引 用 格 式：Zhang Kun，Liu Naiwen，Gao Shuai，*et al*.Data-Driven Estimation of Gross Primary Production[J].Remote Sensing Technology and Application，2020，35(4)：943-949.[张坤，刘乃文，高帅，等.数据驱动的植被总初级生产力估算方法研究[J].遥感技术与应用，2020，35(4)：943-949.]
doi：10.11873/j.issn.1004-0323.2020.4.0943

数据驱动的植被总初级生产力估算方法研究

张 坤[1]，刘乃文[2]，高 帅[3]，赵书慧[1]
（1.山东师范大学信息科学与工程学院，山东 济南 250358；
2.山东管理学院信息工程学院省高校重点实验室，山东 济南 250357；
3.中国科学院遥感与数字地球研究所 遥感科学国家重点实验室，北京 100101）

摘要：植被总初级生产力（Gross Primary Production，GPP）是指在单位时间和单位面积上，绿色植物通过光合作用固定二氧化碳所产生的全部有机物同化量，对GPP的准确估算有助于碳循环的研究。为了提高GPP的估算精度，将机器学习技术与遥感技术相结合，首先利用GEE平台下的遥感数据以及中国陆地生态系统通量观测研究网络的通量塔实测GPP数据，建立数据集。然后使用随机森林作为估算模型，建模后根据数据特点对模型调参。最后获得模型的预测结果，决定系数R^2为0.87，均方根误差RMSE的值为1.132 $gC\cdot m^{-2}\cdot d^{-1}$。这说明随机森林模型可以较为精确地估算GPP。结果发现，以大数据以及人工智能为代表的计算机技术飞速发展，将为遥感技术注入新的活力，使遥感技术走向更加成熟的发展应用阶段。

关 键 词：随机森林模型；碳循环；GPP；大数据；GEE
中图分类号：TP701 **文献标志码**：A **文章编号**：1004-0323(2020)04-0943-07

1 引 言

随着全球工业的持续发展，化石燃料的使用量越来越大，在这个过程中释放了大量二氧化碳，引起了温室效应等一系列全球气候问题[1]，陆地生态系统可以通过碳循环影响全球气候变化[2]，因此全球陆地生态系统碳循环研究受到人们的广泛重视[3]，成为目前全球变化研究领域的热点[4]。而陆地生态系统碳循环研究中一个重要的指标就是GPP，GPP是指在单位时间和单位面积上，绿色植物通过光合作用固定二氧化碳所产生的全部有机物同化量[5]。它代表了所有进入陆地生态系统的碳和能量，因此，GPP的准确估算具有非常重要的意义。

人类研究植被生产力及其地理分布甚至可以追溯到公元前322年[5]。1882年，德国的林学家Ebermayer通过野外测量，第一次对全世界植被的含碳量进行了估计[6]。随着技术的发展逐渐出现了森林资源清查法、涡动通量观测法等观测方法，这些方法需要依靠在地面建立的调查站点，成本很高，虽然可以对小范围的GPP进行精确的计算，却无法对大空间尺度的GPP进行测算。

遥感技术的迅猛发展发展，为解决这一问题提供了基础，基于遥感数据的GPP过程模型成为估算大空间尺度GPP较为精准的方法。比如MOD17A3H的GPP数据就是利用BIOME-BGC模型与光能利用率模型得到的[7]。光能利用率模型是基于太阳辐射的利用率来估算植被光合作用的固碳量，该模型较为简单，所需的参数可以通过遥感技术大范围获取，因此可以获得大空间尺度和长时间序列的

收稿日期：2018-06-30；**修订日期**：2020-06-06
基金项目：国家重点研发计划项目(2017YFA0603004)，国家自然科学基金项目(41730107)，中国科学院百人计划项目(Y6YR0700QM)，高分项目(30－Y20A34－9010－15/17)。
作者简介：张 坤(1993—)，男，山东烟台人，硕士研究生，主要从事机器学习与空间数据挖掘研究。E-mail：329937012@qq.com
通讯作者：刘乃文(1971—)，男，山东济南人，教授，主要从事网络资源管理、计算机网络及应用等研究。E-mail：sdnwliu@126.com

GPP估算结果。然而受到水分温度等具体环境胁迫因素影响,不同植被功能类型的光能利用率模型也存在较大的时空差异,对估算精度造成影响。

随着大数据技术的发展,利用准确的地面观测资料,例如涡动通量塔的观测数据,与长时间大范围的遥感观测数据,进行碳循环研究已成为可能[8-9]。许多机构开发了工具来促进地理空间数据的大规模处理,比如谷歌地球引擎(Google Earth Engine,GEE)就存储了国际上主要的卫星遥感数据集,为遥感大数据应用提供了平台支持。而机器学习技术能够结合大数据云平台发现空间观测数据与GPP相关观测结果的可靠规律[10],利用这些规律对GPP进行预测和估算,从而形成一种数据驱动的观测新方法。本研究试图使用随机森林模型,利用通量观测数据和GEE平台,对GPP进行估算。

2 研究区与数据

2.1 研究区概况

研究区为长白山等8个站点及周围区域。8个地点均具有较强的代表性。根据IGBP的全球植被分类方案,站点的植被类型等信息如表1所示。

表1 研究区信息

Table 1 Research area information

站点名称	纬度/°N	经度/°E	植被类型
长白山	42.403	128.096	混交林
千烟洲	26.733	115.067	木本热带稀树草原
鼎湖山	23.167	112.530	常绿阔叶林
西双版纳	21.950	101.200	常绿阔叶林
锡林格勒	44.130	116.320	草地
禹城	36.833	116.567	农用地
拉萨当雄站	30.410	91.080	草地
海北站	37.660	101.330	草地

长白山站位于长白山自然保护区,属于温带大陆性气候,春季干旱多风,夏季炎热多雨,冬季干燥寒冷。

千烟洲人工林通量观测站地处江西省吉安市泰和灌溪镇,属于亚热带季风气候,夏季高温多雨,冬季温和少雨,四季分明。

鼎湖山通量观测站位于广东省肇庆市境东北部。该站点属南亚热带季风湿润气候,日照长,终年温暖。

西双版纳热带雨林通量观测站处于云南省西双版纳傣族自治州勐腊县,属于热带季风气候,一年中有雾凉季、干热季、湿热季之分,终年无霜。

锡林郭勒温性典型草原通量观测站位于内蒙古自治区锡林郭勒盟白音锡勒牧场,属于大陆性温带半干旱草原气候,冬春寒冷干燥,夏秋温暖湿润,受季风影响,具有明显的雨热同期特征。

禹城农田通量观测站,位于山东省禹城市西南,属于暖温带半湿润季风气候区,雨热同期,利于农业生产。

当雄高寒草甸碳通量观测站,位于西藏当雄县草原站,代表了藏北高原中部地区高寒草甸向高寒草原过渡的草原化草甸类型,属于高原季风气候。

海北高寒草甸生态系统通量观测站位于青藏高原,属于高原大陆性气候,海拔高气温极低,无明显四季之分,仅冷暖季之别,干湿季分明。

2.2 GEE平台

美国国家航空航天局、欧洲空间局等全球多个政府机构费提供了海量的遥感数据,并开发了相应的工具来进行处理。但是使用这些工具需要很强的计算机专业性,因此对一些科研人员来说灵活使用这些数据还有一些困难。

Google为了解决这个问题,推出了Google Earth Engine平台,用户只需要通过网络访问API接口或使用基于Web的交互式开发环境(IDE)就可以轻松访问和处理大型地理空间数据集,这些操作由Google的大型服务器进行处理,因此实现快速模型设计和结果可视化[11-12]。GEE中包含了200多个数据集,数据库每天增加包含环境、社会等方面超过5PB的数据,有着广泛的应用场景[13-15]。本文的遥感数据获取及预处理均在GEE平台完成。

2.3 数据

中分辨率成像光谱仪(MODIS)是NASA地球观测卫星上的重要传感器[16],自1999年以来每天都在收集地球图像,其数据在积雪面积、湖冰、水灾以及火灾等方面得到了广泛应用[17-19]。本文使用的EVI/NDVI、温度、土地覆盖类型等数据均为MODIS产品。其中EVI/NDVI数据为MCD43A4[20],温度数据为MOD11A2,土地覆盖类型数据为MCD12Q1,它们的时间分辨率均为8天,空间分辨率均为500 m。

统计降雨量时使用了PERSIANN-CDR[21]。PERSIANN-CDR是使用人工神经网络算法生成的一种全球日降水量产品,空间分辨率为0.25°,涵盖了从1993至今的全球降水数据。

GPP数据由中国陆地生态系统通量观测研究

网络(China-FLUX)提供。China-FLUX创建于2001年,它为中国碳循环领域的科学研究提供了重要的基础数据,推动了中国通量观测研究事业的发展[22],引起了国内外的广泛关注[23-26]。

3　方　法

3.1　实验方法

将通量塔实测GPP数据作为真值,从Google Earth Engine获取通量塔站点及周围3个MODIS象元的EVI、NDVI、温度、降水数据平均值作为影响因子,建立随机森林模型。流程如图1所示。

首先收集中国通量观测网2003年到2007年的实测GPP数据,然后从Google Earth Engine平台获取到相同时间段的遥感数据,将两者组成数据集,剔除掉异常数据后,将前三年的数据作为训练集,剩下的数据作为测试集。

研究使用机器学习中随机森林算法,首先使用训练集对模型进行训练,并根据数据特点对模型参数进行调整,选出最优的模型参数后对测试集进行预测,然后将预测结果与通量观测值进行对比,分析预测精度。最后将预测结果与MODIS数据进行对比,分析预测结果的可靠性。

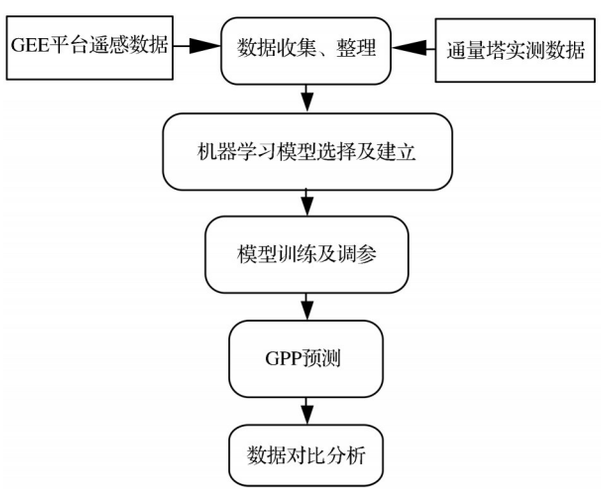


图1　流程图

Fig.1　Flow chart

3.2　随机森林

随机森林有很好的抗噪声能力,并且原理简单,易于实现,因此有着广泛的应用[27]。随机森林算法顾名思义,首先它会根据训练数据生成若干决策树,就像一片森林,当随机森林对一个对象进行分类时,森林中的每棵决策树都对这个对象进行判断,作出分类决定,按多棵决策树投票决定最终结果输出;对于回归问题,由多棵树预测值的均值决定最终预测结果[28]。

4　结果与讨论

4.1　影响因子重要性和模型参数

在使用训练组数据对模型进行训练后,各因素对结果的影响如图2所示。

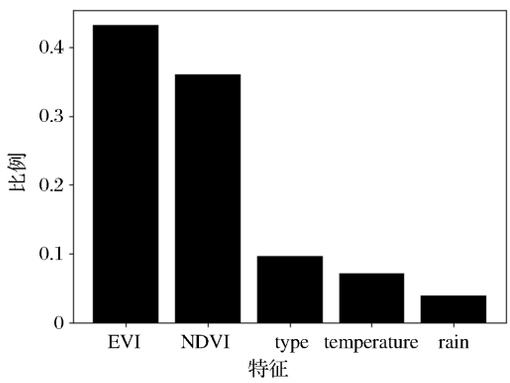


图2　对结果的影响力

Fig.2　Influence on results

图中可以看出对GPP影响最大的因素是EVI,这符合我们的预期,因为影响光合作用强弱最大的因素就是光照。后期试验还表明,当不考虑植被类型数据时,降水和温度的重要性将大幅提高,这也符合预期,因为植被类型的差异主要是由温度和降水导致的。

通过对模型调参可以获得更好的实验结果。每种机器学习模型的参数都不尽相同,随机森林模型参数主要有决策树最大深度(max_depth)、最大特征数(max_features)和最大迭代次数(n_estimators)。

本文模型的准确率与最大深度(max_depth)、最大特征数(max_features)关系如图3所示。由于数据量不算庞大,因此本模型不限制这两个参数。

本模型的另一个重要的参数是弱学习器的最大迭代次数(n_estimators),预测准确率与最大迭代次数关系如图4所示,采用10折交叉验证的方法,分别计算模型第10、20、…、100次的迭代后的均方根误差RMSE,经过计算,得到得到的最优迭代次数为80,RMSE的值为2.62,如表2所示。

4.2　预测结果

调参完成后,将测试组数据导入模型中进行预测,本文将通量塔数据作为实测数据,与预测数据进行对比,验证其精度。所预测的GPP结果与通量塔实测值对比如图5所示,经计算,预测结果的R^2为

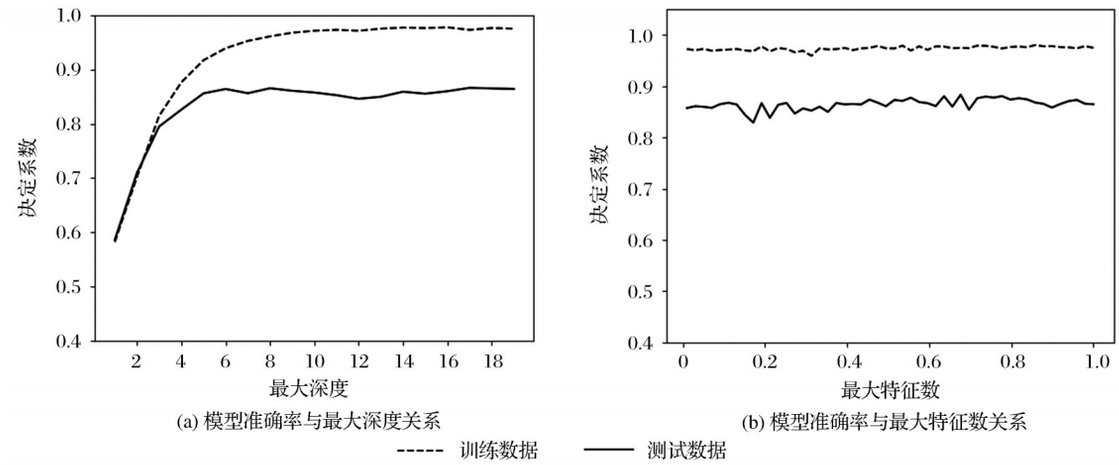


图 3 参数与准确率关系

Fig.3 Relationship between parameters and accuracy

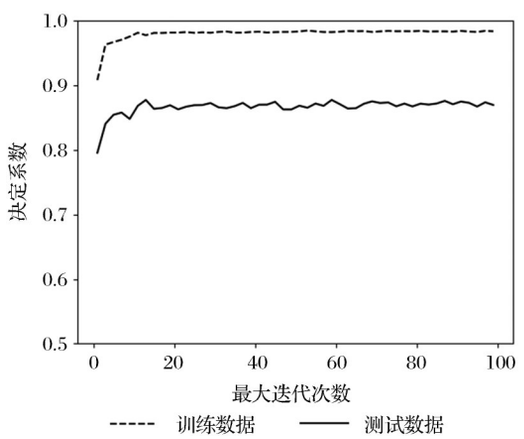


图 4 最大迭代次数与决定系数关系

Fig.4 The relationship between n_estimators and R^2

表 2 最大迭代次数调参表

Table 2 N_estimators adjustable parameter table

最大迭代次数	RMSE	排名
10	3.06	10
20	2.82	9
30	2.73	8
40	2.70	7
50	2.66	6
60	2.64	4
70	2.63	2
80	2.62	1
90	2.64	3
100	2.64	5

0.87，RMSE 为 1.132 gC·m^{-2}·d^{-1}，表明模型的模拟结果准确度较高。

4.3 分析对比讨论

为了验证模型预测的可靠性，将 MODIS 的 GPP 产品 MOD17A3H 与预测结果进行对比。MOD17A3H 是 MODIS 陆地 4 级标准数据产品，基于 BIOME-BGC 模型与光能利用率模型建立，准确度较高，在国际上具有很高的认可度，已在全球 GPP 与碳循环研究中得到广泛应用。

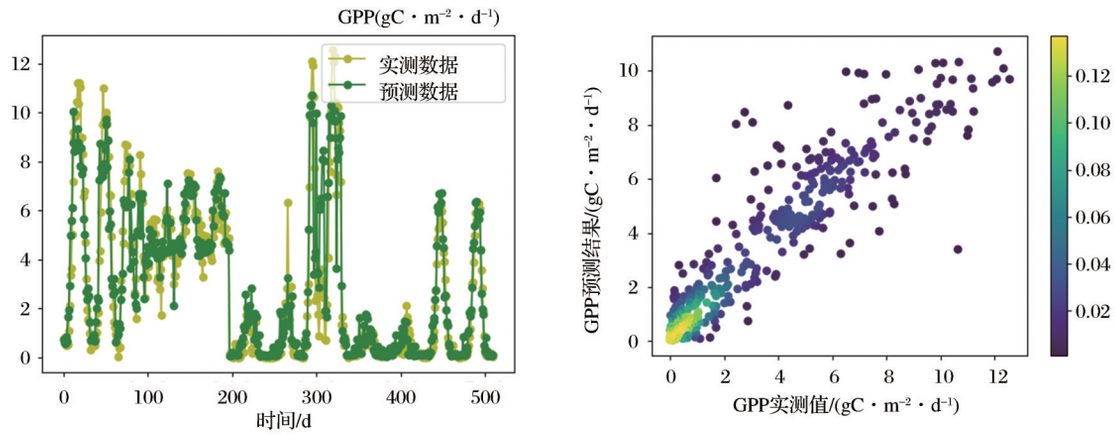
将 2006～2007 年 MOD17A3H 数据与模型预测数据进行对比，结果如图 6 所示，决定系数 R^2 为 0.66，RMSE 为 1.92 gC·m^{-2}·d^{-1}。由表 3 可知，模型基于通量塔测量数据学习而得到，因此模型预测值跟通量塔实测数据相关性很强，具有很高的预测准确度。同时，虽然 MOD17 数据的估算方法与本研究完全不同，但是通过对比可以看出随机森林模型预测值与 MOD 产品仍然具有很强的相关性，误差较小。

5 结 语

随着近年来云计算和大数据的发展，遥感数据的获取变得越来越方便，更多的科研人员可以将遥感数据应用到自己的研究工作当中，有利于遥感技术与多学科交叉发展。

本文对遥感技术与计算机技术的融合应用进行探索，将随机森林算法和 Google Earth Engine 平台应用在 GPP 的估算研究中，利用机器学习算法在回归预测中所具备的优势和 GEE 平台中的海量遥感数据，建立 GPP 估算模型，形成了一种全新的基于数据驱动的 GPP 估算方法。在对计算结果进行分析的基础上，将最终预测结果与 MODIS 数据进行对比，说明了本研究所建模型具有很高的精确度和可靠性。

但是以上实验所用到的通量数据只是从中国的通量塔站点获取，能否将模型进行更大范围的应



(a) GPP实测与预测值　　(b) GPP实测与预测值散点图

图 5　预测值与实测值对比

Fig.5　Comparison chart of predicted and measured values

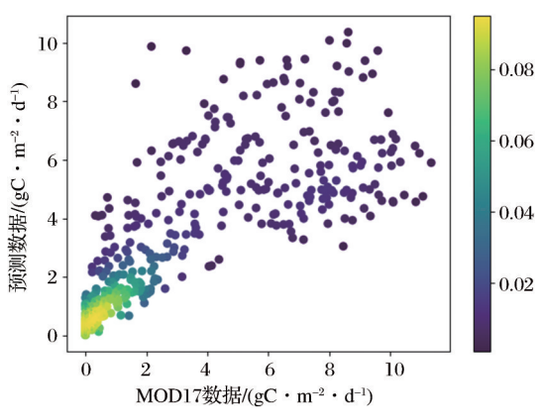


图 6　MODIS 数据与预测数据对比图

Fig.6　Comparison between MODIS data and measured values

表 3　模型效果对比表

Table 3　Model effect comparison

	MOD17A3	通量测量数据
R^2	0.66	0. 87
RMSE	1.92	1.132

用,还需要更多的实验验证。另外,可在以后的研究工作中进一步拓展,如增加更为多样的遥感数据[29],尝试更多的机器学习算法,以期取得更好的结果,实现对更多碳源汇数据的准确估算。

参考文献(References):

[1] Solomon S, Qin D, Manning M, *et al*. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to The Fourth Assessment Report of The Intergovernmental Panel on Climate Change. Summary for Policymakers[J]. Intergovernmental Panel on Climate Change Climate Change, 2007, 18(2):95-123.

[2] Bonan G B. Forests and Climate Change: Forcings, Feedbacks, and The Climate Benefits of Forests[J]. Science, 2008, 320(5882): 1444-1449.

[3] Cannell M G R, Milne R, Hargreaves K J, *et al*. National Inventories of Terrestrial Carbon Sources and Sinks: The U. K. Experience[J]. Climatic Change, 1999, 42(3):505-530.

[4] Yu Guirui.Global Change and Terrestrial Ecosystems Carbon Cycle and Carbon Accumulation[M]. Beijing: Meteorology Press, 2003[于贵瑞. 全球变化与陆地生态系统碳循环和碳蓄积[M]. 北京:气象出版社, 2003.]

[5] Lieth H. Primary Production: Terrestrial Ecosystems[J]. Human Ecology, 1973, 1(4):303-332.

[6] Li Wenhua, Zhou Xingmin. Ecological System and Optimal Utilization Mode of Qinghai-Tibet Plateau[M]. Guangzhou: Guangdong Science and Technology Press, 1998.[李文华, 周兴民. 青藏高原生态系统及优化利用模式[M]. 广州:广东科技出版社, 1998.]

[7] Mao J, Wang B, Dai Y, *et al*. Improvements of A Dynamic Global Vegetation Model and Simulations of Carbon and Water at An Upland-Oak Forest[J]. Advances in Atmospheric Sciences, 2007, 24(2): 311-322.

[8] Wang J, Wu C, Zhang C, *et al*. Improved Modeling of Gross Primary Productivity (GPP) by Better Representation of Plant Phenological Indicators from Remote Sensing Using A Process Model[J]. Ecological Indicators, 2018, 88: 332-340.

[9] Wu C, Peng D, Soudani K, *et al*. Land Surface Phenology Derived from Normalized Difference Vegetation Index (NDVI) at Global FLUXNET Sites[J]. Agricultural and Forest Meteorology, 2017, 233(Complete):171-182.

[10] Tramontana G, Ichii K, Camps-Valls G, *et al*. Uncertainty Analysis of Gross Primary Production Upscaling Using Random Forests, Remote Sensing and Eddy Covariance Data[J]. Remote Sensing of Environment, 2015, 168: 360-373.

[11] Chen B, Xiao X, Li X, *et al*. A Mangrove Forest Map of China in 2015: Analysis of Time Series Landsat 7/8 and Sentinel-1A Imagery in Google Earth Engine Cloud Computing Platform[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2017, 131: 104-120.

[12] Alonso A, Muñoz-carpena R, Kennedy R E, *et al*. Wetland Landscape Spatio-temporal Degradation Dynamics Using The New Google Earth Engine Cloud-based Platform: Opportunities for Non-Specialists in Remote Sensing[J]. Transactions of The ASABE, 2016, 59(5): 1331-1342.

[13] Gorelick N, Hancher M, Dixon M, *et al*. Google Earth Engine: Planetary-scale Geospatial Analysis for Everyone[J]. Remote Sensing of Environment, 2017,202: 18-27.

[14] Dong J, Xiao X, Menarguez M A, *et al*. Mapping Paddy Rice Planting Area in Northeastern Asia with Landsat 8 Images, Phenology-based Algorithm and Google Earth Engine[J]. Remote Sensing of Environment, 2016,185:142-154..

[15] Tian H F, Meng M, Wu M Q, *et al*. Mapping Spring Canola and Spring Wheat Using Radarsat-2 and Landsat 8 Images with Google Earth Engine[J]. Current Science, 2019, 116: 291-298.

[16] Cheng Yanpei, Zhang Fawang, Dong Hua, *et al*. Dynamic Monitoring of Water Bodies in Central Asia based on Modis Satellite Data [J]. Hydrogeological Engineering Geology, 2010, 37(5): 33-37.[程彦培, 张发旺, 董华, 等. 基于MODIS卫星数据的中亚地区水体动态监测研究[J]. 水文地质工程地质, 2010, 37(5):33-37.]

[17] Giglio L, Van d W G R, Randerson J T, *et al*. Global Estimation of Burned Area Using MODIS Active Fire Observations[J]. Atmospheric Chemistry and Physics Discussions, 2005, 5(6):11091-11141.

[18] Reed B, Budde M, Spencer P, *et al*. Integration of MODIS-derived Metrics to Assess Interannual Variability in Snowpack, Lake Ice, and NDVI in Southwest Alaska[J]. Remote Sensing of Environment, 2009, 113(7):1443-1452.

[19] Maurer E P, Rhoads J D, Dubayah R O, *et al*. Evaluation of The Snow-covered Area Data Product from MODIS[J]. Hydrological Processes, 2003, 17: 59-71.

[20] Huete A, Didan K, Miura T, *et al*. Overview of The Radiometric and Biophysical Performance of The MODIS Vegetation Indices[J]. Remote Sensing of Environment, 2002, 83 (1-2):195-213.

[21] Ashouri H, Hsu K L, Sorooshian S, *et al*. PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies[J]. Bulletin of the American Meteorological Society, 2015, 96 (1):69-83.

[22] Yu Guirui, Sun Xiaomin. The Principle and Method of Flux Observation of Terrestrial Ecosystem[M]. Beijing: Advanced Education Publishing House,2006.[于贵瑞, 孙晓敏. 陆地生态系统通量观测的原理与方法[M]. 北京:高等教育出版社, 2006.]

[23] Leuning R, Yu G R. Carbon Exchange Research in China-FLUX[J]. Agricultural & Forest Meteorology, 2006, 137(3-4):0-124.

[24] Baldocchi D.Breathing of The Terrestrial Biosphere: Lessons Learned from A Global Network of Carbon Dioxide Flux Measurement Systems[J]. Australian Journal of Botany, 2008, 56 (1):1-26.

[25] Doherty S J, Bojinski S, Henderson-Sellers A, *et al*. Lessons Learned from IPCC AR4: Scientific Developments Needed to Understand, Predict, and Respond to Climate Change[J]. Bulletin of the American Meteorological Society, 2009, 90(4): 497-514.

[26] Stoy P C, Mauder M, Foken T, *et al*. A Data-driven Analysis of Energy Balance Closure Across FLUXNET Research Sites: The Role of Landscape Scale Heterogeneity[J]. Agricultural and Forest Meteorology, 2013, 171: 137-152.

[27] Liu X, Guanter L, Liu L, *et al*. Downscaling of Solar-induced Chlorophyll Fluorescence from Canopy Level to Photosystem Level Using A Random Forest Model[J]. Remote Sensing of Environment, 2019, 231: 110772. doi: 10.1016/j.rse.2018.05.035.

[28] Yang Siqi, Zhao Lihua. Application of Random Forest Algorithm in Urban Air Quality Prediction[J]. Statistics and Decision Making,2017(20):83-86.[杨思琪, 赵丽华. 随机森林算法在城市空气质量预测中的应用[J]. 统计与决策, 2017 (20):84-87.]

[29] Chen Y, Shen W, Gao S, *et al*. Estimating Deciduous Broadleaf Forest Gross Primary Productivity by Remote Sensing Data Using A Random Forest Regression Model[J]. Journal of Applied Remote Sensing, 2019, 13(3): 038502.

Data-Driven Estimation of Gross Primary Production

Zhang Kun[1], Liu Naiwen[2], Gao Shuai[3], Zhao Shuhui[1]

(1.*School of Information Science & Engineering, Shandong Normal University, Jinan* 250358, *China*;
2.*Provincial Key Laboratory of Information Technology, School of Information Engineering, Shandong Management University, Jinan* 250357, *China*;
3.*The State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract: Gross Primary Production (GPP) of vegetation refers to the assimilation of all organic matter produced by green plants through photosynthesis and fixed carbon dioxide per unit time and unit area. Accurate estimation of GPP is helpful for the study of carbon cycle. In order to improve the estimation accuracy of GPP, this study combines machine learning technology and remote sensing technology. First, the remote sensing data under the GEE platform and the flux tower measurement data of the China Terrestrial Ecosystem Flux Observation Research Network are used to establish a data set. Then use random forest as the estimation model, and adjust the model according to the data characteristics after modeling. Finally, the prediction results of the model are obtained, the determination coefficient R^2 is 0.87, and the root mean square error RMSE is 1.132 $gC \cdot m^{-2} \cdot d^{-1}$. This shows that the random forest model can estimate GPP more accurately.From the results of this study, we can see that the rapid development of computer technology represented by big data and artificial intelligence will inject new vitality into remote sensing technology and make remote sensing technology enter a more mature stage of development and application.

Key words: Random forest regression; The carbon cycle; GPP; Big data; GEE platform