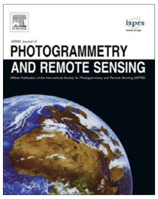
ISPRS Journal of Photogrammetry and Remote Sensing 135 (2018) 173–188

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



# A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning

Rasmus Houborg [a,b,*], Matthew F. McCabe [a]

[a] King Abdullah University of Science and Technology (KAUST), Water Desalination and Reuse Center (WDRC), Biological and Environmental Science & Engineering (BESE), Thuwal, Saudi Arabia

[b] Geospatial Sciences Center of Excellence, South Dakota State University, Brookings, SD 57007-3510, USA

ARTICLE INFO

*Article history:*
Received 10 June 2017
Received in revised form 1 October 2017
Accepted 4 October 2017

*Keywords:*
Machine-learning
LAI
Cubist
Random forests
Hybrid
REGFLEC

ABSTRACT

With an increasing volume and dimensionality of Earth observation data, enhanced integration of machine-learning methodologies is needed to effectively analyze and utilize these information rich datasets. In machine-learning, a training dataset is required to establish explicit associations between a suite of explanatory 'predictor' variables and the target property. The specifics of this learning process can significantly influence model validity and portability, with a higher generalization level expected with an increasing number of observable conditions being reflected in the training dataset. Here we propose a hybrid training approach for leaf area index (LAI) estimation, which harnesses synergistic attributes of scattered in-situ measurements and systematically distributed physically based model inversion results to enhance the information content and spatial representativeness of the training data. To do this, a complimentary training dataset of independent LAI was derived from a regularized model inversion of RapidEye surface reflectances and subsequently used to guide the development of LAI regression models via Cubist and random forests (RF) decision tree methods. The application of the hybrid training approach to a broad set of Landsat 8 vegetation index (VI) predictor variables resulted in significantly improved LAI prediction accuracies and spatial consistencies, relative to results relying on in-situ measurements alone for model training. In comparing the prediction capacity and portability of the two machine-learning algorithms, a pair of relatively simple multi-variate regression models established by Cubist performed best, with an overall relative mean absolute deviation (rMAD) of ~11%, determined based on a stringent scene-specific cross-validation approach. In comparison, the portability of RF regression models was less effective (i.e., an overall rMAD of ~15%), which was attributed partly to model saturation at high LAI in association with inherent extrapolation and transferability limitations. Explanatory VIs formed from bands in the near-infrared (NIR) and shortwave infrared domains (e.g., NDWI) were associated with the highest predictive ability, whereas Cubist models relying entirely on VIs based on NIR and red band combinations (e.g., NDVI) were associated with comparatively high uncertainties (i.e., rMAD ~ 21%). The most transferable and best performing models were based on combinations of several predictor variables, which included both NDWI- and NDVI-like variables. In this process, prior screening of input VIs based on an assessment of variable relevance served as an effective mechanism for optimizing prediction accuracies from both Cubist and RF. While this study demonstrated benefit in combining data mining operations with physically based constraints via a hybrid training approach, the concept of transferability and portability warrants further investigations in order to realize the full potential of emerging machine-learning techniques for regression purposes.

## 1. Introduction

Space-borne capacities for vegetation monitoring at high spatial resolutions (i.e., on the order of 30 m or finer) have expanded during the past decade as a result of government space agency missions such as Landsat 8 (Roy et al., 2014), Sentinel-2 (Drusch et al., 2012) and a growing number of commercial sensor constellations including RapidEye and WorldView (Houborg et al., 2015a). More recently, swarms of CubeSats have further advanced the spatiotemporal observation capacity (Houborg and McCabe, 2016a; McCabe et al., 2017). A number of proposed satellite missions offer

[*] Corresponding author at: Geospatial Sciences Center of Excellence, South Dakota State University, Brookings, SD 57007-3510, USA

*E-mail address:* rasmus.houborg@sdstate.edu (R. Houborg).

https://doi.org/10.1016/j.isprsjprs.2017.10.004

platforms for a range of spectrally enhanced sensors, including the Hyperspectral Precursor of the Application Mission (PRISMA) (Candela et al., 2016), the Environmental Mapping and Analysis Program (ENMAP) (Stuffler et al., 2007), and the Hyperspectral Infrared Imager (HyspIRI) (Roberts et al., 2012). If launched, these satellite programs would dramatically expand observation capacity and add to the already massive volumes of multi-dimensional data streams currently being produced. In order to effectively exploit and interpret the enhanced information content for accurate, robust and computationally efficient mapping of a range of earth surface variables over high resolution space and time domains, these observation sources will require the application of novel data driven retrieval methods (Verrelst et al., 2015a).

Traditionally, vegetation properties have been retrieved from Earth observation data via simple Vegetation Index (VI) relationships that are often established statistically (i.e., by fitting a standard regression function) on the basis of field measurements (Dorigo et al., 2007). As an example, leaf area index (LAI), which is a fundamental biophysical determinant of vegetation growth, health, and function (Anderson et al., 2015; Doraiswamy et al., 2004; Gitelson et al., 2014), has been related empirically to the Normalized Difference Vegetation Index (NDVI) (e.g., Chen and Cihlar, 1996; Turner et al., 1999), which itself has been correlated to plant height (Thenkabail et al., 2000). While the NDVI continues to serve as an important and widely adopted satellite based metric for vegetation growth and function (Candiago et al., 2015; Peters et al., 2002), it is associated with well-known limitations (Carlson and Ripley, 1997) and may not always be the best predictor of LAI dynamics (Kang et al., 2016; Viña et al., 2011).

The interpretation of satellite reflectance signals and subsequent generation of simple and generic variable-driven cause-effect relationships are significantly challenged by limitations in sensor radiometric information content (Houborg et al., 2015a) and confounding factors in highly complex spectrum-trait associations (Colwell, 1974). In reality, simple (e.g., single-variable) VI relationships established on the basis of a limited number of in-situ samples will be associated with significant deficiencies (Baret and Guyot, 1991; Colombo et al., 2003; Price, 1993) and prone to portability and transferability issues (i.e., reduced predictability beyond the realm of the calibration data) (Atzberger et al., 2015; Gobron et al., 1997; Houborg et al., 2007).

Advancements in the understanding and description of light absorption and scattering within vegetation canopies led to important developments in leaf optics and canopy radiative transfer modeling (Jacquemoud and Baret, 1990; Verhoef, 1984), initiating physically sound retrieval of vegetation properties via model inversion with satellite observed reflectances (Jacquemoud et al., 2009). While intrinsically generic and generally applicable (Dorigo et al., 2007), physically based approaches are challenged by competing demands of model realism and inversion feasibility, and simplified treatment of otherwise complex physical processes is typically needed to constrain the inversion process in a remote sensing context (Goel, 1988; Houborg et al., 2015b). Still, ill-posedness resulting from an under-determined problem, model and measurement uncertainties, and the association of near-identical spectra with different combinations of input model variables (Baret and Buis, 2008; Combal et al., 2002; Zurita-Milla et al., 2015), is often inevitable and can lead to significant confusion in the discrimination of model variable contributions (Daughtry et al., 2000). These issues have spawned significant research into the development of regularization strategies for introducing additional information, in an attempt to mitigate the ill-posedness of relevant vegetation properties (e.g., Atzberger and Richter, 2012; Bacour et al., 2002; Dorigo et al., 2009; Lauvernet et al., 2008; Rivera et al., 2013). Despite these efforts, physically based estimation of vegetation properties remains an involved and challenging task. Additionally, inversion approaches can be computationally demanding and complicated to implement, and are not typically readily adaptable within operational frameworks (Verrelst et al., 2015a).

Non-parametric (i.e., machine-learning) regression algorithms have recently emerged as a convenient interface linking the efficiency of standard statistical techniques with the detail and complexity of physically based approaches. They also provide a capacity to efficiently process an expanding volume of Earth observation data (Verrelst et al., 2015a, 2012). Typical learning approaches include decision trees (e.g. random forests; Breiman, 2001), artificial neural networks (Haykin, 1998), kernels (e.g. support vector machines; Vapnik et al., 1996), and Gaussian processes regression (Rasmussen and Williams, 2005). In machine-learning, complex associations (e.g., non-linear relationships) between a target property and a potentially unlimited number of explanatory predictor variables can be unraveled without explicit knowledge of underlying processes by "letting the data speak for itself" (e.g., McCabe et al., 2017). As mapping applications based on machine-learning typically rely on field collected data for model training (Mutanga et al., 2012; Verrelst et al., 2015b), previously mentioned issues regarding model transferability are likely to persist (Vuolo et al., 2013) and generalization capabilities may differ depending on the given algorithm (Verrelst et al., 2013, 2012).

As an alternative to the empirical training approach, machine-learning models have been trained with entirely synthetic datasets generated by forward runs of canopy radiative transfer models over a wide array of realizations (Atzberger, 2004; Doktor et al., 2014; Liang et al., 2015). While this training approach will ensure a higher generalization level, the robustness of resulting image-based retrievals will be affected by the level of consistency between modeled and satellite observed reflectance spectra (Baret and Buis, 2008; Houborg and McCabe, 2016b) and the degree of ill-posedness of the target variable. In addition, entrusting the machine-learning algorithm with the power to decode and reproduce physically based cause-effect relationships can also present risks, as reflected in reduced predictabilities (Vohland et al., 2010) and unrealistic conclusions regarding causality (Papagiannopoulou et al., 2017). A related training strategy was investigated by Houborg et al. (2007), which assumed regularized model inversion results for select pixels as substitutes for in-situ measurements. In contrast to the use of synthetic training data, this approach relies on inversion of actual observation data with flexibility in the handling and implementation of effective regularization strategies.

In this study, we expand upon the concept of using image-based model inversion results to help guide the development and calibration of computationally efficient regression models (Houborg et al., 2007; Vohland et al., 2010). The proposed hybrid model training approach exploits experimental in-situ measurements and physically based inversion results synergistically within a machine-learning context in order to mitigate limitations associated with using each training source individually. The general idea is to combine the unique and complementary attributes of each training source to help inform the learning process of select non-parametric regression models for improved predictability and transferability, with the expectation being that the more conditions covered in the training dataset, the more generic resulting regression models will be. Accordingly, the overall objective of the study is to assess prediction robustness and portability of LAI estimated on the basis of two common decision tree regression algorithms. This is done by using a hybrid multi-day training dataset consisting of in-situ measurements and physical model-based estimates derived from a LUT-based inversion of satellite observed surface reflectances. Investigations into the benefit of utilizing in-situ and physically based training sources synergistically in the context of model portability constitute a particularly novel aspect