Comparison of different machine learning method for GPP estimation using remote sensing data

Comparison of different machine learning method for GPP estimation using remote sensing data

Kun Zhang[1], Naiwen Liu[2], Yue Chen[3] and Shuai Gao[4,*]

[1]School of Information Science & Engineering, Shandong Normal University, Jinan, China
[2]Key Laboratory of TCM Data Cloud Service in Universities of Shandong，Shandong Management University, Jinan, China
[3]School of Earth Sciences and Resources ,China University of Geosciences, Beijing, China
[4]The State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China

*Corresponding author e-mail: gaoshuai@radi.ac.cn

Abstract. This paper selects eight sites with typical characteristics in China (Changbai Mountain, Qianyanzhou, Dinghushan, etc.). Based on remote sensing data acquired from the Google Earth Engine (GEE) big data cloud platform, four machine learning models were established to estimate GPP. Firstly, remote sensing data such as EVI, NDVI, precipitation and temperature were downloaded by GEE, and the flux tower data of 8 sites of China-FLUX was obtained. Secondly, the machine learning algorithm is used to establish the connection between the two types of data. Finally, the machine learning model is used to predict the test group data, and the results are evaluated by using R2, RMSE and other related precision indicators, and the accuracy of the MODIS data is compared. Studies have shown that machine learning models can obtain more accurate GPP predictions.

1. Introduction

GPP（Gross Primary Productivity）refers to the total amount of organic matter produced by photosynthesis in units per unit of time per plant area. It determines the amount of initial energy and material entering the terrestrial ecosystem, and also represents the amount of carbon dioxide fixed by the plant through photosynthesis, so GPP is a very important parameter in the study of terrestrial system carbon cycle and global climate change. Accurate estimation of GPP is important for simulating the carbon cycle and mastering the change of carbon dioxide in the atmosphere[1].

In recent years, technologies such as big data and artificial intelligence have been continuously developed. Data-based machine learning is an important aspect of modern intelligent technology. It is applied in various research fields. Machine learning models look for patterns from observational data (samples) and use these rules to predict future or unobservable data. Applying machine learning algorithms to GPP estimation can make full use of massive remote sensing data to provide a method for GPP's large spatial scale, long-term sequence and high-precision estimation.

In order to establish a high-precision GPP estimation model, this paper uses the Google Earth Engine to obtain remote sensing data sets, combines the ground flux tower to measure GPP data, and uses various machine learning algorithms to carry out modeling research on GPP estimation. A

comparative analysis of various machine learning algorithms in terms of estimation accuracy, etc., attempts to find a machine learning algorithm that is most suitable for GPP estimation, and provides a more accurate model for GPP's global estimation.

2. Research area and data

2.1. Research area

The research areas of this paper are eight sites and surrounding areas such as Changbai Mountain, Qianyanzhou, Dinghushan, and Xilingler, as shown in Figure 1. These eight regions all have carbon flux towers that provide measured carbon flux data, and the altitudes, latitudes, and land types of the eight locations are different and therefore highly representative. According to the global vegetation classification scheme of IGBP in MOD12Q1[2] data, the vegetation types of the sites are shown in Table 1.

Figure 1. Location map of the study area.

Table 1. Research area information.

Site name	Vegetation Types	Climate type	Area
Changbai Mountain	Mixed forest	Temperate continental climate	Changbai Mountain Nature Reserve
Qianyanzhou	Woody savanna	Subtropical monsoon climate	Qianyanzhou Red Soil Hilly Area
Dinghushan	Evergreen broad-leaved forest	Subtropical monsoon humid climate	Dinghushan Nature Reserve, Zhaoqing City
Xishuangbanna	Evergreen broad-leaved forest	tropical monsoon climate	Mengla County National Nature Reserve
Xilin Gol	Grassland	Continental temperate semi-arid grassland climate	Xilin Gol League Baiyin Xile Ranch
Yucheng	Agricultural land	Warm temperate semi-humid monsoon climate	Southwest of Yucheng
Lhasa Dangxiong	Grassland	Plateau monsoon climate	Central Tibet Plateau
Haibei	Grassland	Plateau continental climate	Qinghai-Tibet Plateau

2.2. Data

The remote sensing data such as EVI/NDVI, temperature and land cover type used in this paper are all MODIS products. The EVI/NDVI data is MCD43A4[3], the temperature data is MOD11A2, and the land cover type data is MCD12Q1. Their time resolution is 8 days and the spatial resolution is 500 meters. The precipitation data used herein is PERSIANN-CDR [4]. PERSIANN-CDR is a global daily precipitation product generated by artificial neural network algorithm using GridSat-B1 infrared data

with a spatial resolution of 0.25° . The GPP data used in this paper is provided by the China Terrestrial Ecosystem Flux Observation and Research Network (China-FLUX)[5]. This paper uses the GPP data of China-FLUX's 8 ecosystem flux sites 2003-2006.

Table 2. Prediction model input factor.

Variable name	EVI	NDVI	Temperature	Precipitation	Vegetation Types
unit	—	—	（℃）	（mm）	—

3. Methods

In this paper, the measured data of the flux tower is taken as the true value, and the remote sensing data such as temperature, precipitation and EVI around the flux tower are used as the influencing factors to establish a machine learning model. The model was trained using data from eight sites 2003-2005 and then used to estimate GPP based on data from 2006 . The model estimation accuracy is analyzed based on the flux tower data. The specific process is shown in Figure 2.

Vegetation photosynthesis is greatly affected by temperature, precipitation and light radiation. Therefore, remote sensing data such as EVI, NDVI, temperature and precipitation are selected as the model impact factors. Firstly, the Google Earth Engine is used to obtain the remote sensing data of the flux tower site area. Then combined with the measured GPP data of China Flux Observing Network, the data set is composed. Due to the influence of weather and other factors, there is no effective remote sensing data on individual dates. We will filter and organize the datasets and eliminate the anomalous data. According to the needs of some machine learning models, the parameters need to be normalized.

In this study, the data sets are divided into training group and test group. The training group data is used to build the machine learning model and trained one by one. The parameters of the four machine models were adjusted using R^2 as the evaluation index. Through the method of 10-fold cross-validation, the optimal model parameters are selected to obtain better regression results. Finally, the built-in model is used to predict the test group data. By comparing the true values of the test group data, the prediction accuracy of the four machine learning models is compared and analyzed.

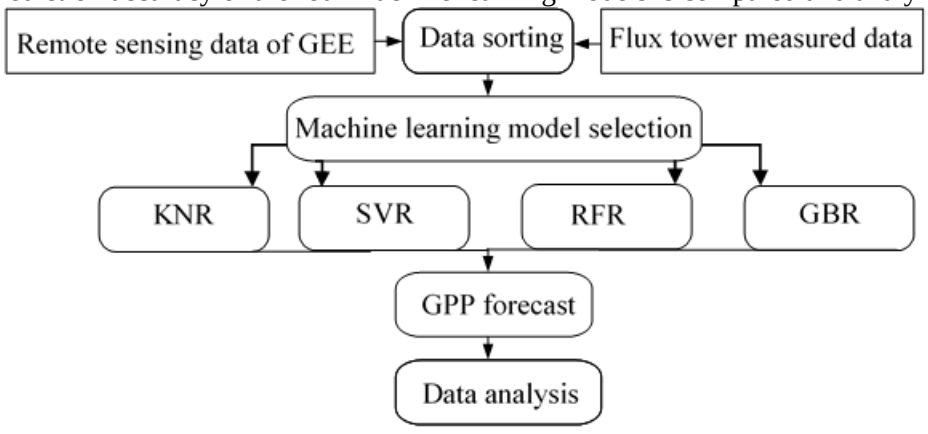

Figure 2. Technical flow chart.

4. Results

4.1. Model establishment

The prediction of GPP belongs to the regression problem of supervised learning in machine learning. Since the correlation between GPP and remote sensing data is complex and nonlinearly related, linear regression models cannot be selected. We selected Support Vector Machine Regression model (SVR), K Nearest Neighbor Regression model (KNR), Random Forest Regression model (RFR), Gradient BoostRegression Tree (GBR) for experiments and comparisons[6].

When using the training group data to build the model, the four models are each verified with ten-fold cross-validation. The effect is shown in Figure 3. It can be seen that the random forest model has higher accuracy.

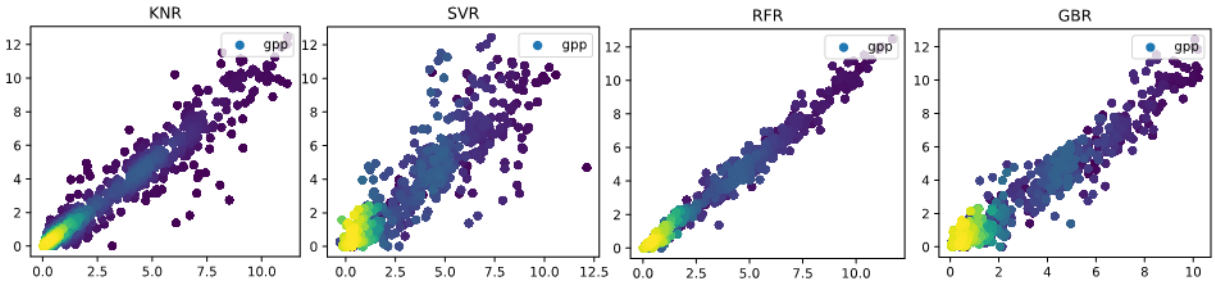

Figure 3. Ten fold cross validation results.

4.2. Model prediction result

The 2006 GPP is predicted using the established four machine learning models. The measured values of the flux tower are compared with the predicted values of the model. The results are shown in Fig. 4. From the figure, all four models have high prediction accuracy.

We use the accuracy indicators such as R^2 and RMSE to evaluate the prediction results of the four models. The results are shown in Table 3. From the data, the best regression model is random forest, and the k-nearest neighbor regression model is very similar. The R^2 values of these four models all exceed 0.8, which indicates that machine learning is very suitable for GPP prediction.

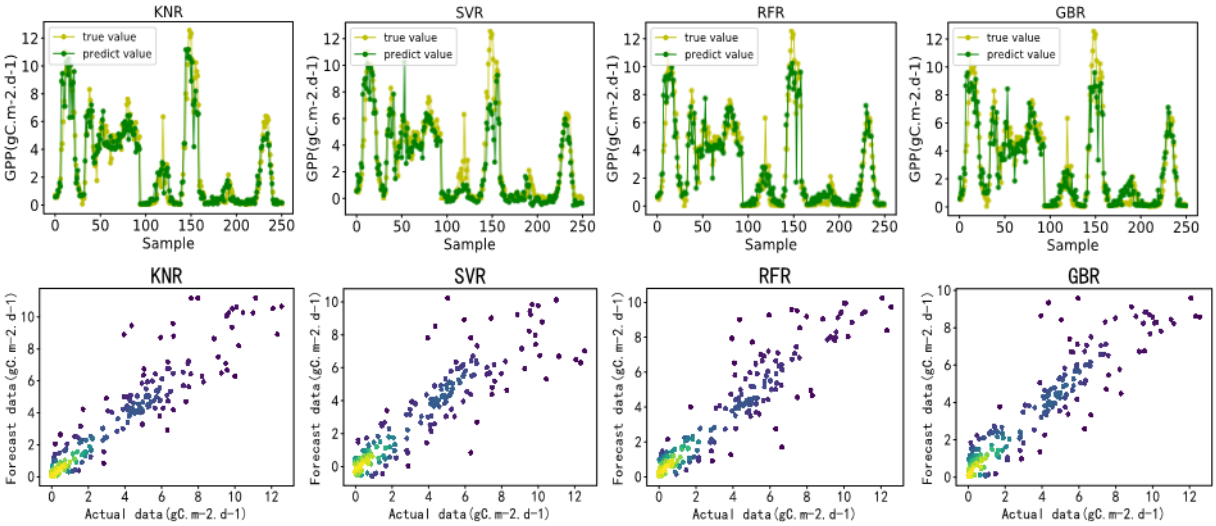

Figure 4. Model prediction result.

Table 3. Precision comparison.

Precision indicators	KNR	SVR	RFR	GBR
R^2	0.88	0.81	0.89	0.86
RMSE	1.08	1.39	1.07	1.17
MAE	0.67	0.87	0.68	0.76

4.3. Contrast with MODIS

In order to illustrate the performance of the model built in this study, it is also necessary to use a relatively accurate remote sensing ecological process model for comparative experiments. We use Modis' GPP product MOD17A3 for experimental demonstration.

MOD17A3 is a GPP estimation data based on the BIOME-BGC model (Biome Biogeochemical Model) and the light energy utilization model. It has high international recognition and has been widely used in global GPP and carbon cycle research[7].

The 2006 MOD17A3H data was compared to the measurement data of the flux tower. The results are shown in Figure 5. As can be seen from the figure, the R^2 between the MODIS data and the measured value is 0.44. The calculated RMSE is 2.37 $gC.m^{-2}.d^{-1}$. It can be seen from Table 4 that although the MODIS products have high accuracy, the experimental results of the machine learning model are significantly better than the Modis products from the comparison of the accuracy evaluation indexes such as R^2 and RMSE. This shows that the machine learning model we built is relatively good.

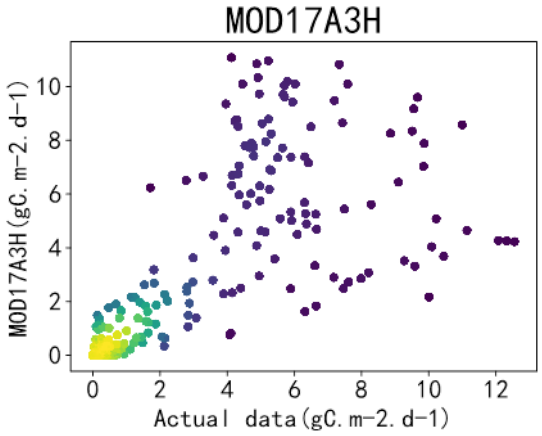

Figure 5. Technical flow chart.

Table 4. Precision comparison.

Precision indicators	MOD17A3H
R^2	0.44
RMSE	2.37
MAE	1.46

5. Conclusion

In this paper, the machine learning and GEE platform are applied in the estimation study of GPP. We analyzed the estimation accuracy of GPP for the four machine learning models and compared the final predictions with MODIS's GPP products. It can be seen from indicators such as R^2 and RMSE. The machine learning model works better than the traditional process model.

Using the advantages of machine learning algorithms in regression prediction and the characteristics of carbon cycle data, a new GPP estimation method can be formed. Because the observation time of China's terrestrial ecosystem flux observation research network is not long enough, the amount of data is still relatively small. In the future, combined with more measured data from global flux sites and more remote sensing data, the accuracy of machine learning for GPP estimation can be further improved, and data can be expanded in time and space.

Acknowledgments

This work was financially supported by National Key R&D Program (2017YFA0603004), Natural Science Fund Project (41730107), Chinese Academy of Sciences Hundred Talents Program (Y6YR0700QM) and High Score Project (30-Y20A34-9010-15/17) .

References

[1] Lin, Shangrong, L. I. Jing, and Q. Liu. "Overview on estimation accuracy of gross primary productivity with remote sensing methods." Journal of Remote Sensing (2018).

[2] Li, Xiaosong, and J. Zhang. "Derivation of the Green Vegetation Fraction of the Whole China from 2000 to 2010 from MODIS Data." Earth Interactions 20.8(2016).
[3] Huete, A., et al. "Overview of the radiometric and biophysical performance of the MODIS vegetation indices." Remote Sensing of Environment 83.1(2002):195-213.
[4] Ashouri, Hamed, et al. "PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies." Bulletin of the American Meteorological Society 96.1(2014):197-210.
[5] Guirui, Y. U., L. Zhang, and X. Sun. "Progresses and prospects of Chinese terrestrial ecosystem flux observation and research network (ChinaFLUX)." Progress in Geography 33.7(2014).
[6] Singh, Amanpreet, N. Thakur, and A. Sharma. "A review of supervised machine learning algorithms." International Conference on Computing for Sustainable Global Development IEEE, 2016.
[7] Plummer, S. "On validation of the MODIS gross primary production product." IEEE Transactions on Geoscience & Remote Sensing 44.7(2006):1936-1938.