文章编号:1001-5078(2022)01-0023-06　　·激光应用技术·

一种基于随机森林的激光大气传输效应评估方法

谭　哲,龚艳春,杨云涛,冷　坤
(陆军工程大学基础部,江苏 南京 211101)

摘　要:针对现有的激光大气传输效应评估方法(定标律评估和波动光学仿真评估)较难有良好的模型正则性和应用普适性的困难,提出一种基于随机森林的激光大气传输效应评估方法。该方法首先以烟台某地大气环境数据(温度、风速、湍流强度(C_n^2)等)的随机采样和激光发射参数(传输距离、激光功率等)作为输入数据和多层相位屏模型仿真生成的环围功率(PIB)作为输出数据,然后利用随机森林进行训练和预测。结果表明,随机森林较支持向量机能更好的表征输入与输出间的多元回归关系,预测均方根误差优于 0.021 %;传输距离、C_n^2 与 PIB 相关性最强,对 PIB 影响最大。本方法可为机器学习在激光大气传输效应评估的应用提供更加完善的理论依据,具备一定的应用价值。

关键词:随机森林;激光传输;效应评估
中图分类号:O436.1;TN249　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2022.01.004

A simulation study on atmospheric transmission effect evaluation of laser based on random forest

TAN Zhe, GONG Yan-chun, YANG Yun-tao, LENG Kun
(Army Engineering University Foundation Division, Nanjing 211101, China)

Abstract: In view of the difficulties in model regularization and application universality of existing methods (scaling law evaluation and wave optical simulation evaluation) for evaluating laser atmospheric transport effect, a method based on random forest for evaluating laser atmospheric transport effect is proposed. This method firstly takes random sampling of atmospheric environmental data (temperature, wind speed and turbulence intensity(C_n^2), etc.) in Yantai and laser parameters (transmission distance, laser power, etc.) as the input data, and multiple phase screen model simulation of turbine power (PIB) as the output data, then with the use of random forests for training and prediction. The results show that the random forest can better represent the multiple regression relationship between input and output than the support vector, and the prediction root mean square error is less than 0.021 %. Transmission distance and turbulence intensity C_n^2 have the strongest correlation with PIB, and have the greatest influence on PIB. This method can provide a more perfect theoretical basis for the application of machine learning in the evaluation of laser atmospheric transport effect, and has a certain application value.

Keywords: random forest; laser transmission; effect assessment

基金项目:国家自然科学基金项目(No. 11804390)资助。
作者简介:谭　哲(1991 -),男,硕士研究生,主要研究方向为激光大气传输。E-mail:562373857@qq.com
通讯作者:杨云涛(1984 -),男,博士,讲师,主要从事大气传输及激光应用技术方面的研究。E-mail:legend08fda@126.com
收稿日期:2021-02-20

1 引 言

激光经过大气传输后到达靶目标处一定面积内的光功率密度分布是衰减效应、湍流效应和非线性热晕效应共同作用的结果,与激光的初始状态、传输路径上的大气环境密切相关。通过对靶目标处的光功率换算,可以得到环围功率(PIB)、远场光斑半径、衍射极限倍数等光束质量评价因子,以实现对激光大气传输效应的评估。对激光大气传输效应进行评估研究,根据分析方法和采用模型的不同,主要有波动光学模型评估方法、定标律模型评估方法和统计分析模型评估方法[1]。波动光学模型评估方法是基于波动光学方程,通过多层相位屏方法建立的激光大气传输仿真软件;定标律模型评估方法是将不同的激光系统参数输入仿真软件经计算获得了激光传输规律,而后加入特征参数进行拟合,得出定标公式。统计分析模型评估方法的原理是基于外场设备时时测量、数据统计分析和机器学习等技术,研究各类输入参数对传输效应的影响。

近年,基于机器学习方法研究激光大气传输效应已发展为一种趋势,采取的方法有随机森林(random forest,RF)[2]、神经网络(BP)[3]和支持向量机(SVM)[4]。文献[2]基于 RF,以低空海洋环境大气参数为输入特征量对 C_n^2 预测,分析得出气温和气压是预报海洋环境中的 C_n^2 最重要的参数。文献[3]通过连续四天收集三亚地区 C_n^2 及温度、风速、相对湿度三种参数,基于 BP 对 C_n^2 进行估算,整体相关系数达到 0.86。文献[4]基于 SVM 将激光发射参数和大气参数作为输入构建好的模型,通过输出光束质量评价因子,以实现对激光大气传输效应评估。得出支持向量机能很好的拟合输入与输出的多元回归关系。上述研究表明,机器学习方法在该领域是可行的。

本文在机器学习的基础上,采用 RF 算法。RF 基于 2014 年 1 月烟台地区某地实际大气环境数据[5]、激光发射参数以及上述两类数据通过相位屏模型[6]获取仿真数据(包括 PIB、远场光斑半径、衍射极限倍数等光束质量评价因子,本文选取 PIB 作研究对象),对 PIB 进行预测,并将预测结果与相位屏生成的 PIB 比较,检验其拟合能力。同时为了评价模型的准确性和可靠性,分别采用 RF、SVM[4]从均方根误差(E_{RMS})、平均绝对误差(E_{MA})和平均相对误差(E_{MR})等方面进行比较,以实现对激光大气传输效应的评估。

2 RF 算法

2.1 RF 的工作原理

RF 算法是一种基于决策树的集成算法,它利用 Bootstrap 重采样技术,以随机的模式来构建森林,采用 Bagging 算法有放回的从原始训练集取样得到多个训练集,而后用每一个训练集进行训练得到相应的决策树模型[7]来组建立“森林”。决策树通过选择最优特征在树的每个节点不停进行分类,直到达到树成型的停止条件[8]。决策树的分支结点所包含的样本尽可能属于同一类别,即结点的“纯度”越高。信息增益和 Gini 是提高样本“纯度”的最佳方式,信息增益使用“信息熵”表征“纯度”的高低,数据集 D 的信息熵公式为:

$$\mathrm{Ent}(d) = -\sum_{k=1}^{|y|} P_k \log_2 P_k \tag{1}$$

$\mathrm{Ent}(d)$ 的值越小,D 的“纯度”的越高。表示任意类别样本 占数据集 D 的概率假定离散属性 A 有 k 个值,用特征 A 对 D 进行划分,D 会被划分为 k 个部分,此时可用特征 A 分割结点的信息增益用 $\mathrm{Gain}(D,A)$ 表示,公式如下:

$$\mathrm{Cain}(D,A) = \mathrm{Ent}(D) - \sum_{k=1}^{k} \frac{|D^k|}{|D|}\mathrm{Ent}(D^k) \tag{2}$$

数据集 D 的纯度还可用基尼指数来衡量,基尼指数越小,D 的纯度越高。公式如下:

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{k} p_i^2 \tag{3}$$

$$\mathrm{Gini}(D,A) = \sum_{j=1}^{k} \frac{|D^k|}{|D|}\mathrm{Gini}(D^k) \tag{4}$$

由于本文处理的是激光发射参数、大气环境参数和 PIB 的关系,因此采用随机森林回归(random forest for regression,RFR)算法。随机森林回归[9]模型是通过与随机向量 θ 有关的决策树构成的,模型的预测结果是 k 棵决策树的 $\{h(X,\theta_i,i=1,2,\cdots,k)\}$ 均值。

$$f_{\mathrm{RFR}} = \frac{1}{k}\sum_{i=1}^{k} h_i(X) \tag{5}$$

式中,f_{RFR} 表示 *RFR* 模型的结果。

2.2 RFR 模型训练

随机森林回归模型的建立主要分以下四步,如

图 1 所示:①数据准备。将仿真数据中的激光初始半径、初始功率、C_n^2、能见度(V)、传输距离(L)、风速(v)、温度(t)作为自变量,将 PIB 作为因变量输入模型;②数据预处理。样本数据为 100 组 800 个,虽然表中的 8 个属性量纲相差较大,但随机森林泛化能力较好,数据无需做规范化处理。而且还可查看样本数据有无缺失,如有可用随机森林回归进行填补;③参数选取。将数据按一定比例划分为训练集和测试集,采取网格搜索法或学习曲线法选取最优的决策树数量和树的最大深度,实现模型全局最优;④构建模型。设定好参数好利用训练集对模型进行训练,并通过交叉验证法检验模型稳定性。如不满足要求,重新调整参数,继续训练模型。

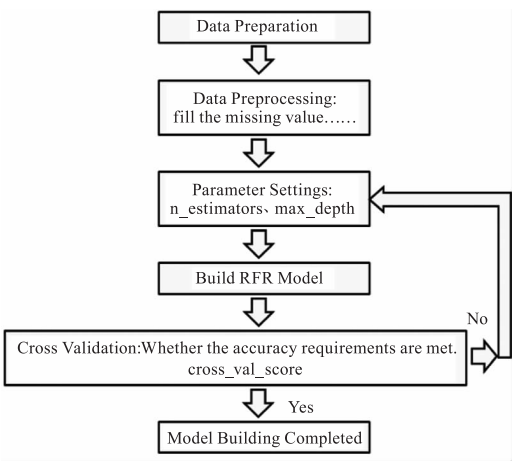

图 1　随机森林建模步骤的结构框图

Fig. 1 Schematic diagram of stochastic forest modeling steps

将模型在测试集上的 E_{RMS}、E_{MA}、E_{MR}作为评价模型预测精度的指标。变量值与预测值的相关系数(R)作为模型模拟结果与实际值的吻合程度的衡量指标。E_{RMS}、E_{MA}和 E_{MR}的值越小,R 的绝对值越接近 1,表明模型的预测效果越好。定义式如下:

$$E_{\mathrm{RMS}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_i-f(x_i)\right]^2} \tag{6}$$

$$E_{\mathrm{MA}}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-f(x_i)\right| \tag{7}$$

$$E_{\mathrm{MR}}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|y_i-f(x_i)\right|}{y_i}\right) \tag{8}$$

$$R=\pm\sqrt{\frac{\sum_{i=1}^{n}\left(f(x_i)-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}} \tag{9}$$

式中,y_i、$f(x_i)$、$\bar{y}$ 分别表示第 i 组数据的仿真结果、预测结果、样本仿真结果的平均值,n 为样本数目。交叉验证是用来观察模型的稳定性的另一种方法,它将原始样本打乱并重复利用,尽可能利用有限的样本资源减少预测偏差,并同时考虑了训练误差和泛化误差,此方法可以检验划分的训练集和测试集的比例是否恰当。本文将数据划分为 10 份,依次使用其中一份作为测试集,其他 9 份作为训练集,多次计算模型的精确性来评估模型的平均准确程度。

3　仿真与结果

3.1　仿真数据生成

训练 RFR 以烟台某地大气环境数据的随机采样和激光发射参数作为输入,以多层相位屏模型仿真计算的 PIB 作为输出,如图 2 所示。先采用 2014 年 1 月烟台地区的大气环境数据(包括 t、v、C_n^2)和激光发射参数,结合相位屏模型生成激光光斑数据,从而计算得 PIB。再将大气环境数据、激光发射参数和相位屏生成的 PIB 作为原始数据集输入 RFR 模型。RFR 模型按照 17∶3 划分训练集与测试集,训练集用于训练学习模型,测试集用于验证模型的预测性能。

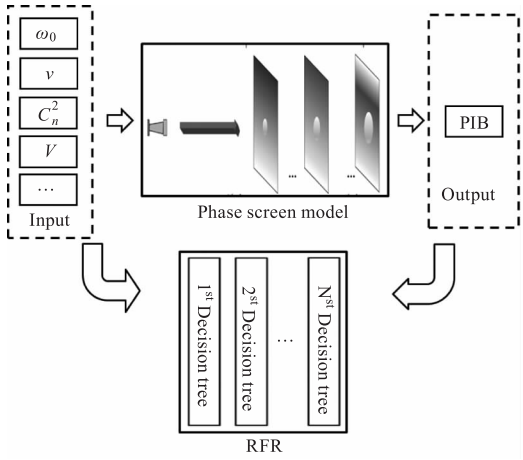

图 2　RFR 中数据来源

Fig. 2 Source of data for training RFR

参数具体设置为:ω_0 的范围为 0.15 m;P_0 的范围为 0.2 ~ 1 kW;C_n^2 的范围为 1×10^{-14} ~ 1×10^{-17} $\mathrm{m}^{-1.5}$;V 的取值范围为 1 ~ 3 km;L 的范围为 100 ~ 1000 m;v 的范围为 4 ~ 8 m/s;t 的范围是 276 ~ 286 K;C_n^2,t,v 的变化趋势如图 3 所示。由图可知,C_n^2 的密度直方图为孤岛型;t 的密度直方图为双峰型和孤岛型;v 的密度直方图为双峰型。三者偏离正常态,符合大气环境随机性的特性。

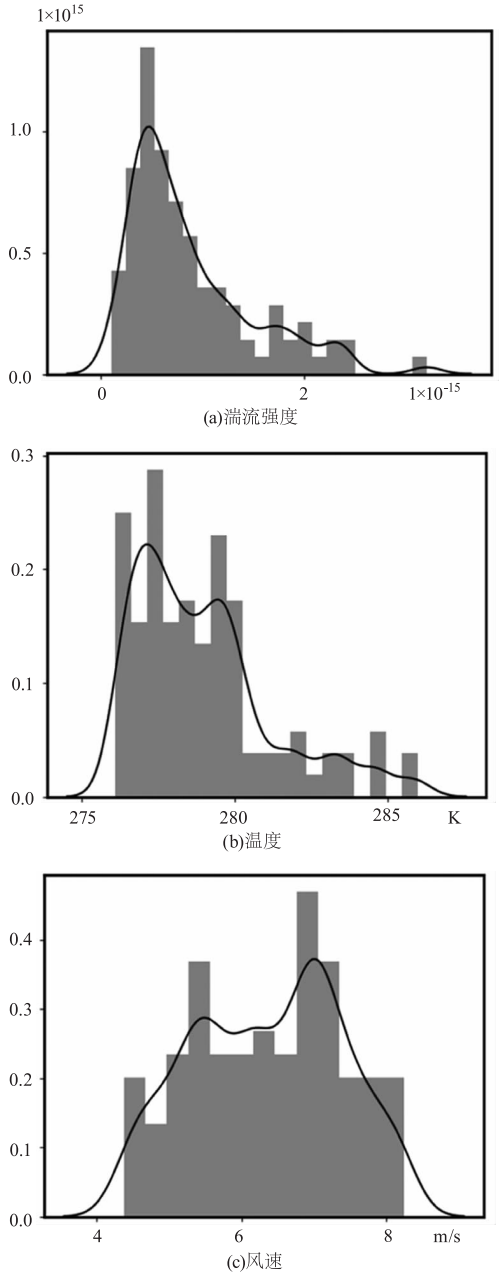

图 3 湍流强度、温度、风速直方密度曲线图

Fig. 3 Linear density curve of turbulence intensity, temperature and wind speed

3.2 RFR 参数筛选结果分析

随机森林模型在构建过程中,两个关键参数决策树数量(n_estimators)和最大的树深度(max_depth)的选取可采用学习曲线法和网格搜索法。具体过程先使 n_estimato – rs 的值使参数局部最优,在前述基础上再求 max_depth 的值使参数局部最优,通过两次参数调整实现全局最优。光束质量评价因子 PIB 预测模型的最优参数学习曲线图法如图 4 所示。

网格搜索法可选用调参算法(如 Grid Search)对 n_estimators、max_depth 一并搜索,得出 n_estimators:79,max_depth:17,模型在测试集上相关系数为 0.878。对比上述两种方法,随机森林模型采用学习曲线法可以得到最优参数 n_estimators:88,max_depth:12,模型相关系数为 0.906。

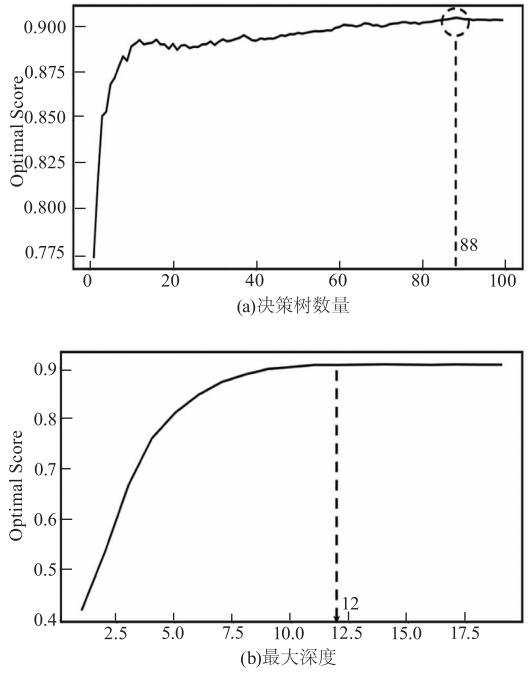

图 4 决策树数量和最大深度的学习曲线图

Fig. 4 Learning curve of decision tree number and maximum depth

3.3 RFR 与 SVM 预测结果对比

图 5(a)、5(b)给出了 RFR 的 PIB 的预测结果散点图,图中直线表示预测值与仿真值相等的情况,数据点越接近图中直线,相关系数越大,代表自变量对因变量的解释程度越高。从图 5 中结果可知,PIB 模型训练集与测试集的数据点基本位于直线上,说明模型的拟合程度较高,预测误差较小。测试集平均绝对误差为 0.015 %,平均相对误差为 0.017 %,均方根误差为 0.021 %。通过 10 次交叉验证的检验分析技术,求得模型均方根误差均值为 0.035 %。因为 0.021 % 小于 0.035 %,在平均均方根误差范围内,即模型构建符合要求。

图 5(c)、5(d)中为 SVM 的 PIB 预测结果散点图,图中模型在预测集上预测精度为 0.901,测试集平均绝对误差为 0.61 %,平均相对误差为 0.53 %,均方根误差为 0.44 %。从图 5(a)、5(b)和图 5(c)、5(d)结果分析中可知,RFR 比 SVM 的相关系数大,且 RFR 的 E_{RMS}、E_{MA}、E_{MR} 更低。表明 RFR 比 SVM 的预测精度高,RFR 较 SVM 能更好地拟合自变量和因变量间的回归关系。

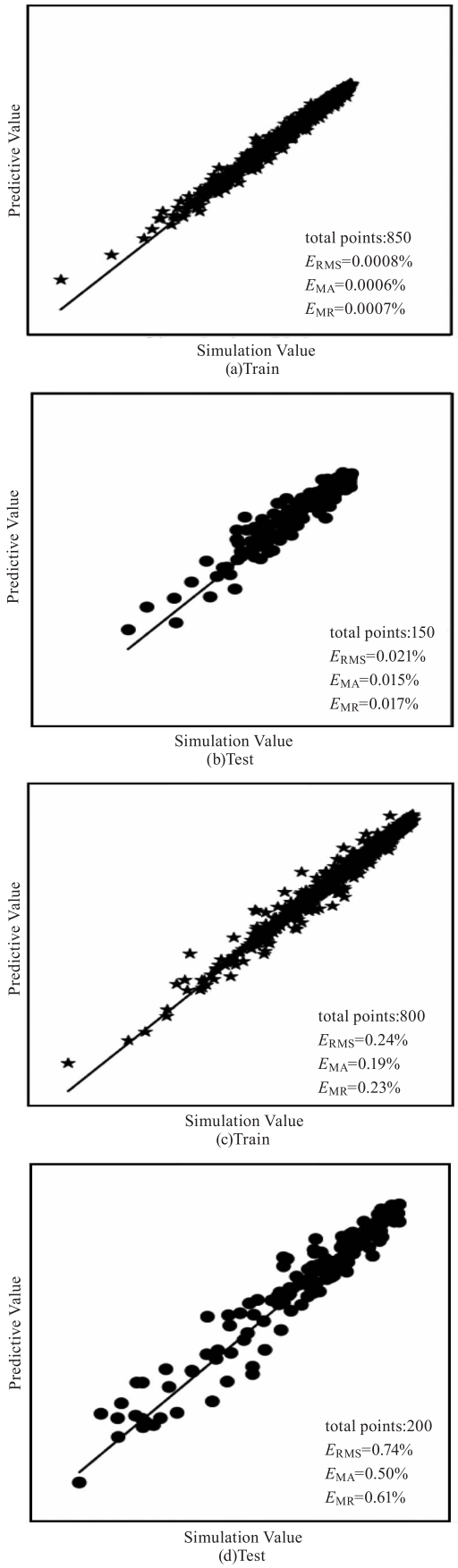

图 5　SVM、RFR 的 PIB 预测结果散点图

Fig. 5 Scatter diagram of PIB prediction results of SVM and RFR

3.4　RFR 中变量的影响分析

图 6(a)为 RFR 中自变量对 PIB 的重要性的排序结果。本文给出的 7 个属性变量,只有 5 个变量产生了一定影响,变量按重要性大小排序:$L > C_n^2 > v > t > V$,其中变量 ω_0 和 P_0 对 PIB 的重要性为零。图 6(b)用皮尔逊相关系数(Pearson coefficient)表征变量两两之间的相关性。从图中可得,PIB 与 V、v、t 呈弱相关性(相关系数位于[0,0.25]),与 C_n^2 呈中等程度相关性(相关系数位于[0.25,0.50]),与 L 为强相关性(相关系数位于[0.50,0.75])。图 6(b)中 ω_0、P_0 与其他变量(包括 PIB)相关性几乎为零,图 6(a)和图 6(b)关于 ω_0、P_0 的结论具有一致性,是因为 ω_0、P_0 经检验其方差为零,这样的 ω_0、P_0 对于样本的区分没有作用,即 RFR 进行样本区分时的 ω_0、P_0 的贡献为零。

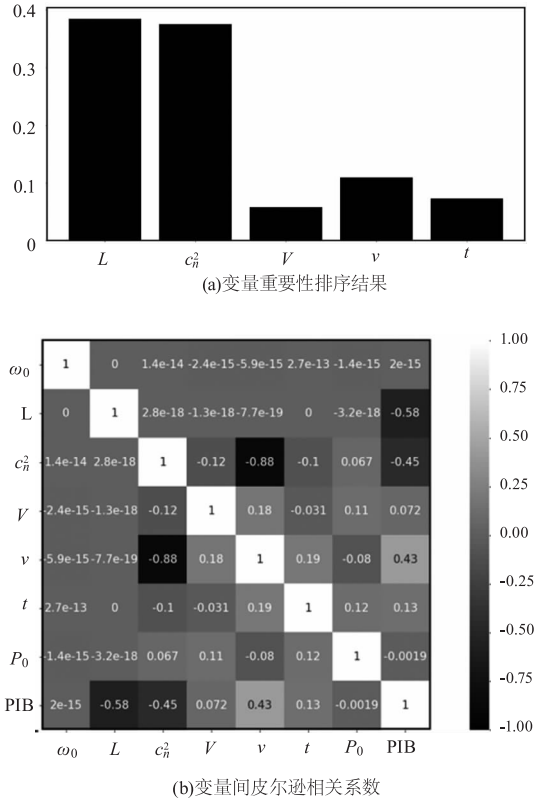

图 6　变量重要性排序结果和变量间皮尔逊相关系数

Fig. 6 Variable importance ranking results and Pearson correlation coefficients between variables

4　结　论

本文参照 2014 年 1 月烟台地区相关数据分布结合随机采样方法,通过多层相位屏模型获取了更接近实际的仿真数据。从仿真数据出发,基于随机森林回归算法构建了激光大气传输效应评估模型,

研究了激光发射参数和大气环境参数对 PIB 的影响。结果表明：

(1)训练集、测试集的预测结果与仿真数据的均方根误差优于 0.021 %，同时在经过交叉验证所得均方根误差均值 0.035 % 范围内，模型符合要求。

(2)随机森林比支持向量机能更好的拟合激光初始功率、能见度、湍流强度、风速、温度、传输距离与 PIB 的多元回归关系，拟合程度为 0.906。

(3)PIB 分别与激光风速、温度、能见度呈弱相关性，与湍流强度呈中等程度相关性，与传输距离呈强相关性。

(4)基于随机森林的输入变量重要性排序为：传输距离 > 湍流强度 > 风速 > 温度 > 能见度。

参考文献：

[1] Zhu Wenyue, Wang Huihua, Chen Xiaowei, et al. Uncertainty analysis of evaluation of high power laser propagation in atmosphere[J]. Chinese Journal of Quantum Electronics, 2020, 37(5): 525 – 532. (in Chinese)
朱文越，王辉华，陈小威，等. 高能激光大气传输评估的不确定性研究[J]. 量子电子学报，2020，37(5)：525 – 532.

[2] Jellen C, Burkhardt J, Brownell C, et al. Machine learning informed predictor importance measures of environmental parameters in maritime optical turbulence[J]. Applied Optics, 2020, 59(21): 6379 – 6389.

[3] Lv Jie, Zhu Wenyue, Cai Jun, et al. Comparison of two approaches for estimating atmospheric optical turbulence intensity near sea[J]. ACTA Optica Sinica, 2017, 37(5): 0501001. (in Chinese)
吕洁，朱文越，蔡俊. 两种估算近海面大气光学湍流强度方法的比较[J]. 光学学报，2017，37(5)：0501001.

[4] Leng Kun, Yang Yuntao, Tan Zhe, et al. Evaluation method of laser atmospheric propagation efficiency based on support vector machine[J]. Journal of Quantum Electronics, 2020, 37(5): 548 – 555. (in Chinese)
冷坤，杨云涛，谭哲，等. 基于支持向量机的激光大气传输效能评估方法[J]. 量子光学学报，2020，37(5)：548 – 555.

[5] Wu Xiaojun, Wang Hongxing, Li Bifeng, et al. Statistical analysis of atmospheric refractive index structure parameter under the sea surface environment[J]. Acta Optica Sinica, 2015, 35(4): 0401002. (in Chinese)
吴晓军，王红星，李笔锋. 近海面大气折射率结构常数统计特性分析[J]. 光学学报，2015，35(4)：0401002.

[6] Yang Yuntao, Leng Kun, Wu Wenyuan, et al. A simulation method of far field laser atmospheric propagation based on multilayer complex phasescreen characterization. Chinese: CN – 109933859A[P]. 2019 – 06 – 25. (in Chinese)
杨云涛，冷坤，武文远. 一种基于多层复数相位屏表征的远场激光大气传输仿真方法. 中国：CN109933859A[P]. 2019 – 06 – 25.

[7] Hang Qi, Yang Jinghui, Huang Guorong. Application of random forest algorithm in air quality evaluation[J]. Journal of Shanghai Second Polytechnic University, 2018, 35(2): 129 – 133. (in Chinese)
杭琦，杨敬辉，黄国荣. 随机森林算法在空气质量评价中的应用[J]. 上海第二工业大学学报，2018，35(2)：129 – 133.

[8] Zhou Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016. (in Chinese)
周志华. 机器学习[M]. 北京：清华大学出版社，2016.

[9] Luo Ren. Towards the retrieval methods for land surface temperature in urban areas[D]. Chengdu: University of Electronic Science and Technology of China, 2019. (in Chinese)
罗仁. 城市下垫面地表温度反演方法研究[D]. 成都：电子科技大学，2019.