Available online at www.sciencedirect.com
ScienceDirect
Remote Sensing of Environment 110 (2007) 109 – 122
Remote Sensing of Environment
www.elsevier.com/locate/rse

Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach

Feihua Yang [a,b,*], Kazuhito Ichii [b,c], Michael A. White [d], Hirofumi Hashimoto [b,e], Andrew R. Michaelis [b,e], Petr Votava [b,e], A-Xing Zhu [f,a], Alfredo Huete [g], Steven W. Running [h], Ramakrishna R. Nemani [b]

[a] *University of Wisconsin, Madison, USA*
[b] *NASA Ames Research Center, USA*
[c] *San Jose State University, USA*
[d] *Utah State University, USA*
[e] *California State University, Monterey Bay, USA*
[f] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, China*
[g] *University of Arizona, Tucson, USA*
[h] *University of Montana, Missoula, USA*

Received 19 October 2006; received in revised form 8 February 2007; accepted 10 February 2007

Abstract

Remote sensing is a potentially powerful technology with which to extrapolate eddy covariance-based gross primary production (GPP) to continental scales. In support of this concept, we used meteorological and flux data from the AmeriFlux network and Support Vector Machine (SVM), an inductive machine learning technique, to develop and apply a predictive GPP model for the conterminous U.S. In the following four-step process, we first trained the SVM to predict flux-based GPP from 33 AmeriFlux sites between 2000 and 2003 using three remotely-sensed variables (land surface temperature, enhanced vegetation index (EVI), and land cover) and one ground-measured variable (incident shortwave radiation). Second, we evaluated model performance by predicting GPP for 24 available AmeriFlux sites in 2004. In this independent evaluation, the SVM predicted GPP with a root mean squared error (RMSE) of 1.87 gC/m^2/day and an R^2 of 0.71. Based on annual total GPP at 15 AmeriFlux sites for which the number of 8-day averages in 2004 was no less than 67% (30 out of a possible 45), annual SVM GPP prediction error was 32.1% for non-forest ecosystems and 22.2% for forest ecosystems, while the standard Moderate Resolution Imaging Spectroradiometer GPP product (MOD17) had an error of 50.3% for non-forest ecosystems and 21.5% for forest ecosystems, suggesting that the regionally tuned SVM performed better than the standard global MOD17 GPP for non-forest ecosystems but had similar performance for forest ecosystems. The most important explanatory factor for GPP prediction was EVI, removal of which increased GPP RMSE by 0.85 gC/m^2/day in a cross-validation experiment. Third, using the SVM driven by remote sensing data including incident shortwave radiation, we predicted 2004 conterminous U.S. GPP and found that results were consistent with expected spatial and temporal patterns. Finally, as an illustration of SVM GPP for ecological applications, we estimated maximum light use efficiency (e_{max}), one of the most important factors for standard light use efficiency models, for the conterminous U.S. by integrating the 2004 SVM GPP with the MOD17 GPP algorithm. We found that e_{max} varied from ∼0.86 gC/MJ in grasslands to ∼1.56 gC/MJ in deciduous forests, while MOD17 e_{max} was 0.68 gC/MJ for grasslands and 1.16 gC/MJ for deciduous forests, suggesting that refinements of MOD17 e_{max} may be beneficial.

Keywords: Gross primary production; AmeriFlux; Support Vector Machines; Light use efficiency; Moderate Resolution Imaging Spectroradiometer (MODIS)

* Corresponding author. Department of Geography, University of Wisconsin-Madison, 426 Science Hall, 550 North Park Street, Madison, WI 53706, USA.
E-mail address: feihuayang@wisc.edu (F. Yang).

0034-4257/$ - see front matter
doi:10.1016/j.rse.2007.02.016

1. Introduction

Gross primary production (GPP), the integral of photosynthesis by all leaves, is a critical component in ecological systems. Carbon accumulated by ecosystem GPP and not used for plant growth and maintenance is returned to the atmosphere via respiration or disturbance (e.g. combustion), or is transported to other ecosystems through flow paths such as dissolved organic carbon. Autotrophic respiration consumes about half of GPP (Chapin et al., 2004); net primary production (NPP) is the residual. Quantitative estimates of the spatial and temporal distribution of GPP and NPP at regional to global scales are critical for the understanding of ecosystem response to increased atmospheric carbon dioxide (CO_2) level and are thus central to policy-relevant decisions (Metz et al., 2006).

Currently, GPP is estimated via process-based models and satellite-data based models. Process-based models such as BIOME-BGC (Thornton, 1998; White et al., 2000) can dynamically simulate vegetation physiology (e.g. rubisco activities and soil water stress), leading to potentially useful application at watershed scales (Zhu & Scott, 2001). However, process-based models are difficult to extend to large regions due to their complex structures and requirements for complete coverage of frequently poorly known land surface state variables such as specific leaf area and respiration coefficients. On the other hand, models based on or ingesting remote sensing data have two central advantages over purely process-based models: (1) satellite remote sensing offers broad spatial coverage and regular temporal sampling; and (2) requirements for spatial and temporal parameterization of vegetation physiological variables are reduced or eliminated. Remote sensing models are thus theoretically capable of accurately predicting actual carbon fluxes at regional to continental scales.

Methods using remote sensing data for carbon flux calculation are mostly based on the concept of light use efficiency (LUE). Monteith (1972) suggested that the NPP of well-watered and fertilized annual crop plants was linearly related to the absorbed photosynthetically active radiation (APAR). This formulation simplifies photosynthesis calculation over large region and is the basis of many remote sensing-based GPP algorithms such as the Moderate Resolution Imaging Spectroradiometer (MODIS) GPP/NPP algorithm (MOD17; Running et al., 2004), the Vegetation Photosynthesis Model (VPM; Xiao et al., 2005) and the Biosphere Model Integrating Eco-physiological and Mechanistic Approaches using Satellite Data (BEAMS; Sasai et al., 2005). In LUE-based GPP models, APAR (the product of photosynthetically active radiation (PAR) and the fraction of PAR absorbed by plant canopies, FPAR) is usually linearly converted to GPP using biome-specific maximum light use efficiency (e_{max}, gC/MJ) attenuated by temperature and water stress status. However, recent studies have suggested that this approach may lead to considerable errors in modeled GPP (Heinsch et al., 2006; Turner et al., 2003, 2005; Zhao et al., 2005). Likely causes of the error include uncertainties in PAR, FPAR, and the conversion from incident shortwave radiation to PAR. Perhaps most centrally, though, uncertainty in the spatiotemporal variation of e_{max} is a crucial limiting factor for LUE models.

While remote sensing has been used to extrapolate field-measured NPP since the late 20th century (Paruelo et al., 1997), the increasing availability of near-real time observations of water and carbon exchange from eddy covariance flux towers (e.g. AmeriFlux; Baldocchi et al., 2001) has sparked a renewed interest in extrapolative techniques, especially through the use of empirical modeling techniques using remote sensing explanatory variables. For example, Rahman et al. (2005) observed a strong correlation between across-site tower-GPP and enhanced vegetation index (EVI). Wylie et al. (2003) related coarse resolution normalized difference vegetation index (NDVI) to 14-day average daytime CO_2 fluxes in a sage–brush–steppe ecosystem. Xiao et al. (2005) integrated e_{max} derived from tower-based net ecosystem exchange (NEE) and PAR into the VPM model. Gilmanov et al. (2005) found that NDVI was statistically significantly correlated with tower-based GPP and ecosystem respiration leading to potential scaling-up of tower fluxes to larger areas. Drolet et al. (2005) reported strong correlations between MODIS-derived photochemical reflectance index and tower-based LUE. These studies established the potential of using statistical and machine learning methods to extrapolate tower-based GPP to a regional to continental scale.

The goal of this paper is to explore the application of Support Vector Machine (SVM) learning techniques for GPP prediction at a continental scale. To do so, we tuned and trained SVMs driven by ground measured and remotely sensed explanatory variables to predict AmeriFlux GPP, tested the SVMs using a withheld portion of the flux data, and applied the final model for GPP prediction over the conterminous U.S. In the following sections, we present: (1) a brief description of the SVM technique; (2) SVM tuning and training, including a description of the AmeriFlux GPP observations and the selection of explanatory variables; (3) results from independent testing of the SVMs; (4) a comparison of SVM GPP to MOD17 GPP; and (5) extrapolation of the SVMs to the conterminous U.S. Finally, to demonstrate the potential use of the SVM GPP for a broad modeling community, we present a method to estimate e_{max} for the conterminous U.S. by coupling the SVM GPP with the MOD17 GPP algorithm.

2. Methods

2.1. SVM for regression

Regression methods attempt to construct an approximate function which maps an input domain to a real valued output domain based on a set of data examples (Cristianini & Shawe-Taylor, 2000). Commonly used regression methods include conventional statistical methods, such as multiple regressions, and machine learning methods, such as neural network and SVM.

Multiple regression is a standard statistical method designed to predict the values of a target concept from two or more explanatory variables. It is conceptually simple but less suited for highly nonlinear problems, especially those outside a prescribed range of nonlinear approaches. Neural network is a computing system motivated by the function of a human brain (Haykin, 1998). It is widely used for regression due to its ability