doi:10.6041/j.issn.1000-1298.2019.05.031

基于 XGBoost-ANN 的城市绿地净碳交换模拟与特征响应

齐建东[1]　黄金泽[1]　贾　昕[2]

(1.北京林业大学信息学院,北京 100083;2.北京林业大学水土保持学院,北京 100083)

摘要:为分析城市绿地净生态系统碳交换(Net ecosystem exchange,NEE)对环境因子的响应,利用涡度相关法测量了 2013—2016 年生长季白天的 NEE 数据,使用 XGBoost 以及 ANN 模型对 NEE 进行模拟和分析,并通过决定系数(R^2)、平均绝对误差(MAE)、均方根误差(RMSE)和一致性系数(IA)4 个指标评价模拟精度。结果表明,当输入因子为光合有效辐射(PAR)、饱和水汽压差(VPD)、空气温度(Ta)、相对湿度(RH)、土壤温度(Ts)、风速(WS)、10 cm 处土壤含水率(VWC10)时,模拟效果达到最优。其训练集精度 R^2 为 0.712,RMSE 为 4.394 μmol/(m²·s),MAE 为 3.129 μmol/(m²·s),IA 为 0.911;测试集精度 R^2 为 0.748,RMSE 为 4.253 μmol/(m²·s),MAE 为 2.971 μmol/(m²·s),IA 为 0.920。在考虑因子间相互作用后,环境因子对 NEE 的重要性排序从大到小依次为 PAR、VPD、Ta、RH、Ts、WS、VWC10;就单环境因子而言,对 NEE 的重要性由大到小依次为 Ta、Ts、RH。通过计算生态系统净生产力(Net ecosystem productivity,NEP,即-NEE)对主要环境因子(PAR、VPD、Ta)的偏导数可知,生态系统光合作用表观量子效率最大值为 0.087,并且当 PAR 大于 1 200 μmol/(m²·s)时,其不再是影响光合作用的主要因素;VPD 偏导数的变化趋势表明,VPD 对植物光合作用的影响以抑制性为主,当 VPD 过大时,偏导数趋近于 0,此时植物叶片气孔闭合,抑制光合作用;Ta 偏导数的变化趋势说明,随着温度的升高,光合作用速率逐渐大于呼吸作用的速率。研究表明,基于 XGBoost 与 ANN 模型能够更为精确地模拟 NEE 动态,在相关环境因子中,PAR、VPD、Ta 是影响 NEE 变化的主导因子,NEE 对主要影响因子的生态特征响应趋势可为理解碳循环关键过程提供参考。

关键词:碳通量;XGBoost;人工神经网络;环境因子;涡度协方差

中图分类号:Q148　　**文献标识码**:A　　**文章编号**:1000-1298(2019)05-0269-10

Simulation of NEE and Characterization of Urban Green-land Ecosystem Responses to Climatic Controls Based on XGBoost - ANN

QI Jiandong[1]　HUANG Jinze[1]　JIA Xin[2]

(1. *College of Information Science and Technology*, *Beijing Forestry University*, *Beijing* 100083, *China*
2. *School of Soil and Water Conservation*, *Beijing Forestry University*, *Beijing* 100083, *China*)

Abstract: Aiming to analyze the responses of urban green-land's net ecosystem exchange (NEE) to the climatic controls and provide theoretical and technical support for carbon cycle simulation between land and atmosphere. In growing season, half-hourly daytime NEE based on eddy covariance flux data collected from 2013 to 2016 were simulated by XGBoost and back propagation artificial neural network (ANN) model. Moreover, the accuracy of model was evaluated by using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and index of agreement (IA). The experimental results showed that ANN model presented that seven input variables (photosynthetically active radiation (PAR), vapor pressure deficit (VPD), air temperature (Ta), relative humidity (RH), soil temperature (Ts), wind speed (WS) and volumetric water content at 10 cm depth) performed best, yielding R^2 of 0.712, RMSE of 4.394 μmol/(m²·s), MAE of 3.129 μmol/(m²·s) and IA of 0.911 on train dataset, and R^2 of 0.748, RMSE of 4.253 μmol/(m²·s), MAE of 2.971 μmol/(m²·s) and IA of 0.920 on test dataset. After considering the function and interaction among the factors, the importance score of each environmental factor was decreased in the following order: PAR, VPD, Ta, RH, Ts, WS and VWC10, otherwise Ts would be more important than RH. In particularly, after calculating the numerical partial derivatives of main climatic controls for each half-hourly point, the numerical partial

收稿日期:2019-02-21　修回日期:2019-03-04

基金项目:国家重点研发计划项目(2017YFC0504400、2017YFC0504406)和中央高校基本科研业务费专项资金项目(2015ZCQ-SB-02)

作者简介:齐建东(1976—),男,教授,博士生导师,主要从事智能信息处理、生态监测与模拟研究,E-mail:qijd@bjfu.edu.cn

derivatives of PAR showed the ecosystem quantum yield with the value of 0.087, and it also indicated that PAR was no longer a main impact factor when value was greater than 1 200 μmol/(m^2·s). Besides, the numerical partial derivatives of VPD expressed that VPD could mainly inhibit the photosynthesis, and the higher VPD aggravated the inhibition of photosynthesis by affecting photosynthetic rate. Furthermore, the numerical partial derivatives of Ta demonstrated that the photosynthetic rate was increased bit by bit and made the photosynthetic rate overpass respiration rate gradually. According to the result, PAR, VPD and Ta played an important role in controlling the NEE of urban green-land ecosystem. Also, XGBoost and ANN could be capable in capturing NEE dynamics and simulating the NEE with high accuracy. Meanwhile, the present result provided instant insight in underlying ecosystem physiology.

Key words: carbon flux; XGBoost; artificial neural network; environmental factors; eddy covariance

0 引言

近年来,随着城市化进程的不断推进,城市绿地面积也在不断增加,截止 2016 年,北京城市绿地总面积约为 3×10^4 hm^2,人均公园绿地面积为 16.1 m^2,城市绿地覆盖率达 48.4%[1]。人工林植被作为城市绿地的重要组成部分,对气候具有重要的调控作用。定量分析城市绿地净生态系统碳交换(Net ecosystem exchange,NEE)数据不仅能够促进人们对区域碳源、汇功能的理解,还可为研究不同生态系统对于全球气候变化的反馈机制、预测区域气候变化提供参考。

NEE 数据具有明显的季节性特征,并且与温度、光合有效辐射等各类气象、环境因子存在复杂的非线性关系[2-3],因此,模型模拟难度较大,难以保证模拟效果。虽然通过试凑法或者经验选择法选择环境因子往往能获取较高的精度,但过于复杂的生态学意义不明晰,且不利于模型推广。随着机器学习技术的不断进步,以随机森林、XGBoost 和人工神经网络(Artificial neural network,ANN)为代表的机器学习算法能够有效地克服上述缺陷,广泛适用于生态学领域。在城市绿地净碳交换模拟中,王宏莹[4]使用 BP-ANN 对 NEE 进行插补,评价徐州南部城区碳源、汇情况,MENZER 等[5]通过对比 3 种不同的 ANN 模型,探究风速与风向对于城市生态系统中 NEE 的影响。虽然由于 ANN 强大的非线性映射能力而被广泛使用,但是在输入因子选择等方面则需要研究人员凭经验确定[3],因此具有不确定性。XGBoost 模型是 CHEN 等[6]于 2016 年提出的模型,该算法不仅可以通过计算输入因子的相对重要性对结果进行解释,而且通过内置交叉验证等方法有效防止模型过拟合,使计算结果更为科学可靠,目前已经被广泛用于空气质量预报[7]等领域。充分利用 XGBoost 对结果的可解释性,将其与 ANN 模型相结合,可以有效弥补 ANN 在因子选择方面的缺陷。

在城市生态系统中,地表接收的太阳辐射强度和空气温度和湿度、土壤含水率以及风速等环境因素受到公园内植物多样性、人工林管理方式以及建筑布局等人类活动的影响。这些城市特征性使生态系统对环境因子的响应发生了巨大变化,已经引起科学家广泛关注。目前已经有大量学者开始探究城市生态系统中影响 NEE 的主要因子,认为 NEE 对环境因子的响应主要与光合有效辐射、温度、风速、风向有关[8-9],但是对于城市生态系统中 NEE 对环境因子的响应尚缺乏深入探讨。鉴于此,本文基于北京奥林匹克森林公园 2013—2016 年生长季白天利用涡度相关法连续观测的 NEE 数据,采用 XGBoost 方法分析空气温度(Ta)、土壤温度(Ts)、光合有效辐射(PAR)、风速(WS)、相对湿度(RH)、饱和水汽压差(VPD)、10 cm 深度土壤含水率(VWC10)7 个环境因子对 NEE 的影响程度,并对其进行选择和评价,利用 ANN 神经网络模型解释 NEE 对主要环境因素的响应结果。

1 材料与方法

1.1 研究区概况

研究区位于北京奥林匹克森林公园(40°01′N,116°23′E),海拔 51 m,主要土壤类型为潮褐土,4 年均温变化范围为 11.69 ~ 13.01℃,1 月与 7 月均温变化范围分别为 -5.22 ~ -1.01℃ 与 25.45 ~ 27.82℃,极端低温变化范围 -19.00 ~ -12.01℃,极端高温变化范围为 36.58 ~ 39.82℃。该地区气候属于暖温带半湿润大陆性季风气候,四季分明,4 年均降水量变化范围 458.61 ~ 669.12 mm,无霜期 208 ~ 225 d。降水季节分配不均,主要集中于 6、7、8 月(2013—2016 年北京统计年鉴)。观测地内主要植被为人工营造的乔冠草复层景观林,乔木类代表物种包括油松(*Pinus tabulaeformis*)、侧柏(*Platycladus orientalis*)、国槐(*Sophora japonica*)、白蜡(*Fraxinus chinensis*)、银杏(*Ginkgo biloba*),灌木主要为山桃(*Prunus davidiana*),丛生灌木主要为丁香(*Syzygium*

aromaticum)、地被植物主要为石竹(*Dianthus chinensis*)等[10]。

1.2　数据获取及预处理

实验数据通过涡度通量观测仪测量获取。仪器主要包括三维超声风速仪(CSAT3 型,Campbell Scientific Ltd.,美国)、红外气体分析仪(EC155 型,Campbell Scientific Ltd.,美国)、净辐射仪(CNR－4 型,Kipp & Zonen Inc.,荷兰)、光量子传感器(PARLITE 型,Kipp & Zonen Inc.,荷兰)、空气温湿度传感器(HMP45C 型,Campbell Scientific Ltd.,美国)、土壤温度传感器(Campbell－109 型,Campbell Scientific Ltd.,美国)及土壤含水率传感器(CS616 型,Campbell Scientific Ltd.,美国),分别用于测量风速、CO_2/H_2O 密度脉动、辐射、空气温湿度以及土壤温度等微气象数据。数据采集器(CR3000 型,Campbell Scientific Ltd.,美国)以 10 Hz 频率记录涡度通量观测仪的数据,经过野点剔除、二次坐标轴旋转、频率响应校正和 WPL 校正等操作后,在线计算 30 min 通量值[10]。

采用 2013—2016 年 30 min 通量塔 NEE 与微气象数据,由于恶劣天气、硬件设施故障以及人为因素的影响,数据存在异常值和缺失值。针对上述情况,参考中国通量网通量数据标准处理流程与净生态系统交换量标准处理流程[4,11],将数据进行如下操作:

(1)利用 3 倍标准差法剔除野点,检查数据范围。

(2)存储通量计算。考虑到奥林匹克森林公园植被类型空间分布均匀,植被较高,具有涡度相关系统高度下冠层内存储通量不为零的特点。NEE 定义为

$$P_{\mathrm{NEE}} = F_c + F_s \tag{1}$$

式中　P_{NEE}——净生态系统碳交换量

F_c——通量观测塔在植被上部观测值

F_s——涡度通量观测仪安装高度下冠层内存储通量

(3)使用分段平均值检验法计算的摩擦风速阈值为 0.2 m/s,因此剔除夜间 NEE 数据中摩擦风速小于 0.2 m/s 的碳通量数据。

(4)考虑到不同量纲的数据序列会增加数据处理成本与模型拟合时间,对数据进行归一化处理。最后,选取 4 年中生长季(6—9 月)白天[14],即 PAR 大于 10 μmol/(m²·s)的数据作为研究对象[12]。各年有效数据统计结果如表 1 所示。

表 1　各年生长季白天 NEE 有效数据统计

Tab. 1　Summary of effective growing season daytime NEE data for each year　　条

年份	NEE 有效数据量				合计
	6 月	7 月	8 月	9 月	
2013	693	716	773	674	2 856
2014	587	559	462	532	2 140
2015	545	676	579	536	2 336
2016	745	773	724	665	2 907

1.3　模拟方法

1.3.1　XGBoost 模型与因子选择

XGBoost 模型属于集成学习模型,通过构建并结合多个回归树模型来完成学习任务。首先通过自举法(Bootstrap)生成 N 个训练集。其次,对于每个训练集均建立回归树模型并进行训练。最后,计算所有回归树模拟结果,加权后作为输入变量的预测值。具体流程如图 1 所示。由于 XGBoost 模型在训练过程中根据每个输入因子的信息增益选择最好的特征进行分裂,因此,通过计算所有回归树中第 i 个输入因子 x_i 出现的次数,可得到该输入因子在整个 XGBoost 模型中的重要性得分。计算公式为

$$\mathrm{score}(x_i) = \sum_{n=1}^{N_t} \mathrm{score}_n(x_i)/N_t \tag{2}$$

式中　n——XGBoost 中树的编号

N_t——XGBoost 中树的数量

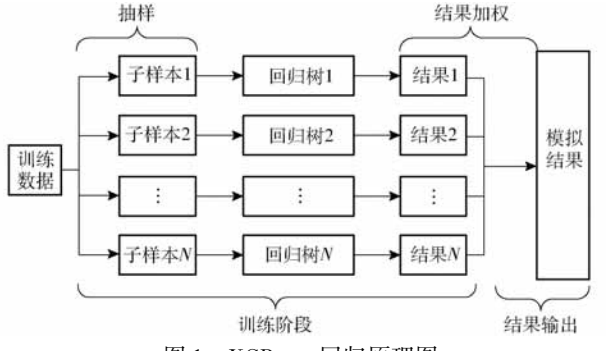


图 1　XGBoost 回归原理图

Fig. 1　XGBoost regression principle

1.3.2　ANN 神经网络及其偏导数

神经网络主要是通过自学习寻找目标值与输入变量之间的映射关系。一个神经网络主要由输入层、输出层以及中间的隐含层构成(图 2)。输入层负责接收输入数据,输出层负责输出整个神经网络的计算结果,隐含层则负责描述问题的层次关系。神经网络上每个节点称之为神经元,各层之间的神经元通过一定的权重相互连接。其基本原理为:当网络输出层的计算结果与期望结果偏差过大时,通过优化算法对每个神经元的权重进行更新。通常 ANN 神经网络使用反向传播算法作为优化算法,其本质是计算目标函数的梯度来寻求最优值。在数学上,梯度值表示目标函数在该点变化率最大的方向。

在生态学中,则可以表示为 NEE 对于环境因子的响应速率[13]。

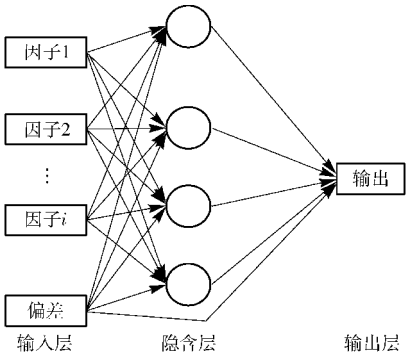


图 2 ANN 分析原理图

Fig. 2 Diagram of ANN analysis principle

本文采取自动调整学习率以及早停策略防止 ANN 过拟合或者欠拟合。

1.3.3 基于 XGBoost 与 ANN 神经网络的 NEE 分析模型

基于上述算法,为了能够更为准确地分析影响城市生态系统 NEE 的主要影响因子与其响应关系,本文将两种算法结合进行分析。其分析流程如图 3 所示,首先对采集到的 NEE 数据进行质量控制和归一化预处理;其次利用 XGBoost 模型筛选出 NEE 主要影响因子,并将其作为 ANN 模型的输入因子;通过 ANN 模型对各输入因子的偏导数,探究 NEE 对各环境因子的响应关系,实现城市生态系统 NEE 对主要环境因子的响应分析。

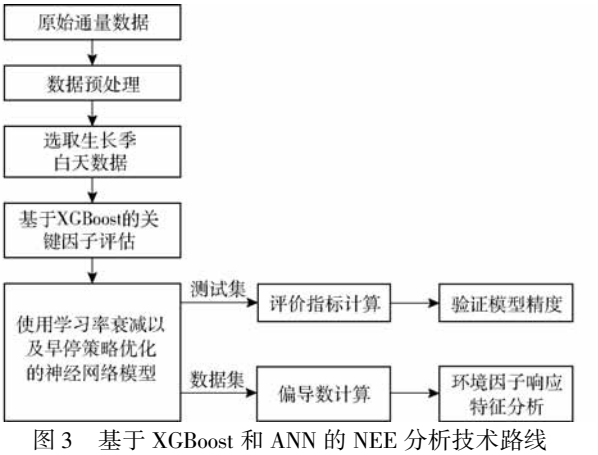


图 3 基于 XGBoost 和 ANN 的 NEE 分析技术路线

Fig. 3 CO_2 flux analysis flow chart based on XGBoost and ANN

1.4 性能评估

为评估 ANN 模型的拟合效果,本研究中使用了目前最常用的评估标准[14]:决定系数 R^2、平均绝对误差(Mean absolute error, MAE)、均方根误差(Root mean square error, RMSE)和一致性系数(Index of agreement, IA)。

研究所用程序设计语言为 Python 3.6(64-bit),集成开发环境为 Anaconda 3。程序设计中,XGBoost 模型基于 xgboost 包实现,ANN 模型则由 Keras 2.2.0 和 TensorFlow 1.6.0 完成编写。为保证结果可靠性,每组实验在相同条件下重复 100 次,实验结果取平均值。

2 结果与分析

2.1 模型参数优化

2.1.1 XGBoost 模型

在 XGBoost 模型中,主要参数有 3 个:每棵树的最大深度、树的数量、最小叶子节点权重之和。本文通过网格寻优策略测试了 3 种参数的 240 种组合,最终确定当树的最大深度为 8、树的数量为 1 500、最小叶子节点权重之和为 200 时,XGBoost 模型的目标损失函数值最小,达到最优。

2.1.2 ANN 神经网络

本文设定初始学习率为 1.5,最小学习率为 0.001;迭代次数经早停策略优化后平均次数为 473 次;批大小为 16,所用隐含层神经元数量为 4,输入层与隐含层共用 bias 单元,如图 2 所示。另外,激活函数采用 sigmoid 函数。

2.2 环境因子变化特征

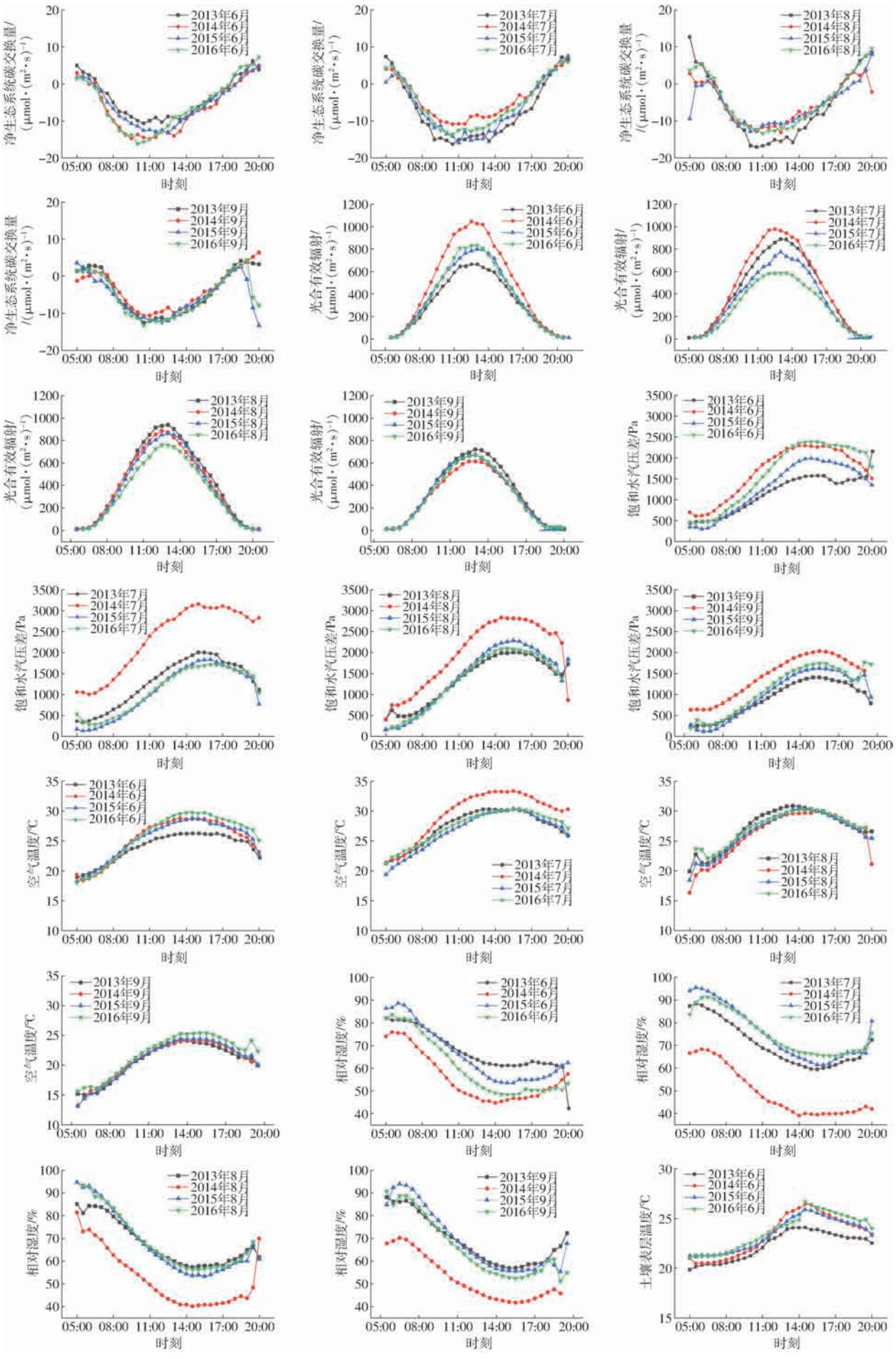
4 年内生长季白天时间段内环境变化的月平均日变化基本特征如图 4 所示。本研究数据时间段,北京奥林匹克森林公园的 PAR、Ta、Ts、VPD 在 12:00—14:00 达到极值,PAR 极大值出现在 2013 年 7 月 16 日,为 1 533 μmol/(m²·s),月平均日变化极值出现在 2014 年 6 月,为 1 045 μmol/(m²·s);气温的月平均日变化极小值为 24.12℃,极大值为 33.35℃;由于 VPD 主要受到水热条件以及植物蒸腾作用的影响,生长季期间饱和水汽压差明显波动,且变化剧烈。2014 年的 VPD 显著高于其他年份;土壤含水率反映了区域降水情况[15],2014 年 9 月的降水使得 VWC10 指标与其他 3 年呈明显差异,VWC10 与 WS 4 年内月平均日变化范围分别为 16.54% ~35.36% 与 0.93 ~1.56 m/s。

2.3 环境因子重要性得分

图 5 为计算得到的环境因子对 NEE 影响的重要性得分。可以看出,环境因子对 NEE 影响的重要性得分由大到小表现依次为 PAR、VPD、Ta、RH、Ts、WS、VWC10。由此可知,PAR、VPD、Ta 是影响奥林匹克森林公园植物生长季 NEE 变化的重要因素。

2.4 NEE 模拟结果以及对主要环境因子的响应过程

表 2 为不同组合的输入指标在测试数据集上的评价结果。计算结果表明,随着环境因子输入数量


净生态系统碳交换量/(μmol·(m²·s)⁻¹)
光合有效辐射/(μmol·(m²·s)⁻¹)
饱和水汽压差/Pa
空气温度/℃
相对湿度/%
土壤表层温度/℃
时刻
05:00 08:00 11:00 14:00 17:00 20:00
2013年6月
2014年6月
2015年6月
2016年6月
2013年7月
2014年7月
2015年7月
2016年7月
2013年8月
2014年8月
2015年8月
2016年8月
2013年9月
2014年9月
2015年9月
2016年9月

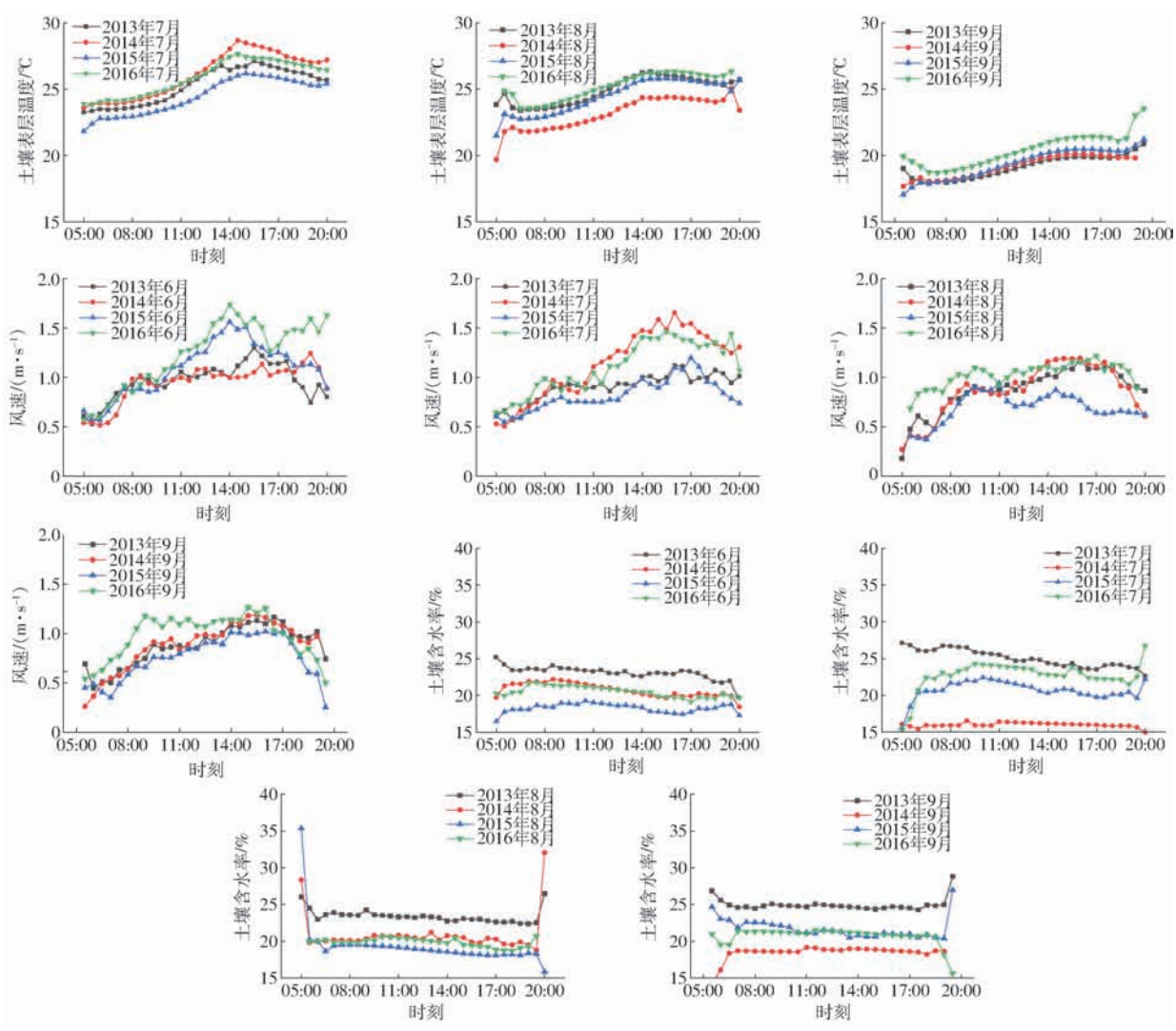


图 4 2013—2016 年生长季白天各指标的月平均日变化量

Fig. 4 Growing season daytime monthly diurnal variations of input variables from 2013 to 2016

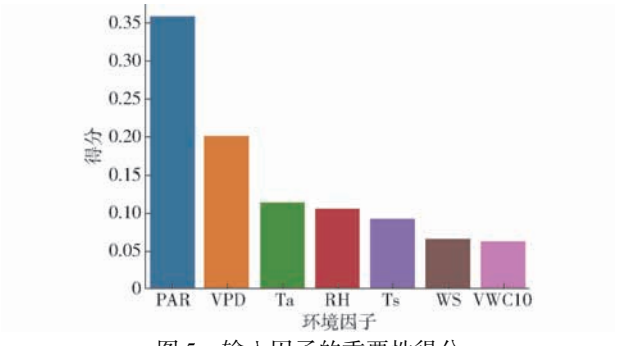


图 5 输入因子的重要性得分

Fig. 5 Importance score of input variables

的增加,R^2总体逐渐增加,当输入因子为 PAR、VPD、Ta、RH、Ts、WS、VWC10 时,训练集 R^2为 0.712,MAE 为 3.129 μmol/(m²·s),RMSE 为 4.349 μmol/(m²·s),一致性指数为 0.911,测试集决定系数 R^2为 0.748,RMSE 与 MAE 分别为 4.253、2.971 μmol/(m²·s),IA 为 0.920,相较于其他组合为最优结果。训练集模拟值与观测值的拟合结果如图 6 所示,测试集的拟合结果如图 7 所示。

观察图 5 与表 2 结果可知,当输入因子依次加入 VPD 以及 Ta 时,模型测试集 R^2增加显著(R^2从 0.587 提升到 0.741);若继续增加环境因子数量,R^2提升效果并不明显。因此,本文选取 PAR、VPD、Ta 3 个环境因子进行分析讨论。

2.4.1 PAR 对 NEP 的影响

NEP(Net ecosystem productivity)为生态系统净生产力。由图 5 可知,在生长季,PAR 是影响白天净生态系统交换量的决定性因素。图 8 显示了 NEP 随 PAR 的变化情况以及对 PAR 的偏导数,即表观量子效率。从图中可知,$\partial P_{NEP}/\partial P_{PAR}>0$($P_{NEP}$为生态系统净生产力,$P_{PAR}$为光合有效辐射),光合作用随着 PAR 的增大而逐渐增强,生态系统对于 CO_2的吸收量逐渐增大,碳汇能力增强。当 PAR 大于 1 200 μmol/(m²·s)时,表观量子效率逐渐趋于 0 并保持稳定,说明此时光合有效辐射已经不是部分植被进行光合作用的主要因素。图 8 中,最大表观量子效率为 0.087。

表 2　不同输入因子模拟结果评估

Tab. 2　Evaluation indices of different combinations

编号	输入因子							R^2		MAE/（μmol·（m²·s）⁻¹）		RMSE/（μmol·（m²·s）⁻¹）		IA	
	PAR	VPD	Ta	RH	Ts	WS	VWC10	训练集	测试集	训练集	测试集	训练集	测试集	训练集	测试集
CIV. 1	Y	N	N	N	N	N	N	0. 548	0. 587	4. 114	4. 054	5. 452	5. 452	0. 836	0. 846
CIV. 2	Y	Y	N	N	N	N	N	0. 684	0. 719	3. 301	3. 201	4. 555	4. 503	0. 898	0. 907
CIV. 3－1	Y	Y	Y	N	N	N	N	0. 704	0. 741	3. 184	3. 059	4. 404	4. 302	0. 905	0. 916
CIV. 3－2	Y	Y	N	Y	N	N	N	0. 694	0. 729	3. 240	3. 133	4. 480	4. 308	0. 903	0. 911
CIV. 3－3	Y	Y	N	N	Y	N	N	0. 702	0. 733	3. 195	3. 096	4. 421	4. 307	0. 904	0. 914
CIV. 4－1	Y	Y	Y	Y	N	N	N	0. 707	0. 742	3. 141	3. 040	4. 381	4. 292	0. 909	0. 917
CIV. 4－2	Y	Y	Y	N	Y	N	N	0. 705	0. 742	3. 166	3. 043	4. 393	4. 306	0. 907	0. 917
CIV. 5	Y	Y	Y	Y	Y	N	N	0. 708	0. 743	3. 139	3. 025	4. 374	4. 292	0. 908	0. 917
CIV. 6	Y	Y	Y	Y	Y	Y	N	0. 711	0. 746	3. 131	3. 017	4. 354	4. 278	0. 909	0. 918
CIV. 7	Y	Y	Y	Y	Y	Y	Y	0. 712	0. 748	3. 129	2. 971	4. 349	4. 253	0. 911	0. 920

注：Y 为输入因子中包括该因子，N 为未包括该因子。

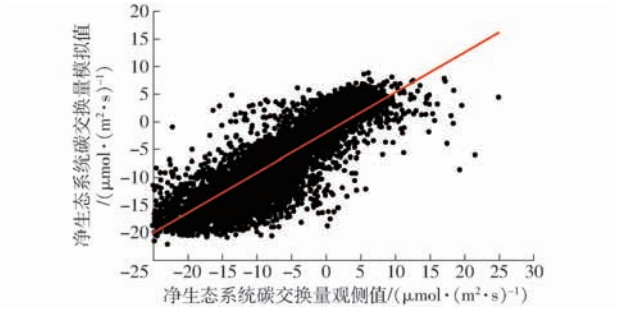


图 6　训练集观测值与模拟值的关系

Fig. 6　Relationship between observed and simulation results in train dataset

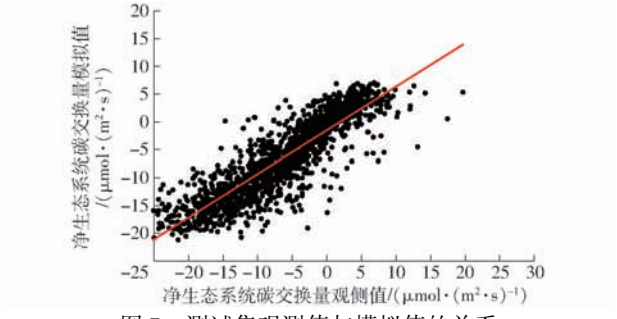


图 7　测试集观测值与模拟值的关系

Fig. 7　Relationship between observed and simulation results in test dataset

2. 4. 2　饱和水汽压差对 NEP 的影响

VPD 可由 RH、Ta 通过公式估算得出，该指标能够体现出温度与相对湿度的特点[16]，因此，其重要性得分要高于 RH、Ta。当 VPD 过高时，会对植物叶片气孔闭合产生压力，影响植物的光合和蒸腾作用。由图 9 可知，当 VPD 较小时，植被生长环境适宜的情况下，$\partial P_{NEP}/\partial P_{VPD} > 0$（$P_{VPD}$ 为饱和水汽压差），NEP 与 VPD 正相关，促进生态系统碳汇，但促进能力逐渐减弱。随着 VPD 的增大，研究区域内大部分植物光合作用受到抑制，但仍有一部分时刻的偏导数大于 0，对 NEP 的变化起促进作用，但总体趋近于 0。这说明即使在较高 VPD 时，由于研究区域内植物种类多样，其中包括侧柏等耐高温、抗旱性强植物，生态系统总体依然呈碳汇现象，但是随着 VPD 的继续增加，对植物叶片气孔闭合产生压力，影响植物的光合和蒸腾作用[8,22]。

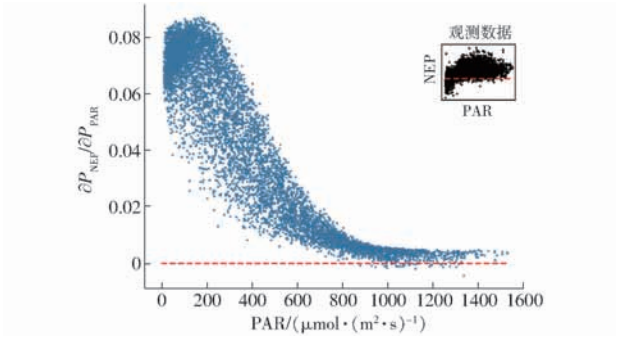


图 8　NEP 对 PAR 的偏导数

Fig. 8　Numerical partial derivatives of daytime NEP response to PAR for each half-hourly data

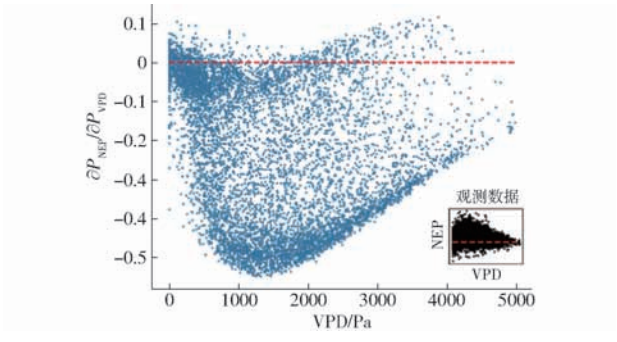


图 9　NEP 对 VPD 的偏导数

Fig. 9　Numerical partial derivatives of daytime NEP response to VPD for each half-hourly data

2. 4. 3　空气温度对 NEP 的影响

空气温度主要通过影响生态系统呼吸和光合作用来影响生态系统 CO_2 的交换。在本文中，当温度大于 15℃ 时，大部分时刻空气温度与 NEP 呈正相

关,小部分呈负相关。在合适的温度范围内,随着温度的增加,逐渐转变为正相关(图 10,P_{Ta}为空气温度)。该现象说明温度较低时,生态系统光合作用弱于呼吸作用,随着温度的逐渐升高,光合作用速率逐渐加快,大于呼吸作用速率。这与陈文婧等[10]关于奥林匹克森林公园温度与生态系统碳交换影响的研究结果类似。

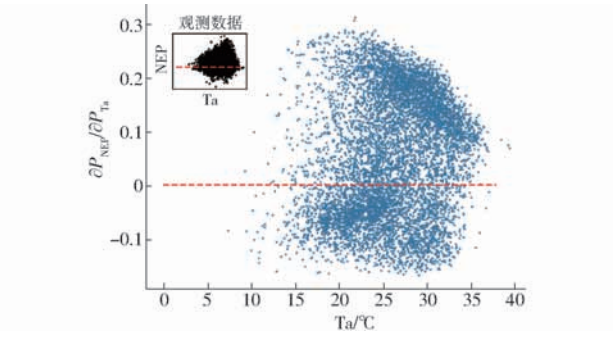


图 10 NEP 对 Ta 的偏导数

Fig. 10 Numerical partial derivatives of daytime NEP response to Ta for each half-hourly data

3 讨论

3.1 样本大小及模型参数对模拟精度的影响

与随机森林类似,XGBoost 模型的样本容量较小时,难以保证极高的模拟精度与泛化能力,因此机器学习模型需要有足够的样本数量保证结果的合理性。本文模型在训练、测试与独立验证 3 个阶段样本数量分别为 5 544、1 109、2 376,样本容量较大。

模型模拟的结果除了与样本数量有关外,还受到模型参数的影响。在 XGBoost 模型中,树的深度、数量设置过大会导致模型运行时间过长,效率低下;若设置过小则无法有效地表达 NEE 的特征重要性,合理地设置最小权重之和可以避免模型学习到局部的特殊样本[6]。本文采用的网格寻优策略,是目前较常用的寻优算法,与机器学习模型相结合,广泛应用于煤炭灰熔点预测[17],工矿复垦区土地利用分类等各个领域[18]。通过该算法优化 XGBoost 模型参数,可以较好地避免由于参数设置不恰当所造成的误差。

在神经网络模型中,除网络结构外,学习率、迭代次数对模拟结果有巨大影响。学习率过大,导致模型无法收敛,学习率过小,会导致模型收敛速度过慢等。本文采取学习率衰减的方法,并设定最小学习率为 0.001,训练过程中每隔 50 代学习率降低为原来的 1/2,直至达到最小学习率,该方法不仅加快了网络收敛速率,也较好地避免了振荡现象,已经被广泛应用于许多改进的 BP 算法中[19];迭代次数对于结果的影响体现为:过多的迭代次数导致模型产生过拟合,过少的迭代次数则会使模型在未收敛时训练结束,导致欠拟合。本文通过观察测试集损失函数,采用早停策略[20],保证在该参数配置下获取最优的拟合结果。

3.2 数据质量对模拟结果的影响

在数据收集过程中,由于大气降水、大气湍流、传感器自身误差等原因,造成数据出现缺失,极端噪声等问题,此类问题会对模型的精度造成极大的影响。在本文中,主要评估了包括光合有效辐射、饱和水汽压差在内的 7 个环境因素,未考虑生物因素,如气孔导度、水分利用效率等,因此无法更加有效模拟 NEE 的变化情况。另外,由于人类活动对城市生态系统的影响要远大于对其他生态系统的影响,因此,在未来的 NEE 的研究中,应当充分考虑植被生理因素、生长环境的气象因素以及城市交通[21]等因素的特点,选用更加精确的模型,以期提高 NEE 模拟精度。总体来说,XGBoost 模型能准确地选取影响生态环境 NEE 的主要因素,ANN 模型能够较好拟合 NEE 的大小以及变化趋势。

3.3 共线性变量对模型的影响

由图 5 可知,Ta、RH、Ts 3 个环境因子重要性得分接近。为了进一步探究 3 个因子的主导性,本文评估了 3 种组合的模拟效果(表 2,CIV.3-1,CIV.3-2,CIV.3-3),模拟优度从大到小依次为 CIV.3-1(PAR、VPD、Ta)、CIV.3-3(PAR、VPD、Ts)、CIV.3-2(PAR、VPD、RH)。说明在输入因子包括 PAR 和 VPD 的情况下,Ta 最重要,其次为 Ts,RH 重要性最小。与图 5 结果存在差异,出现这一问题的原因是,XGBoost 模型是决策树模型的组合,在模拟过程中输入因子共线性不会影响其预测能力,但是对数据的解释性影响巨大[22]。即当 Ta 特征被使用时,XGBoost 模型对于 Ts 特征的权重将会减少,此时 Ts 相对于 NEP 来说,增加的有效信息少于 RH,因此,在图 5 显示的重要性得分中,由大到小依次为 Ta、RH、Ts。另外,CIV.4-1(PAR、VPD、Ta、RH)与 CIV.4-2(PAR、VPD、Ta、Ts)组合的评估结果中,CIV.4-1 的精度高于 CIV.4-2,这一结果也佐证了上述结论。图 11(图中 P_{Ts}为土壤温度的值)展示了当输入因子组合为 CIV.3-3 时,NEP 对 Ts 的偏导数,观察可知,其变化趋势与空气温度对于 NEP 的影响相同,但数值更小,对 NEP 的响应能力弱于 Ta。综上所述,就单因子而言,重要性由大到小依次为 Ta、Ts、RH;但综合考虑 Ta 后,RH 对 NEP 影响大于 Ts。

3.4 环境因子对 NEE 的影响

在本研究中,当 PAR 增大时,植物光合作用增

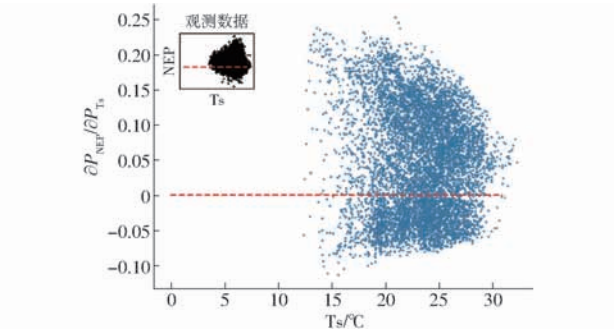


图 11　NEP 对 Ts 的偏导数

Fig. 11　Numerical partial derivatives of daytime NEP response to Ts for each half-hourly data

强，生态系统中植被固碳能力增强，最大表观量子效率为 0.087，高于其他生态系统中的计算值，出现这一现象是因为城市中气溶胶浓度较高使得散射辐射的比例增大，而碳收支对散射辐射敏感度较高所导致的[20]。当 PAR 大于 1 200 μmol/(m²·s)后，其不再是影响光合作用的主要因素，如图 8 所示。PAR 除了在城市生态系统中是 NEE 的主要影响因素外，在其他生态系统也起主要调控作用，WEN 等[23]通过遗传神经网络筛选出 PAR 是湖南会同杉木人工林生态系统的决定性因子，在半干旱荒漠生态系统中，PAR 对 NEE 的影响最为显著[24]，唐祥等[25]对北京八达岭 NEE 的研究表明，PAR 是影响该生态系统生长季 NEE 变化的主导因子。除 PAR 外，VPD、Ta 是影响生长季碳吸收的重要控制因素，该结论与陈文婧[8]针对城市绿地神态系统碳水通量研究结论一致，MOFFAT[20]对影响 Hainich 森林的环境因素重要性进行了等级划分，其研究结果也证明，VPD 与 Ta 是重要的非辐射性环境因子，KUNWOR 等[12]通过改进 Michaelis－Menten 关系式，在 PAR 与 Ta 的基础上，引入新的环境变量 VPD，不仅提高了数据模拟的精度，还在一定程度上保留了模拟数据与观测数据的方差结构。另外，温度主要通过影响生态系统光合作用以及土壤微生物活性等方式，来影响生态系统 NEE[26－27]。这也验证了 PAR、VPD、Ta 对于生态系统 NEE 模拟的重要性。综上，准确解释环境因子对于 NEE 的影响，有利于更好地认识植被对于区域气候变化的响应。

4　结论

(1)XGBoost－ANN 模型能够较好地捕捉北京奥林匹克森林公园生长季 NEE 数据的变化特点，在训练集上 R^2 为 0.712，MAE 为 3.129 μmol/(m²·s)，RMSE 为 4.349 μmol/(m²·s)，IA 为 0.911；测试集 R^2 为 0.748，RMSE 与 MAE 分别为 4.253、2.971 μmol/(m²·s)，IA 为 0.920，说明其在模拟和分析森林 NEE 方面具有较好的适用性。

(2)模型测试结果表明，在训练过程中，通过机器学习自动化调参以及网格寻优等操作，可以避免由于参数设置不合理带来的误差。

(3)环境因子重要性得分表明，在考虑各因子间相互作用的情况下，影响生长季白天 NEE 的主要环境特征重要性由大到小依次为 PAR、VPD、Ta、RH、Ts、WS、VWC10。该地区 NEE 变化主要受 PAR、VPD、Ta 3 个主要因素调控。

(4)各主要环境因子偏导数表明，北京奥林匹克森林公园生长季白天的光合作用表观量子效率最大值为 0.087，并且当 PAR 大于 1 200 μmol/(m²·s)时，PAR 不再是影响 NEP 的主要因素；对 VPD 的偏导数说明，随着 VPD 的增加，会对植物叶片气孔闭合产生压力，影响其光合和蒸腾作用；对温度的偏导数说明，随着温度的增加，光合速率逐渐大于呼吸速率。

参　考　文　献

[1]　北京统计局. 北京统计年鉴[J]. 北京：中国统计出版社，2017.

[2]　LONGDOZ B，YERNAUX M，AUBINET M. Soil CO_2 efflux measurements in a mixed forest：impact of chamber disturbances，spatial variability and seasonal evolution[J]. Global Change Biology，2000，6(8)：907－917.

[3]　MOFFAT A M，PAPALE D，REICHSTEIN M，et al. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes[J]. Agricultural & Forest Meteorology，2007，147(3－4)：209－232.

[4]　王宏莹. 基于涡度观测和遥感技术的城市碳源/汇研究[D]. 徐州：中国矿业大学，2016.
WANG Hongying. Study of urban carbon sources/sinks based on eddy covariance technology and remote sensing[D]. Xuzhou：China University of Mining and Technology，2016.（in Chinese）

[5]　MENZER O，MEIRING W，KYRIAKIDIS P C，et al. Annual sums of carbon dioxide exchange over a heterogeneous urban landscape through machine learning based gap-filling[J]. Atmospheric Environment，2015，101：312－327.

[6]　CHEN T，GUESTRIN C. XGBoost：a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. San Francisco：ACM，2016：785－794.

[7]　赵俊日，肖昕，吴涛，等. 空气质量数值预报优化方法研究[J]. 中国环境科学，2018，38(6)：2047－2054.
ZHAO Junri，XIAO Xin，WU Tao，et al. A revised approach to air quality forecast based on Models－3/CMAQ [J]. China Enviromental Science，2018，38(6)：2047－2054.（in Chinese）

[8]　陈文婧. 城市绿地生态系统碳水通量研究[D]. 北京：北京林业大学，2013.

CHEN Wenjing. Carbon and water fluxes of urban green-land ecosystem-case study of Beijing Olympic Forest Park[D]. Beijing: Beijing Forestry University, 2013. (in Chinese)

[9] 郭智明. 长沙市城市生态系统碳通量时间变化与影响因子研究[D]. 长沙: 中南林业科技大学, 2015.
GUO Zhiming. The study on temporal variation and impact factors of carbon flux between urban ecosystem and atmosphere in Changsha[D]. Changsha: Central South University of Forestry and Technology, 2015. (in Chinese)

[10] 陈文婧, 李春义, 何桂梅, 等. 北京奥林匹克森林公园绿地碳交换动态及其环境控制因子[J]. 生态学报, 2013, 33(20): 6712 - 6720.
CHEN Wenjing, LI Chunyi, HE Guimei, et al. Dynamics of CO_2 exchange and its environmental controls in an urban green-land ecosystem in Beijing Olympic Forest Park[J]. Acta Ecologica Sinica, 2013, 33(20): 6712 - 6720. (in Chinese)

[11] PAPALE D, REICHSTEIN M, AUBINET M, et al. Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation[J]. Biogeosciences, 2006, 3(4): 571 - 583.

[12] KUNWOR S, STARR G, LOESCHER H W, et al. Preserving the variance in imputed eddy-covariance measurements: alternative methods for defensible gap filling[J]. Agricultural & Forest Meteorology, 2017, 232: 635 - 649.

[13] MOFFAT A M, BECKSTEIN C, CHURKINA G, et al. Characterization of ecosystem responses to climatic controls using artificial neural networks[J]. Global Change Biology, 2010, 16(10): 2737 - 2749.

[14] 窦贤明. 机器学习方法在陆地生态系统碳水通量模拟中的应用研究[D]. 徐州: 中国矿业大学, 2018.
DOU Xianming. Applications of machine learning methods in modeling carbon and water fluxes of terrestrial ecosystems[D]. Xuzhou: China University of Mining and Technology, 2018. (in Chinese)

[15] JING X, XIN J, HE G, et al. Environmental control over seasonal variation in carbon fluxes of an urban temperate forest ecosystem[J]. Landscape & Urban Planning, 2015, 142: 63 - 70.

[16] 张红梅, 吴炳方, 闫娜娜. 饱和水汽压差的卫星遥感研究综述[J]. 地球科学进展, 2014, 29(5): 559 - 568.
ZHANG Hongmei, WU Bingfang, YAN Nana. Remote sensing estimates of vapor pressure deficit: an overview[J]. Advances in Earth Science, 2014, 29(5): 559 - 568. (in Chinese)

[17] 李清毅, 周昊, 林阿平, 等. 基于网格搜索和支持向量机的灰熔点预测[J]. 浙江大学学报(工学版), 2011, 45(12): 2181 - 2187.
LI Qingyi, ZHOU Hao, LIN Aping, et al. Prediction of ash fusion temperature based on grid search and support vector machine[J]. Journal of Zhejiang University (Engineering Science), 2011, 45(12): 2181 - 2187. (in Chinese)

[18] 陈元鹏, 罗明, 彭军还, 等. 基于网格搜索随机森林算法的工矿复垦区土地利用分类[J]. 农业工程学报, 2017, 33(14): 250 - 257.
CHEN Yuanpeng, LUO Ming, PENG Junhai, et al. Classification of land use in industrial and mining reclamation area based grid-search and random forest classifier [J]. Transactions of the CSAE, 2017, 33(14): 250 - 257. (in Chinese)

[19] 李翱翔, 陈健. BP 神经网络参数改进方法综述[J]. 电子科技, 2007(2): 79 - 82.
LI Aoxiang, CHEN Jian. Summarize of parameter improve methods for BP neural network[J]. Electronic Science and Technology, 2007(2): 79 - 82. (in Chinese)

[20] MOFFAT A M. A new methodology to interpret high resolution measurements of net carbon fluxes between terrestrial ecosystems and the atmosphere[D]. Berlin: Friedrich-Schiller-Universität, 2012.

[21] 张诗青, 王建伟, 郑文龙. 中国交通运输碳排放及影响因素时空差异分析[J]. 环境科学学报, 2017, 37(12): 4787 - 4797.
ZHANG Shiqing, WANG Jianwei, ZHENG Wenlong. Spatio-temporal difference of transportation carbon emission and its influencing factors in China [J]. Acta Scientiae Circumstantiae, 2017, 37(12): 4787 - 4797. (in Chinese)

[22] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81 - 106.

[23] WEN X, ZHAO Z, DENG X, et al. Applying an artificial neural network to simulate and predict Chinese fir (*Cunninghamia lanceolata*) plantation carbon flux in subtropical China[J]. Ecological Modelling, 2014, 294: 19 - 26.

[24] JIA X, ZHA T, GONG J, et al. Multi-scale dynamics and environmental controls on net ecosystem CO_2 exchange over a temperate semiarid shrubland[J]. Agricultural and Forest Meteorology, 2018, 259: 250 - 259.

[25] 唐祥, 陈文婧, 李春义, 等. 北京八达岭林场人工林净碳交换及其环境影响因子[J]. 应用生态学报, 2013, 24(11): 3057 - 3064.
TANG Xiang, CHEN Wenjing, LI Chunyi, et al. Net carbon exchange and its environmental affecting factors in a forest plantation in Badaling, Beijing of China[J]. Chinese Journal of Applied Ecology, 2013, 24(11): 3057 - 3064. (in Chinese)

[26] ICHII K, UEYAMA M, KONDO M, et al. New data-driven estimation of terrestrial CO_2 fluxes in asia using a standardized database of eddy covariance measurements, remote sensing data, and support vector regression[J]. Journal of Geophysical Research-Biogeosciences, 2017, 122(4): 767 - 795.

[27] 姜海梅, 张德广, 王若静, 等. 不同生态系统呼吸模型在半干旱草原生长季碳循环研究中的比较及应用[J]. 北京大学学报(自然科学版), 2018, 54(3): 593 - 604.
JIANG Haimei, ZHANG Deguang, WANG Ruojing, et al. Comparison of different ecosystem respiration models and its application in carbon cycle research over semi-arid grassland during growing season[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54(3): 593 - 604. (in Chinese)