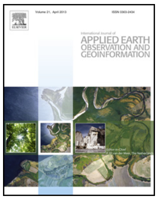
International Journal of Applied Earth Observation and Geoinformation 52 (2016) 554–567

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Spectral band selection for vegetation properties retrieval using Gaussian processes regression

Jochem Verrelst[a,*], Juan Pablo Rivera[a,b], Anatoly Gitelson[c], Jesus Delegido[a], José Moreno[a], Gustau Camps-Valls[a]

[a] Image Processing Laboratory (IPL), Parc Científic, Universitat de València, 46980 Paterna, València, Spain
[b] Departamento de Oceanografía Física, CICESE, 22860 Ensenada, Mexico
[c] Israel Institute of Technology, Technion, Haifa, Israel

ARTICLE INFO

Article history:
Received 10 May 2016
Received in revised form 20 July 2016
Accepted 21 July 2016
Available online 10 August 2016

Keywords:
Gaussian processes regression (GPR)
Machine learning
Band selection
ARTMO
Vegetation properties
Hyperspectral

ABSTRACT

With current and upcoming imaging spectrometers, automated band analysis techniques are needed to enable efficient identification of most informative bands to facilitate optimized processing of spectral data into estimates of biophysical variables. This paper introduces an automated spectral band analysis tool (BAT) based on Gaussian processes regression (GPR) for the spectral analysis of vegetation properties. The GPR-BAT procedure sequentially backwards removes the least contributing band in the regression model for a given variable until only one band is kept. GPR-BAT is implemented within the framework of the free ARTMO's MLRA (machine learning regression algorithms) toolbox, which is dedicated to the transforming of optical remote sensing images into biophysical products. GPR-BAT allows (1) to identify the most informative bands in relating spectral data to a biophysical variable, and (2) to find the least number of bands that preserve optimized accurate predictions. To illustrate its utility, two hyperspectral datasets were analyzed for most informative bands: (1) a field hyperspectral dataset (400–1100 nm at 2 nm resolution: 301 bands) with leaf chlorophyll content (LCC) and green leaf area index (gLAI) collected for maize and soybean (Nebraska, US); and (2) an airborne HyMap dataset (430–2490 nm: 125 bands) with LAI and canopy water content (CWC) collected for a variety of crops (Barrax, Spain). For each of these biophysical variables, optimized retrieval accuracies can be achieved with just 4 to 9 well-identified bands, and performance was largely improved over using all bands. A PROSAIL global sensitivity analysis was run to interpret the validity of these bands. Cross-validated R^2_{CV} (NRMSE$_{CV}$) accuracies for optimized GPR models were 0.79 (12.9%) for LCC, 0.94 (7.2%) for gLAI, 0.95 (6.5%) for LAI and 0.95 (7.2%) for CWC. This study concludes that a wise band selection of hyperspectral data is strictly required for optimal vegetation properties mapping.

1. Introduction

A new era of optical remote sensing science is emerging with forthcoming space-borne imaging spectrometer missions such as EnMAP (Environmental Mapping and Analysis Program) (Guanter et al., 2015), HyspIRI (Hyperspectral Infrared Imager) (Roberts et al., 2012), PRISMA (PRecursore IperSpettrale della Missione Applicativa) (Labate et al., 2009) and ESA's 8th Earth Explorer FLEX (Fluorescence Explorer) (Kraft et al., 2012). Having access to operationally acquired imaging spectroscopy data with hundreds of bands paves the path for a wide variety of monitoring applications, such as the quantification of structural and biochemical vegetation properties (Schaepman et al., 2009; Ustin and Gamon, 2010; Homolová et al., 2013).

Facing such exciting new technological opportunity poses, however, an important methodological challenge. Imaging spectroscopy data include highly correlated and noisy spectral bands, and frequently create statistical problems (e.g., the Hughes effect) due to small sample sizes compared to the large number of available, possibly redundant, spectral bands. These characteristics may lead to a violation of basic assumptions behind statistical models or may otherwise affect the model outcome. Models fitted with such multi-collinear data sets are prone to over-fitting, and transfer to other scenarios may thus be limited. Naturally, these issues affect the prediction accuracy as well as the interpretability of the regression (retrieval) models (Curran, 1989; Grossman et al., 1996). It may therefore be desirable to reduce the spectral dimension, either through spectral dimensionality reduction techniques

[*] Corresponding author.
E-mail address: jochem.verrelst@uv.es (J. Verrelst).

http://dx.doi.org/10.1016/j.jag.2016.07.016
0303-2434/© 2016 Elsevier B.V. All rights reserved.

(e.g., Van Der Maaten et al., 2007; Arenas-García et al., 2013) or to select particular spectral regions that are most helpful to describe targeted biophysical variables. Apart from improving the fit and processing speed of regression models, selecting specific spectral regions may allow clarification of the relationships of spectral signatures to leaf and canopy optical properties while minimizing the signal from secondary responses (Feilhauer et al., 2015).

From a pure statistical signal processing point of view, band selection is cast as an optimization problem by which one wants to select a subset of spectral bands that capture most of the information for a particular problem. The search for the best bands out of the available is known to be an NP-complete problem (it cannot be solved in polynomial time) (Blum and Langley, 1998) and the number of local minima can be quite large. This poses both numerical and computational difficulties.

Following a general taxonomy, *band selection* can be divided into two major categories: filter methods (Liu and Motoda, 1998) and wrapper (Kohavi and John, 1997) methods. *Filter methods* use an indirect measure of the quality of the selected bands, so a faster convergence of the regression algorithm is obtained. *Wrapper methods* use the output of the regression algorithm as selection criteria. This approach guarantees that in each step of the algorithm, the selected subset improves the performance of the previous one. Filter methods might fail to select the right subset of bands if the used criterion deviates from the one used for training the regression algorithm, whereas wrapper methods can be computationally intensive since the regression algorithm has to be retrained for each new set of bands.

Spectral band selection for quantifying vegetation properties have used both filter and wrapper band selection methods. Although there are many works of feature (spectral band) selection in imaging spectroscopy and remote sensing, the vast majority of them are related to classification problems (e.g., Bazi and Melgani, 2006; Archibald and Fann, 2007; Pal and Foody, 2010); very few are concerned with regression (retrieval) problems, and in particular with vegetation properties estimation.

On the one hand, filter methods have long been restricted to the systematic calculation of all possible band combinations, e.g. through generic vegetation indices where all bands are combined into two-band indices and then applied to regression (e.g., Heiskanen et al., 2013; Rivera et al., 2014b). However, these are brute-force techniques that usually do not go beyond searching for (linear or polynomial) combinations of two or at most three bands that maximize a fitting criterion (typically linear correlation). On the other hand, wrapper methods have been also applied in the field of chemometrics (Forina et al., 2004; Andersen and Bro, 2010). Here, the focus is on non-parametric, multivariate regression methods. These are full-spectrum statistical methods, and some of them have band ranking properties through wrapper methods. There is a large evidence of their successful performance. For instance, Feilhauer et al. (2015) compared three multivariate regression techniques (partial least square regression, random forests regression, and support vector regression) in their suitability for the identification and selection of spectral bands. A multi-method ensemble strategy, i.e. decision fusion, using these three methods was proposed in order to crystallize a more robust band selection. Among preferred univariate regression methods we find random forests (Genuer et al., 2010) mostly embedded in genetic algorithm procedures (Jung and Zscheischler, 2013), or via permutation analyses.

A drawback of the above wrapper methods is that they are often perceived as complex, e.g. they require software packages and parameter tuning is mostly needed, and not all of these methods performed equally well (Feilhauer et al., 2015). Using a regression method with few hyper-parameters to be tuned is perhaps the main problem here, and alternatives exist. Actually, various alternative non-parametric multivariate methods in the field of machine learning regression algorithms (MLRAs) equally possess band selection/ranking features, which some of them are very competitive. Comparison studies have demonstrated that the above-mentioned methods may not always be most powerful regression algorithms (Rivera et al., 2014a; Verrelst et al., 2012b, 2015c). In these studies, it was shown that Gaussian processes regression (GPR) (Rasmussen and Williams, 2006) outperformed other MLRAs for the retrieval of biophysical variables from airborne and satellite images (Verrelst et al., 2012b, 2015c). Of interest is that GPR also provides band ranking feature, which reveals the bands that contribute most to the development of a GPR model (Camps-Valls et al., 2016). Given its powerful performance, GPR may be a first choice to exploit band ranking features.

Altogether, apart from above and a few more experimental studies (e.g., Verrelst et al., 2012b,a; Van Wittenberghe et al., 2014), band ranking has not been fully exploited in retrieval applications. So far all these studies are experimental, and – while having their scientific merits – none of these methods are directly applicable to operational processing of hyperspectral data streams. For instance, in view of optimized vegetation properties mapping, no user-friendly software package enabling automated identification of most important spectral bands for a given biophysical variable is available to the broader community. Such kinds of tools may become critical when forthcoming unprecedented hyperspectral data stream will become freely accessible.

The objectives of this work are therefore threefold: (1) to develop a GPR-based band analysis tool, further referred to as "GPR-BAT", that analyzes the band-specific information content of spectral data for a given biophysical variable with little user interaction; (2) to demonstrate GPR-BAT's utility by applying it to two extensive hyperspectral datasets (biophysical variables and associated spectra) in order to identify the optimal number of bands and their spectral location; and finally, (3) to apply GPR-BAT to an airborne hyperspectral image for automated and optimized vegetation properties mapping. GPR-BAT will be operated as a graphical user interface (GUI) within ARTMO's (automated radiative transfer models operator) (Verrelst et al., 2012c) machine learning regression algorithm (MLRA) toolbox (Rivera et al., 2014a). To assess the optimality of the identified bands, we will run a global sensitivity analysis applied to the physically based PROSAIL canopy radiative transfer model (RTM).

2. Gaussian processes regression

Estimation, regression and function approximation are old, largely studied problems in statistics and machine learning. The problem boils down to optimize a loss (cost, energy) function over a class of functions. A large class of regression problems in particular are defined as the joint minimization of a loss function accounting for errors of the function $f \in \mathcal{H}$ to be learned, and a regularization term, $\Omega\left(\|f\|_{\mathcal{H}}^2\right)$, that controls its capacity (excess of flexibility). The problem can be approached within a Bayesian nonparametric framework, and several algorithms are available, such as the relevance vector machine (Tipping, 2001; Camps-Valls et al., 2006) or Gaussian Processes regression (GPR) (Rasmussen and Williams, 2006; Camps-Valls et al., 2016), in which we will focus here.

GPR is equivalent in nature to kernel ridge regression (aka least square support vector machine) and kriging. However, due to their high computational complexity they did not become widely applied tools in machine learning until recently. GPR can be interpreted as a family of kernel methods with the additional advantage of providing a full conditional statistical description for the predicted variable, which can be primarily used to establish confidence intervals and to set hyper-parameters (Rasmussen and Williams, 2006). In short, GPR assumes that a Gaussian process prior governs