Computers and Electronics in Agriculture 144 (2018) 232–240

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Generalizability of gene expression programming and random forest methodologies in estimating cropland and grassland leaf area index

Sepideh Karimi[a], Ali Ashraf Sadraddini[a,*], Amir Hossein Nazemi[a], Tongren Xu[b], Ahmad Fakheri Fard[a]

[a] Water Engineering Department, Faculty of Agriculture, University of Tabriz, Iran
[b] State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Keywords:
Meteorological data
Leaf area index
Gene expression programming
Random forest

ABSTRACT

Leaf Area Index (LAI) is a very important structural attribute of ecosystems which affects the energy, water and carbon exchanges between the land surface and atmosphere. Direct measurement of LAI is costly and time consuming so indirect measurement approaches have been developed for determining its magnitude. The present paper aimed at modeling LAI in cropland and grassland sites using the available meteorological data through two heuristic data driven techniques, namely, gene expression programming (GEP) and random forest (RF). Different data set organizations were designed using local (temporal) and external (spatial) norms to provide a thoroughgoing data scanning strategy. The results showed that the external GEP and RF models (EGEP and ERF) might be suitable approaches for modeling LAI by average scatter index (*SI*) values of 0.275 and 0.270 (for cropland) and 0.273 and 0.279 (for grassland) when compared to the local GEP and RF models with average *SI* values of 0.207 and 0.204 (cropland), and 0.249 and 0.204 (grassland), respectively. The presented methodology allowed the evaluation in each site of models developed (trained) using local patterns and the models developed using the exogenous data (patterns from ancillary sites).

1. Introduction

Leaf area index (LAI) is a dimensionless variable defined as the total one-sided area of photosynthetic tissues per unit ground surface area (Watson, 1947). LAI is an important structural characteristic of ecosystem as it influences the exchanges of water, energy, and carbon between the land surface and atmosphere (Sellers et al., 1988; Wulder et al., 1998; Sonnentag et al., 2007). It determines the size of the plant–atmosphere interface, and therefore plays a key role in the energy and mass exchanges between the canopy and the atmosphere (Weiss et al., 2004). Xu et al. (2014) showed that the inclusion of LAI in their variational data assimilation model improved the simulation of surface water and energy fluxes. Li et al. (2009) indicated that using LAI in Xinanjiang hydrologic modeling improved rainfall-runoff modeling. Coopersmith et al. (2014) and Chen et al. (2015) obtained a better soil moisture prediction by incorporating LAI in Integrated Biosphere Simulator (IBS) and HYDRUS-1D models.

The ground-based measurement methods have been developed to measure LAI accurately (Asner et al., 2003; Jonckheere et al., 2004; Qu et al., 2014). However, those methods can only obtain LAI at point scale during limited time periods due to their high cost and time consumption. Therefore, different models have been developed to acquire LAI over large spatial scales based on remotely sensed data.

Currently, there are mainly three kinds of methods for retrieving LAI from remotely sensed data, i.e., the empirical relationships, radiative transfer models, and heuristic data driven models. The empirical methods are used to link LAI with remotely sensed vegetation index (i.e. NDVI) or with reflectance data with regression equations (Combal et al., 2003), which are relatively simple and accurate. However, these methods are sensor dependent and site specific, and have a major drawback of local calibration need. The physical laws are used in radiative transfer models to explicitly describe associations between the vegetation properties and canopy spectra, and produce reasonable LAI at regional scale (Meroni et al., 2004; He et al., 2013). However, the radiative transfer models are usually complicated and time consuming (Jacquemoud et al., 2000). Thus, heuristic data driven techniques are used to estimate LAI at larger scales (Xiao et al., 2014; Liang et al., 2014). The GLASS-LAI product is produced via the reflectance data in the visible and infrared bands based on heuristic data driven techniques (Liang et al., 2014; Xiao et al., 2014).

In recent years, heuristic data driven techniques (e.g., gene expression programming (GEP) and random forest (RF)) have been

* Corresponding author.
E-mail address: sadraddini@tabrizu.ac.ir (A.A. Sadraddini).

https://doi.org/10.1016/j.compag.2017.12.007
Received 10 August 2017; Received in revised form 4 December 2017; Accepted 7 December 2017

utilized for modeling hydrological and eco-hydrological parameters. Genetic programming (GP), a generalization of genetic algorithm (GA) (Goldberg, 1989), was proposed by Koza (1992). It engages a "parse tree" structure for exploring the solutions. Gene expression programming (GEP) is equivalent to GP. The chromosomes in GEP collect multiple genes, each gene converting a smaller subprogram. Moreover, the systematic organization of the linear chromosomes provides the unrestrained behavior of important genetic operators such as mutation, transposition and recombination (Ferreira, 2006). Major dominances of GP (i.e., GEP) are that it can be applied to areas where (a) the interrelationships among the pertinent factors are less clarified, (b) finding the conclusive solution is difficult, (c) normal mathematical investigation cannot supply analytical solutions, (d) a rough solution is acceptable, (e) small improvements in the performance are routinely measured and highly valued, and (f) there is a large amount of data which require evaluation, classification, and integration (Banzhaf et al., 1998). One of the major advantages of GEP is that it can generate an explicit equation between input(s) and output of the underlying problem. Such an equation might be subjected to some interpretation to find the governing rules of the studied process.

A number of studies (e.g., Walthal et al., 2004; Dunea and Moise, 2008; Xiao et al., 2014) applied data driven neural networks models for obtaining LAI from remotely sensed data and filling the gaps between the recorded data. Everingham et al. (2009) applied heuristic techniques to forecast regional sugarcane crop production. Torres et al. (2011) applied support vector machine to estimate daily potential evapotranspiration with limited climatic data. Shiri et al. (2014a) used heuristic techniques to model dew point temperature. Shiri et al. (2014b) showed the generalizability of GEP in modeling daily evapotranspiration in local and regional scales. Karimi et al. (2017) used GEP for simulating daily evapotranspiration through a cross-station approach.

Commonly, lots of heuristic-based applications contemplate only a single data set assignment where models are developed and validated utilizing data of the same site. Apart from not executing a perfect performance evaluation of the local patterns, another important drawback of this data set assignment type is that the generalization ability of the achieved models is not evaluated outside the locations that have been used to train the models (Marti et al., 2013; Shiri et al., 2014b).

The present study aimed at assessing the performances of GEP and RF techniques in local and external cross-station scales for simulating LAI, using available meteorological and NDVI data. By relying only on meteorological data, LAI can be obtained in the long past time and future when no remotely sensed data exists. To the best of authors' knowledge, this is the first assessment of heuristic methods in estimating LAI in local and cross-station scales. The robust k-fold testing cross validation technique was used for assessing the applied methodologies in both local and external scales.

2. Materials and methods

2.1. Data

The GEP and RF-based models were trained and tested extensively over ten experimental sites (with five cropland and five grassland sites). The meteorological data of the ten experimental sites were obtained via Fluxnet website (http://www.fluxnet.ornl.gov/). The site locations and data temporal coverage were summarized in Table 1. Half-hourly or hourly micrometeorological data such as wind speed, air temperature and humidity, atmospheric pressure, solar radiation, and incoming longwave radiation were measured at the ten experimental sites. The 16- day MODIS NDVI/EVI (MOD13A2) (Huete et al., 2002) and the 8-day MODIS FPAR/LAI (MOD15A2) products (Myneni et al., 2002) with 1 km spatial resolution were also collected in this research. The daily NDVI and LAI values were temporally interpolated from the 16- day or 8- day averages using linear interpolation.

In this study, the meteorological data including maximum air temperature (T_{max}), minimum air temperature (T_{min}), mean air temperature (Ta) and mean relative humidity (RH), as well as the remotely sensed normalized difference vegetation index (NDVI) were used as input parameters of the applied models for simulating the LAI. With these parameters, the vegetation growth could be well controlled, and thus could provide useful dynamic information for the LAI estimation (Stockli et al., 2008; Yao et al., 2008; Qu et al., 2012). Table 2 presents the statistical characteristics of the applied LAI records.

2.2. Study flowchart

The performances of the applied GEP and RF-based models were assessed via a cross-validation k-fold test procedure. Accordingly, the available patterns were divided into k blocks and the train-test process was repeated k times till a complete data scanning was achieved (Marti et al., 2013; Roushangar et al., 2014a). Each time, a separate set of data was reserved for testing. In the present study, two different criteria were developed for defining the minimum test set size for executing the k-fold assessment, e.g. external (spatial) and temporal (local) criteria. First, train-test processes (5-fold spatial assessment) per land cover type were performed leaving each time the whole available patterns of one site for testing. This procedure allowed evaluating the external generalizability of the GEP and RF- based models (Shiri et al., 2015). Nonetheless, the local (temporal) performance was assessed per location through a k-fold temporary assessment. This case was carried out separately at each site, where one year was used for testing and the rest data were utilized for training. Finally, a global approach was developed where the applied models were trained using all data from cropland stations, and tested in grassland sites, and vice versa.

Four statistical criteria were used for assessing the applied models, namely, the coefficient of determination (r^2), the scatter index (*SI*), the mean absolute error (*MAE*), and the Nash Sutcliffe coefficient (*NS*), expressions for which are given below:

$$r^2 = \left[\frac{\sum_{i=1}^{n} (LAI_{io} - \overline{LAI_o})(LAI_{iM} - \overline{LAI_M})}{\sqrt{\sum_{i=1}^{n} (LAI_{io} - \overline{LAI_o})^2 \sum_{i=1}^{n} (LAI_{iM} - \overline{LAI_M})^2}} \right]^2 \quad (1)$$

$$SI = \frac{RMSE}{\overline{LAI_o}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} (LAI_{i0} - LAI_{iM})^2}}{\overline{LAI_o}} \quad (2)$$

$$MAE = \frac{1}{n} \left| \sum_{i=1}^{n} (LAI_{i0} - LAI_{iM}) \right| \quad (3)$$

$$NS = 1 - \frac{\sum_{i=1}^{n} (LAI_{i0} - LAI_{iM})^2}{\sum_{i=1}^{n} (LAI_{i0} - \overline{LAI_o})^2} \quad (4)$$

where LAI_{i0} denotes the target (observed) LAI value at the i^{th} time step and LAI_{iM} represents the corresponding simulated value. n stands for the number of time steps, $\overline{LAI_o}$ shows the mean of the observed values and $\overline{LAI_M}$ is the mean value of the simulations.

2.3. Gene expression programming (GEP)

GEP is similar to genetic programming (GP) where it additionally evolves computer programs with wide size domains and shapes encoded in linear chromosomes with fixed lengths. The chromosomes in GEP consist of multiple genes, where every gene is encoding a subprogram. In addition, the structural and functional organization of the linear chromosomes provide the unconstrained operation of important genetic operators, e.g. mutation, transposition and recombination. One strength of the GEP approach is that the creation of genetic diversity is extremely simplified as genetic operators work at the chromosomes level. Another strength of GEP is its uniqueness of multigenic nature