_{(s)}-\nu)^2\Big\}\Big|_{\nu=v_k}=0 \tag{5}$$

由式(5)计算可得：

$$v_k=\frac{\sum_{s=1}^{S}(1-\tau)y_{(s)}I_{(s\leqslant k)}+\tau y_{(s)}I_{(s\geqslant k+1)}}{\sum_{s=1}^{S}(1-\tau)I_{(s\leqslant k)}+\tau I_{(s\geqslant k+1)}} \tag{6}$$

③ 对于 $k=0,1,\cdots,S$，有唯一的 k^* 满足 $y_{(k^*)}\leqslant \hat{v}_{k^*}\leqslant y_{(k^*+1)}$；④ Y_m^z 的期望分位数为 $\hat{\mu}_\tau(y|w\in L_m^z)=\hat{v}_{k^*}$。

3) 将 Z 个函数型决策树获得的条件期望分位数进行平均，通过期望分位数回归森林模型获得的条件期望分位数计算公式为：

$$\hat{\mu}_\tau(y\mid z)=\frac{1}{Z}\sum_{z=1}^{Z}\hat{\mu}_\tau(y\mid w\in L_m^z) \tag{7}$$

在期望分位数回归森林模型中，本文存在 3 个超参数需要进行反复调整：① 决策树的个数(n_{tree})[14]在一定情况下，随机森林模型的误差会随着决策树数量的增加而减少，当决策树的个数达到临界值时，预测误差不再减少计算时间反而会随着决策树个数的增加而增加；② 指定节点中用于二叉树的变量个数(m_{try})[15]设置为解释变量总数的 1/3 时，预测结果更好；③ 每个决策树中叶子节点的最少个数(n_{size})[16]当节点值在 5 左右时，随机森林模型具有更好的预测性能。为了解决以上 3 个超参数问题，应用如下 out-of-bag(记为集合 o)公式[13]：

$$\Omega_{\text{MSE}}^{o}(n_{\text{tree}},m_{\text{try}},n_{\text{size}})=\frac{1}{n_o}\sum_{i\in o}[y_i-\hat{\mu}_\tau(y\mid C_i)]^2 \tag{8}$$

当式(8)最小时，所选取 3 个超参数即为最优。

3　数据分析

3.1　数据来源与说明

目前我国的空气质量标准来自《环境空气质量标准》《环境空气质量指数(AQI)技术规定(试行)》，污染物包括 SO_2、NO_2、PM_{10}、$PM_{2.5}$、CO、O_3 6 种。AQI 是定量描述空气质量状况的无量纲指数，它将多种污染物的质量浓度经过一定的转化变成单一数值形式来反映空气质量情况。本文选取合肥市每日 AQI 为研究对象，数据来自 www.aqistudy.cn。

3.2　空气质量变化特征

合肥市 2015—2022 年各年每日 AQI 的走势折线图如图 1 所示。

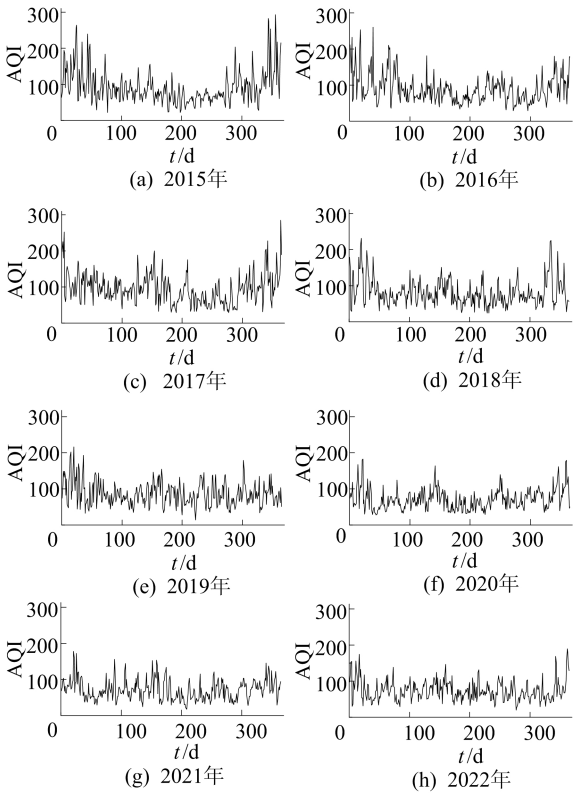


图 1　2015—2022 年合肥市各年每日 AQI 折线图

2015—2022 年合肥市每日 AQI 箱线图如图 2所示。从图 2 可以看出，合肥市的 AQI 在 40～300 之间，空气质量呈现逐年好转的趋势，2015—2022 年出现极端空气质量的天数逐年减少，且极端天气的 AQI 有所下降，特别自 2018 年起开展污染防治攻坚战以来，效果显著。另外由图 1 可知，AQI 较高的时期大多在冬季，冬季处于取暖的高峰期，各类能源消耗较多导致空气质量下降，而夏季的空气质量普遍较好。

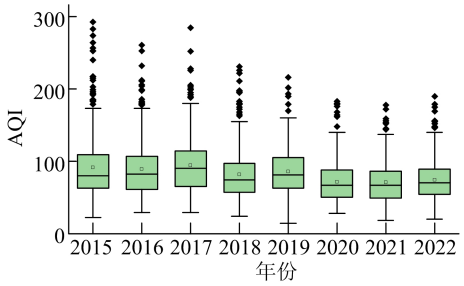
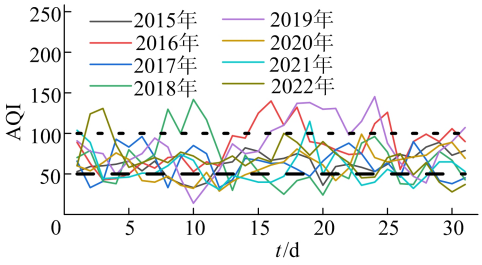


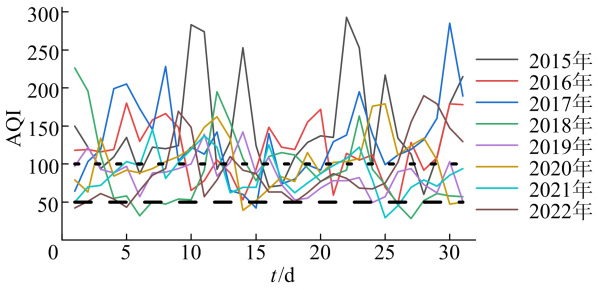
图 2　2015—2022 年合肥市每日 AQI 箱线图

2015—2022 年 8 月和 12 月当月每日 AQI 的

变化情况如图 3 所示。其中:黑色虚线为空气质量指数(AQI=100)良的达标线;黑色划线为空气质量指数(AQI=50)优的达标线。



(a) 8月份每日AQI



(b) 12月份每日AQI

图 3 2015—2022 年 8 月和 12 月当月每日 AQI 的变化

由图 3 可知,夏季空气质量大多数达到了良好的标准,而冬季多出现极端的空气质量不达标天气。另外,随着时间的推移以及人们对于生活品质的注重,可以看出近几年合肥市的空气质量大多数保持在良或轻微污染状态。

3.3 空气质量指数预测

3.3.1 模型的构建

由于 2018 年国家颁布了《打赢蓝天保卫战三年行动计划》,空气质量问题得到全面关注,空气质量有所好转,本文考虑 2018—2022 年合肥市每日 AQI,将其分为训练集和测试集 2 个部分。为了方便计算将 AQI 数据取对数,参考文献[6],并根据季节各选取 2 组训练集和测试集,见表 1 所列,以避免预测结果存在偶然性。以 4 周的每日 AQI 为 1 组,即以 28 d 为 1 个周期,并对未来 28 d的 AQI 进行预测。

表 1 训练集和测试集划分

训练集	测试集
2018-01-01—2021-01-23	2022-01-24—2022-02-20
2018-01-01—2022-07-10	2022-07-11—2022-08-07

本文参考文献[17]中的划分方法,将合肥市每日 AQI 数据视为一组时间序列 $\{Z_j,1\leqslant j\leqslant N\}$,假设 $N=n\alpha$,其中:$n\in\mathbf{N}^*$;$\alpha=28$。构建如下样本:

$$X_i=\{Z_{(i-1)\alpha+t},t=1,2,\cdots,28\},\quad Y_i=Z_{i\alpha+s}\tag{9}$$

其中:$i=1,2,\cdots,n$;$s=1,2,\cdots,28$。

预测问题则变为给定一个函数型变量 X 对响应变量 Y 的预测问题。以 $\{(x_i(t),y_i(s)),i=1,2,\cdots,n-1\}$ 为训练集建立模型,然后以 $(x_n(t),y_n(s))$ 为预测集,分别考虑 τ 为 2.5%、50.0%、97.5% 点处的期望分位数预测结果。

针对式(9)建立的训练集,根据式(2)采用 B 样条基进行拟合,获得基函数的系数组成式(3),根据式(3)建立函数型决策树,并根据 2.3 节所建立的函数型期望分位数回归森林模型对 AQI 进行预测。

3.3.2 预测结果

本文分别应用期望分位数回归树(expectile regression tree,ERT)模型和期望分位数回归森林(ERF)模型进行预测,并采用均方根误差(root mean square error,RMSE)评价估计量的精度结果。均方差误差计算公式为:

$$\Omega_{\mathrm{RMSE}}(\tau)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_\tau(y_i\mid C_i))^2}\tag{10}$$

其中:y_i 为真实值;$\hat{\mu}_\tau(y_i\mid C_i)$ 为 τ 分位点处的预测值。

期望分位数回归树模型和期望分位数随机森林模型预测的均方根误差见表 2 所列,可以看出,期望分位数回归森林模型的精确度更高。

表 2 不同分位点处 ERF 和 ERT 模型的 RMSE

τ/%	ERF 模型	ERT 模型
2.5	0.525 7	0.535 9
50.0	0.365 9	0.484 0
97.5	0.716 8	0.796 5

为了更清晰地展现预测效果,本文分析了 2022-01-24—2022-02-20 和 2022-07-11—2022-08-07 两组 AQI 预测结果,分别如图 4、图 5 所示。

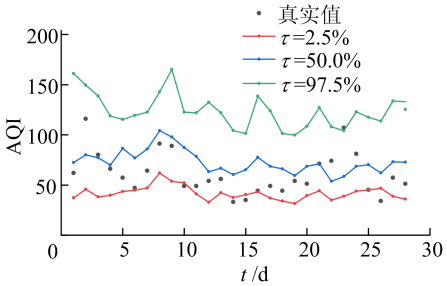


图 4 2022-01-24—2022-02-20 的 AQI 预测结果

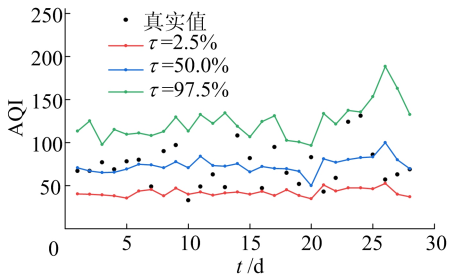


图 5　2022-07-11—2022-08-07 的 AQI 预测结果

从图 5 可以看出，2 个模型均基本捕捉到了 AQI 的波动，而 2.5%和 97.5%分位数预测结果给出了预测值的区间范围，可以看出大多数真实值均落在预测区间内。

4　结　　论

本文结合函数型数据的分析方法，采用 B 样条基对 AQI 数据平滑生成函数变量，结合期望分位数回归森林模型，对合肥市空气质量进行了分析，结果表明：合肥市的空气质量在污染治理下逐年好转，极端污染天气明显减少，同大部分地区一样，合肥市的空气污染呈现冬季污染较严重，夏季空气质量较好的特征；通过建立期望分位数回归树模型和期望分位数回归森林模型对 2022 年部分时段的合肥市每日 AQI 进行预测，经比较发现，真实值大多落在函数型期望分位数回归森林模型所预测出的估计区间中，特别是当出现一些极端值时，期望分位数回归森林模型也表现出良好的预测效果。

［参　考　文　献］

［1］ RAMSAY J，DALZELL C. Some tools for functional data analysis［J］. Journal of the Royal Statistical Society：Series B（Methodological），1991，53（3）：539-561.

［2］ RAMSAY J，SILVEMAN B. Functional data analysis［M］. New York：Springer，1997.

［3］ 程彦茹，凌能祥. 随机缺失函数型数据的 k 近邻估计及其应用［J］. 合肥工业大学学报（自然科学版），2020，43（3）：429-432.

［4］ 梁银双. 基于函数型数据分析的京津冀空气污染问题研究［D］. 北京：首都经济贸易大学，2017.

［5］ RACHDI M，LAKSACI A，Al-KANDARI M N. Expectile regression for spatial functional data analysis（sFDA）［J］. Metrika：International Journal for Theoretical and Applied Statistics，2022，85（5）：627-655.

［6］ 朱佳，冯峥晖，陈正宇. 基于函数型数据分析和广义分位数的 PM2.5 数据探究［J］. 数理统计与管理，2021，40（5）：771-784.

［7］ MATURO F，VERDE R. Pooling random forest and functional data analysis for biomedical signals supervised classification：theory and application to electrocardiogram data［J］. Statistics in Medicine，2022，41（12）：2247-2275.

［8］ NEWEY W，POWELL J L. Asymmetric least squares estimation and testing［J］. Econometrica，1987，55（4）：819-847.

［9］ YANG Y，ZOU H. Nonparametric multiple expectile regression via ER-Boost［J］. Journal of Statistical Computation and Simulation，2015，85（7）：1442-1458.

［10］ JIANG C，JIANG M，XU Q，et al. Expectile regression neural network model with applications［J］. Neurocomputing，2017，247：73-86.

［11］ YIN Y，ZOU H. Expectile regression via deep residual networks［J］. Stat，2021，10（1）：e315.

［12］ PEI H M，LIN Q，YANG L R，et al. A novel semi-supervised support vector machine with asymmetric squared loss［J］. Advances in Data Analysis and Classification，2021，15（1）：159-191.

［13］ CAI C，DONG H T，WANG X Y. Expectile regression forest：a new nonparametric expectile regression model［J］. Expert Systems，2023，40（1）：e13087.

［14］ OSHIRO T M，PEREZ P S，BARANAUSKAS J A. How many trees in a random forest？［C］//Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition. Berlin：Springer-Verlag，2013：154-168.

［15］ MERINSHAUSEN N. Quantile regression forests［J］. Journal of Machine Learning Research，2006，7（6）：983-999.

［16］ LIN Y，JEON Y. Random forests and adaptive nearest neighbors［J］. Journal of the American Statistical Association，2006，101（474）：578-590.

［17］ FERRATY F，VIEU P. Nonparametric functional data analysis：theory and practice［M］. New York：Springer，2006.

（责任编辑　李　凯）