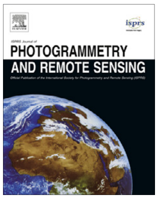
ISPRS Journal of Photogrammetry and Remote Sensing 132 (2017) 88–101

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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



Hyperspectral dimensionality reduction for biophysical variable statistical retrieval

Juan Pablo Rivera-Caicedo [a,b], Jochem Verrelst [a,*], Jordi Muñoz-Marí [a], Gustau Camps-Valls [a], José Moreno [a]

[a] *Image Processing Laboratory (IPL), Parc Científic, Universitat de València, 46980 Paterna, Spain*
[b] *Secretary of Research and Postgraduate, CONACYT-UAN, 63155 Tepic, Nayarit, Mexico*

ARTICLE INFO

Article history:
Received 28 February 2017
Received in revised form 21 August 2017
Accepted 25 August 2017

Keywords:
Spectral dimensionality reduction methods
Machine learning regression algorithms
Biophysical parameter retrieval
ARTMO
Hyperspectral
Vegetation properties

ABSTRACT

Current and upcoming airborne and spaceborne imaging spectrometers lead to vast hyperspectral data streams. This scenario calls for automated and optimized spectral dimensionality reduction techniques to enable fast and efficient hyperspectral data processing, such as inferring vegetation properties. In preparation of next generation biophysical variable retrieval methods applicable to hyperspectral data, we present the evaluation of 11 dimensionality reduction (DR) methods in combination with advanced machine learning regression algorithms (MLRAs) for statistical variable retrieval. Two unique hyperspectral datasets were analyzed on the predictive power of DR + MLRA methods to retrieve leaf area index (LAI): (1) a simulated PROSAIL reflectance data (2101 bands), and (2) a field dataset from airborne HyMap data (125 bands). For the majority of MLRAs, applying first a DR method leads to superior retrieval accuracies and substantial gains in processing speed as opposed to using all bands into the regression algorithm. This was especially noticeable for the PROSAIL dataset: in the most extreme case, using the classical linear regression (LR), validation results R^2_{CV} (RMSE$_{CV}$) improved from 0.06 (12.23) without a DR method to 0.93 (0.53) when combining it with a best performing DR method (i.e., CCA or OPLS). However, these DR methods no longer excelled when applied to noisy or real sensor data such as HyMap. Then the combination of kernel CCA (KCCA) with LR, or a classical PCA and PLS with a MLRA showed more robust performances (R^2_{CV} of 0.93). Gaussian processes regression (GPR) uncertainty estimates revealed that LAI maps as trained in combination with a DR method can lead to lower uncertainties, as opposed to using all HyMap bands. The obtained results demonstrated that, in general, biophysical variable retrieval from hyperspectral data can largely benefit from dimensionality reduction in both accuracy and computational efficiency.

1. Introduction

Spatio-temporally explicit, quantitative retrieval methods for Earth surface are a requirement in a variety of Earth system applications. Optical Earth observing satellites, endowed with a high spectral resolution, enable the retrieval and hence monitoring of continuous bio-geophysical variables (Schaepman et al., 2009). With forthcoming operational imaging spectrometers, such as EnMAP (Guanter et al., 2015), HyspIRI (Roberts et al., 2012), PRISMA (Labate et al., 2009) and ESA's 8th Earth Explorer FLEX mission (Drusch et al., 2016), an unprecedented data stream for land monitoring will soon become available to a diverse user community. These massive data streams will require enhanced processing techniques that are accurate, robust and fast. One of the major challenges with these data streams is the large amount of spectral data that has to be processed.

Over the last few decades, a wide diversity of bio-geophysical retrieval methods have been developed, but only a few of them made it into operational processing chains and many of them are still in its infancy and not fully adapted to hyperspectral data (Verrelst et al., 2015a). Essentially, we may find four main approaches for the inverse problem of estimating biophysical variables from spectra: statistical, i.e. (1) parametric and (2) nonparametric regression; (3) physically-based; and (4) hybrid regression methods. Hybrid methods combine elements of non-parametric regression and physically-based methods. These methods exploit the generic properties of physically-based models combined with the flexibility and computational efficiency of non-parametric, non-linear regression models (Verrelst et al., 2015a). They proved

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

http://dx.doi.org/10.1016/j.isprsjprs.2017.08.012
0924-2716/© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

to be particularly successful in operational generation of land products such as leaf area index (LAI). However, current hybrid methods rely exclusively on neural networks (NN), typically trained by a very large amount of simulated data as generated by radiative transfer models (RTMs) (e.g., Baret et al., 2007, 2013; Verger et al., 2008). For instance, when it comes to LAI retrieval then commonly the PROSAIL model (PROSPECT + SAIL) is used to generate training data (Jacquemoud et al., 2009). This approach works fine to multispectral data but becomes challenging when applied to hyperspectral data due to the computational cost in training a NN with many bands.

Beyond NN, various alternative nonparametric methods in the field of machine learning regression algorithms (MLRAs) have been recently introduced, many of them with interesting properties. Especially bagging/boosting of regression trees (RT), random forests (RF) and kernel-based methods such as kernel ridge regression (KRR) have proven to be simpler and faster to train, providing competitive accuracies. Some of these kernel-based MLRAs such as Gaussian processes regression (GPR) even provide associated uncertainties in a Bayesian framework (Verrelst et al., 2012b, 2015b). A drawback of these advanced statistical regression algorithms (including NN) for retrieving biophysical variables, however, is that they also come with a computational cost, especially when large datasets are involved in the training phase, such as when simulated data are used typically in hybrid schemes. Consequently, reduction of the training data space while retaining as much information as possible would enable to alleviate these computational drawbacks.

Reduction of the training dataset can essentially take place in two domains: (1) in the sampling domain, i.e. by selecting only the most informative samples, e.g. through active learning techniques (MacKay, 1992; Tuia et al., 2011; Crawford et al., 2013; Verrelst et al., 2016a), and (2) in the spectral domain, i.e. by making use of feature (band) selection and feature extraction or dimensionality reduction (DR) techniques (Van Der Maaten et al., 2009). While the first type of methods aim to minimize the amount of samples while preserving high accuracies, the second type of methods aim to bypass the so-called "curse of dimensionality" (Hughes phenomenon) (Hughes, 1968) that is commonly observed in hyperspectral data. Adjacent hyperspectral bands carry highly correlated information which may result in redundant data and possible noise and potentially suboptimal performances. In feature (band) selection, the aim is to define a subset of the original bands that maintains the useful information to apply regression with highly correlated and redundant bands excluded from the regression analysis. In parametric regression, this is typically done by systematically calculating all possible two-band combinations in vegetation indices formulations, (e.g., le Maire et al., 2008; Rivera et al., 2014b). More elegant methods exist by making use of band ranking properties provided by regression methods, such as in GPR or random forests, e.g. (Van Wittenberghe et al., 2014; Feilhauer et al., 2015). For instance, (Verrelst et al., 2016b) recently developed an automated sequential band removal procedure to identify most sensitive bands based on GPR band ranking.

Alternatively, in DR methods the original spectral data is transformed in some way that allows the definition of a small set of new features (components) in a lower-dimensional space which contain the vast majority of the original data set's information (Liu and Motoda, 1998; Lee and Verleysen, 2007). As such, there is no need to search for most relevant spectral bands, and thus simplifies the retrieval problem. Especially in data classification a plethora of feature extraction and DR methods are available in the literature (e.g., Arenas-Garcia et al., 2013; Damodaran and Nidamanuri, 2014). Surprisingly less progress in DR methods has been presented when it comes to biophysical variable retrieval (regression). If a DR method at all is applied, then it is by the classical principal component analysis (PCA) (Jolliffe, 1986; Liu et al., 2016). Although PCA has proven its use in a broad diversity of applications, and continues to be the first choice in vegetation properties mapping based on hyperspectral data, situations may occur where PCA is not the best choice and alternatives have to be sought. As an extension of PCA, partial least squares (PLS) introduces some refinements by looking for projections that maximize the covariance and correlations between spectral information and input variables (Wold, 1966). PLS regression (PLSR) became a popular regression method in chemometrics and remote sensing applications (e.g. see Verrelst et al. (2015a) for review), however, the regression part of PLSR and principal component regression (PCR) has always been restricted to multiple linear regression. It remains to be questioned how well PLS combines with more advanced, nonlinear regression methods. Beyond PCA and PLS, only a few DR-regression studies have been presented, including a semi-supervised DR where the data distribution resides on a low-dimensional manifold has been proposed (Uto et al., 2014). But this method was only applied to linear regression. Apart from Laparra et al. (2015) and Arenas-Garcia et al. (2013) where a few alternative DR methods were proposed, the combined use of DR with advanced regression methods for biophysical variable estimation has been largely left unexplored. Nonetheless, there is no doubt DR methods may become prevalent within the context of introducing advanced regression methods into new generation hybrid retrieval processing chains. This especially holds for LAI retrieval; LAI is characterized by a broad sensitive spectral range (e.g. see global sensitivity analysis Verrelst et al. (2015c)) and thus perfectly suited for a DR conversion step.

In this respect, apart from PCA and PLS, in this work we evaluate 9 alternative DR methods into regression, including canonical correlation analysis (CCA), orthonormalized PLS (OPLS) and minimum noise fraction (MNF), as well as their nonlinear extensions derived by means of the theory of reproducing kernel Hilbert spaces. All these methods have been put together into an in-house developed MATLAB library called SIMFEAT (Arenas-Garcia et al., 2013), which has been now included in a free graphical user interface (GUI) retrieval toolbox.

This brings us to the following objectives: (1) to implement multiple DR methods into a software framework that enables semi-automatic development and validation of (hybrid) statistical retrieval strategies, and (2) to evaluate the efficacy of the SIMFEAT DR methods in combination with advanced regression methods in optimizing statistical LAI retrieval from hyperspectral data. Two experiments are presented. First, a hybrid scheme where the regression algorithms are trained by simulated data coming from PROSAIL. Second, an experimental dataset where the regression algorithms are trained by data coming from ESA's SPARC campaign (Barrax, Spain).

In the following, we will explain the implemented DR methods and used regression techniques (Section 2). This is followed by a description of the developed software and experimental setup (Section 3) and a presentation of the results (Section 4). The work closes with a discussion (Section 5) and a conclusion (Section 6).

2. Function approximation as a multivariate data analysis problem

The problem of regression and variable retrieval aims to *learn* a function $f(\cdot)$ that, based on input hyperspectral data $x \in \mathcal{X}$ can predict an output target variable or biophysical parameter $y \in \mathcal{Y}$. The problem can be approached *directly* with nonlinear regression methods implementing $f(\cdot)$, e.g. with neural networks, random forests, or kernel machines. Despite its efficiency, this approach leads to hidden representations that are hard to analyze, understand and visualize. Alternatively, one can approach the problem by learning

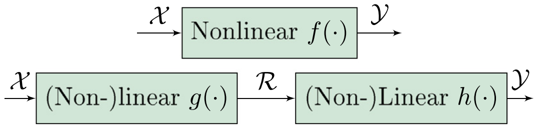


Fig. 1. Nonlinear function approximation can be approached (top) *directly* by developing a nonlinear approximation function *f*, or (bottom) *indirectly* by learning a (non-) linear feature extraction transform *g* plus a (non-) linear transform *h* for fitting, i.e. the nonlinear step can either be applied in the feature extraction or in the regression, or in both.

an intermediate transformation $g(\cdot)$ from the original, potentially high-dimensional feature space $\mathcal{X}$, to an accessible representation space of fewer dimensions, $\mathcal{R}$. From there, one only has to project data into $\mathcal{R}$ to perform a simple linear transform, $h(\cdot)$, to infer the output variable (see Fig. 1). This second approach delivers two important advantages: (1) the first nonlinear step leads to an accessible feature space of lower dimensionality, and (2) the second linear step typically involves solving simpler, faster convex optimization problems. Accordingly, we focus on developing the *g* feature extractors, and to evaluate different linear and nonlinear regression methods for *h*.

2.1. Dimensionality reduction

Extracting meaningful features (or components) from multidimensional data is typically done using the canonical principal component analysis (PCA), aka Empirical Orthogonal Functions (EOF) (Pearson, 1901). Nevertheless, many other DR methods are available in the literature. See Van Der Maaten et al. (e.g., 2009); Arenas-Garcia et al. (e.g., 2013) for a comparative review. Multivariate analysis (MVA) constitutes a family of methods for DR (Arenas-Garcia et al., 2013). The goal of MVA algorithms is to exploit correlations among the variables to find a reduced set of features that are relevant for the learning task. Among the most well-known MVA methods are PCA (Jolliffe, 1986), partial least squares (PLS) (Wold, 1966), canonical correlation analysis (CCA) (Hotelling, 1936), and minimum noise fraction (MNF) algorithm (Green et al., 1988). PCA disregards the target data and exploits correlations between the input variables to maximize the variance of the projections, while PLS and CCA look for projections that maximize, respectively, the covariance and the correlation between the features and the target data. Therefore, they should, in principle, be preferred to PCA for regression or classification problems. A fifth MVA method known as orthonormalized PLS (OPLS) optimizes the projection to achieve optimal results in least squares terms (Borga et al., 1997). A common advantage of all these DR methods is that they can be formulated using standard linear algebra and can be implemented as standard (or generalized) eigenvalue problems.

No matter how refined the various MVA methods are, they are still constrained to account for linear input-output relations. Hence, they can be severely challenged when features exhibit nonlinear relations between them or with the observed target variable. To address these problems, nonlinear versions of MVA methods have been developed, and these can be classified into two fundamentally different approaches (Rosipal, 2010): (1) The modified methods in which the linear relations among the latent variables are substituted by nonlinear parametric relations (Wold et al., 1989; Qin and McAvoy, 1992); and (2) variants in which the algorithms are reformulated to fit a kernel-based approach (Scholkopf et al., 1998; Shawe-Taylor and Cristianini, 2004; Nielsen, 2011). We will focus here on the latter approach. A central property of the kernel approach is the exploitation of the ‘‘kernel trick,” by which the inner products between training samples in the transformed space are replaced by a kernel function working solely with input space data, so knowing the nonlinear mapping is not explicitly necessary. Table 1 provides a summary of the MVA methods.

While the above MVA methods are well-known (Arenas-Garcia et al., 2013), one additional kernel method is briefly explained below, being kernel entropy component analysis (KECA) (Jenssen, 2010). The goal of KECA is to extract features according to the entropy components. As in KPCA, KECA is based on the kernel similarity matrix. However, while KPCA tries to preserve the second-order statistics of the data set, KECA is based on the information theory and tries to preserve the maximum Rényi entropy of the input data set. KECA has been successfully used in remote sensing data processing (Gómez-Chova et al., 2012; Luo and Wu, 2012; Luo et al., 2013). We implemented and evaluated the above-described linear MVA methods (PCA, PLS, CCA, MNF, OPLS), as well as their kernel versions (KPCA, KPLS, KCCA, KMNF, KOPLS, KECA) in a MATLAB library called SIMFEAT.[1]

After learning the $g(\cdot)$ transformation, one can actually project data onto a subspace $\mathcal{R}$. If *g* is implemented with a nonlinear (kernel) MVA method, one should have ideally captured all nonlinear relations in the data and then $h(\cdot)$ could be *optimally* implemented with linear fitting. Alternatively, one could exploit nonlinear machine learning regression for $h(\cdot)$ as well, in order to further account for remaining nonlinear feature dependencies. We will assess both pathways and the motivating hypothesis experimentally.

2.2. Regression and function approximation

Apart from the ordinary least squares (OLS) linear regression, we tested 7 advanced nonlinear MLRAs, i.e. bagging and boosting decision trees, random forests, neural networks, kernel ridge regression and Gaussian processes regression. These MLRAs can be categorized into three groups: (1) decision trees, (2) neural networks, and (3) kernel methods, and are briefly outlined below.

Decision tree learning is based on decision tree predictive modeling. A decision tree is based on a set of hierarchical connected nodes. Each node represents a linear decision based on a specific input feature. A classical decision tree algorithm cannot cope with strong non-linear input-output transfer functions. In that case, a combination of decision trees can improve results, such as bagging (Breiman, 1996), boosting (Friedman et al., 2000) and random forests (Breiman, 2001).

Artificial neural networks (ANNs) are essentially fully connected layered structures of artificial neurons (AN) (Haykin, 1999). A NN is a (potentially fully) connected structure of neurons organized in layers. Neurons of different layers are interconnected with the corresponding links (weights). Training a NN implies selecting a structure (number of hidden layers and nodes *per* layer), initialize the weights, shape of the nonlinearity, learning rate, and regularization parameters to prevent overfitting. The selection of a training algorithm and the loss function both have an impact on the final model. In this work, we used the standard multi-layer perceptron, which is a fully-connected network. We selected just one hidden layer of neurons. By default, we optimized the NN structure using the Levenberg-Marquardt learning algorithm with a squared loss function. However, in case this algorithm takes too long computational time then the option is provided to switch to a faster optimization based on a conjugate gradient backpropagation algorithm.

Kernel methods in machine learning owe their name to the use of kernel functions. Kernels quantify similarities between input samples of a dataset (Shawe-Taylor and Cristianini, 2004). Similar-

[1] http://isp.uv.es/soft_feature.html.

Table 1
Summary of linear and kernel MVA methods brought together in SIMFEAT library. Vectors $\mathbf{u}$ and $\boldsymbol{\alpha}$ are column vectors in matrices $\mathbf{U}$ and $\mathbf{A}$, respectively. $r(\cdot)$ denotes the rank of a matrix. For each method it is stated the objective to maximize (1st row), constraints for the optimization (second row), and maximum number of features (last row). More information can be found in Arenas-Garcia et al. (2013).

PCA	PLS	CCA	OPLS	MNF	KPCA	KPLS	KCCA	KOPLS	KMNF
Pearson (1901)	Wold (1966)	Hotelling (1936)	Borga et al. (1997)	Green et al. (1988)	Scholkopf et al. (1998)	Shawe-Taylor and Cristianini (2004)	Shawe-Taylor and Cristianini (2004)	Arenas-Garcia et al. (2013)	Nielsen (2011)
$\mathbf{u}^\top\mathbf{C}_x\mathbf{u}$	$\mathbf{u}^\top\mathbf{C}_{xy}\mathbf{v}$	$\mathbf{u}^\top\mathbf{C}_{xy}\mathbf{v}$	$\mathbf{u}^\top\mathbf{C}_{xy}\mathbf{C}_{xy}^\top\mathbf{u}$	$\mathbf{u}^\top\mathbf{C}_{xx}\mathbf{u}/\mathbf{u}^\top\mathbf{C}_{nn}\mathbf{u}$	$\boldsymbol{\alpha}^\top\mathbf{K}_x^2\boldsymbol{\alpha}$	$\boldsymbol{\alpha}^\top\mathbf{K}_x\mathbf{Y}\mathbf{v}$	$\boldsymbol{\alpha}^\top\mathbf{K}_x\mathbf{Y}\mathbf{v}$	$\boldsymbol{\alpha}^\top\mathbf{K}_x\mathbf{Y}\mathbf{Y}^\top\mathbf{K}_x\boldsymbol{\alpha}$	$\boldsymbol{\alpha}^\top\mathbf{K}_x^2\boldsymbol{\alpha}/\boldsymbol{\alpha}^\top\mathbf{K}_{xn}\mathbf{K}_{nx}\boldsymbol{\alpha}$
$\mathbf{U}^\top\mathbf{U}=\mathbf{I}$	$\mathbf{U}^\top\mathbf{U}=\mathbf{I}$ $\mathbf{V}^\top\mathbf{V}=\mathbf{I}$	$\mathbf{U}^\top\mathbf{C}_x\mathbf{U}=\mathbf{I}$ $\mathbf{V}^\top\mathbf{C}_y\mathbf{V}=\mathbf{I}$	$\mathbf{U}^\top\mathbf{C}_x\mathbf{U}=\mathbf{I}$	$\mathbf{U}^\top\mathbf{C}_{nn}\mathbf{U}=\mathbf{I}$	$\mathbf{A}^\top\mathbf{K}_x\mathbf{A}=\mathbf{I}$	$\mathbf{A}^\top\mathbf{K}_x\mathbf{A}=\mathbf{I}$ $\mathbf{V}^\top\mathbf{V}=\mathbf{I}$	$\mathbf{A}^\top\mathbf{K}_x^2\mathbf{A}=\mathbf{I}$ $\mathbf{V}^\top\mathbf{C}_y\mathbf{V}=\mathbf{I}$	$\mathbf{A}^\top\mathbf{K}_x^2\mathbf{A}=\mathbf{I}$	$\mathbf{A}^\top\mathbf{K}_{xn}\mathbf{K}_{nx}\mathbf{A}=\mathbf{I}$
$r(\mathbf{X})$	$r(\mathbf{X})$	$r(\mathbf{C}_{xy})$	$r(\mathbf{C}_{xy})$	$r(\mathbf{C}_{xn})$	$r(\mathbf{K}_x)$	$r(\mathbf{K}_x)$	$r(\mathbf{K}_x\mathbf{Y})$	$r(\mathbf{K}_x\mathbf{Y})$	$r(\mathbf{K}_x\mathbf{K}_{xn})$

Table 2
Evaluated non-parametric regression algorithms of the MLRA toolbox. More information can be found in Verrelst et al. (2015a,b).

Name algorithm	Principle
Linear regression (LR) (Hagan and Menhaj, 1994)	Least squares fit with ℓ_2 regularization
Bagging trees (BaT) (Breiman, 1996)	Bootstrap aggregation (bagging) + regression trees (RT)
Boosting trees (BoT) (Friedman et al., 2000)	Boosting + RT
Random forests (RF) (Breiman, 2001)	Bootstrap on samples and features + RT
Neural Network (NN) (Haykin, 1999)	Levenberg-Marquardt algorithm
Kernel ridge regression (KRR) (Suykens and Vandewalle, 1999)	Matrix inversion
Gaussian processes regression (GPR) (Rasmussen and Williams, 2006)	Bayesian statistical inference

ity reproduces a linear dot product (scalar) computed in a possibly higher dimensional feature space, yet without ever computing the data location in the feature space. The following two methods are gaining increasing attention: (1) Kernel ridge regression (KRR), also known as least squares support vector machines (Suykens and Vandewalle, 1999), and (2) Gaussian processes regression (GPR), based on Gaussian processes, which generalize Gaussian probability distributions in function spaces (Rasmussen and Williams, 2006). For both we used a standard radial basis function (RBF) kernel.

The evaluated regression methods are provided in Table 2. A more detailed comprehensive description of the methods is given in Rivera et al. (2014a); Verrelst et al. (2015a) and a MATLAB implementation in Camps-Valls et al. (2013).

2.3. Automating model analysis and vegetation mapping

The previous SIMFEAT dimensionality reduction (Arenas-Garcia et al., 2013) and regression (Camps-Valls et al., 2013) toolboxes were integrated in an in-house developed MATLAB package named ARTMO (Automated Radiative Transfer Models Operator) (Verrelst et al., 2012c).[2] ARTMO embodies a suite of leaf and canopy radiative transfer models (RTMs) including PROSAIL (i.e. the leaf model PROSPECT coupled with the canopy model SAIL (Jacquemoud et al., 2009)) and several retrieval toolboxes, i.e. a spectral indices toolbox, (Rivera et al., 2014b), a LUT-based inversion toolbox (Rivera et al., 2013), and a machine learning regression algorithms (MLRA) toolbox (Rivera et al., 2014a; Camps-Valls et al., 2013). These retrieval toolboxes enable the user to optimize and validate retrieval algorithms and subsequently process optical remote sensing data into maps with little user interaction. We have updated the MLRA toolbox (v1.19) with implementation of SIMFEAT in order to evaluate and compare ensembles of DR methods with MLRAs in a semi-automatic fashion.

In practice, we follow a two-step approach: first a DR method is applied, and then a regression algorithm. Regarding the implementation of the kernelized DR methods, an additional regularization step was introduced in order to tune the kernel hyperparameters as a function of the regression method employed. Another major difficulty when tackling one variable prediction (e.g. LAI) is that *supervised* DR methods can only extract a maximum of one feature (the output space rank). This is the case of powerful methods such as CCA and OPLS and their kernel variants. We solved this by discretizing the output space via clustering using k-means. A 1-of-k encoding was used for the output space. This allows us to apply CCA, OPLS, KCCA or KOPLS to obtain a maximum of k components (the rank of the output space). By default the number of components was set equal to the number of clusters, although the user can opt to split the data into more clusters. This simple strategy allows in turn to develop *local* feature extraction by learning class-dependent subspace projections.

2.4. Evaluation of results

For the validation of the trained models we used different goodness-of-fit statistical indicators: coefficient of determination, R^2; root mean square error: RMSE; and normalized RMSE: NRMSE. Additionally, to ensure robust identification of validation results, we combined the methods with a k-fold CV sub-sampling scheme. This scheme first splits randomly the training data into k mutually exclusive subsets (folds) of equal size and then by training k times a regression model with variable-spectra pairs. Each time, we left out one of the subsets from training and used it (the omitted subset) only to obtain an estimate of the regression accuracy (R^2, RMSE, NRMSE). From k times of training and validation, the resulting validation accuracies were averaged and basic statistics calculated (standard deviation (SD), dynamic range) to yield a more robust validation estimate of the considered regression model (see also (Verrelst et al., 2015b)). Finally, for each retrieval strategy, i.e. combination of DR and regression method, the training and validation processing time is tracked.

3. Data and methodology

3.1. PROSAIL dataset

Two datasets are analyzed. The first dataset involves simulated data as generated by the widely used PROSAIL RTM. PROSAIL is a coupled leaf reflectance model PROSPECT with a canopy reflectance model SAIL (Jacquemoud et al., 2009). At the leaf scale, the PROSPECT-4 model (Feret et al., 2008) is currently one of the most widely used leaf optical models and is based on earlier PROSPECT versions. The model calculates leaf reflectance and transmittance as a function of its biochemistry and anatomical structure. It consists of four parameters, those being leaf structure (N), chlorophyll content (Cab), equivalent water thickness (Cw) and dry matter con-

[2] http://ipl.uv.es/artmo/.

tent (*Cd*). PROSPECT-4 simulates directional reflectance and transmittance over the solar spectrum from 400 to 2500 nm at the fine spectral resolution of 1 nm. This output serves as input into the canopy model SAIL (Verhoef, 1984). SAIL is easy to use due to its low number of input variables. The model is based on a four-stream approximation of the RT equation, in which case one distinguishes two direct fluxes (incident solar flux and radiance in the viewing direction) and two diffuse fluxes (upward and downward hemispherical flux) (Verhoef et al., 2007). SAIL inputs consist of leaf area index (LAI), leaf angle distribution (LAD), ratio of diffuse and direct radiation, soil coefficient, hot spot and sun-target-sensor geometry, i.e. solar/view zenith angle and relative azimuth angle (SZA, VZA and RAA, respectively). PROSAIL generates hemispherical and bidirectional top-of-canopy (TOC) reflectance in the 400–2500 spectral range at 1 nm as output, i.e. 2101 spectral bands.

A look-up table (LUT) of 500 samples was generated by means of Latin hypercupe sampling (McKay et al., 1979) within the PROSAIL variable space with minimum and maximum boundaries of vegetation properties as given in Table 3. The LUT size is considered as an acceptable trade-off between sufficiently sampling the parameter space while keeping the sampling size low enough to enable fast processing. All the PROSPECT-4 leaf variables have been ranging, whereas regarding SAIL only the vegetation properties, i.e. LAD and LAI have been ranging. Finally, only LAI was retrieved from bi-directional TOC reflectance data based on the synergistic use of DR and MLRA methods.

3.2. Field and HyMap data

The second dataset involves an experimental dataset with real spectral data. The widely used SPARC dataset (Delegido et al., 2013) was chosen to evaluate the performances of the SIMFEAT-MLRA retrieval strategies. The SPectra bARrax Campaign (SPARC) field dataset encompasses different crop types, growing phases, canopy geometries and soil conditions. The SPARC-2003 campaign took place from 12 to 14 July in Barrax, La Mancha, Spain (coordinates 30°3′N, 28°6′W, 700 m altitude). Bio-geophysical parameters have been measured within a total of 108 Elementary Sampling Units (ESUs) for different crop types (garlic, alfalfa, onion, sunflower, corn, potato, sugar beet, vineyard and wheat). An ESU refers to a plot, which is sized compatible with pixel dimensions of about 20 m × 20 m. In the analysis no differentiation between crops was made. Green LAI has been derived from canopy measurements made with a LiCor LAI-2000 digital analyzer. Each ESU was assigned one LAI value, obtained as a statistical mean of 24 measurements (8 data readings × 3 replica) with standard errors ranging from 5% to 10% (Fernández et al., 2005). LAI values ranged between 0.4 and 6.2 m^2/m^2. During the campaign, airborne hyperspectral HyMap flight-lines were acquired for the study site, during the month of July 2003. HyMap flew with a configuration of 125 contiguous spectral bands, spectrally positioned between 430 and 2490 nm. Spectral bandwidth varied between 11 and 21 nm. The pixel size at overpass was 5 m. The flight-lines were corrected for radiometric and atmospheric effects according to the procedures of Alonso and Moreno (2005) and Guanter et al. (2005). Finally, a calibration dataset was prepared, referring to the pixel that covers the centre point of each ESU and its corresponding LAI values. Additionally 20 bare soil spectra were added.

3.3. Experimental setup

The pursued analysis for the PROSAIL and the HyMap dataset was alike. First, all MLRA methods are run with a 4-fold ($k = 4$) CV sampling scheme without a DR method (i.e. full spectral data), then with the classical PCA, and then with the alternative DR methods. To start with 5 components are used in the regression analysis. The performances are compared both in terms of accuracy (R^2_{CV}, RMSE_{CV}, NRMSE_{CV}) and processing speed. Following, taking the LR as reference and best performing MLRA, the performances of all DR methods along an increasing number of components are compared, from 1 to 10 components. Additionally, to link between the noise-free simulated spectral data and HyMap spectral data that is inherently noisy, an exercise of adding Gaussian noise to the simulated data has been applied. This is not trivial since simulated data is free from any noise, thus highly collinear and perfectly suited for applying DR methods. Conversely, in case of real sensor data all kinds of noises (e.g. instrumental, environmental) may occur, which may impact the performances of the DR methods. Multiple noise levels, i.e. 0.1, 0.5, 1 and 5 % Gaussian noise have been injected to the PROSAIL simulated spectra, and the DR + MLRA analyses have been repeated for 5 components. The performances of the DR methods are then compared against the previous noise-free 5 components results. Finally, any of the developed DR-MLRA models can be applied to an remote sensing imagery to process it into a LAI map given the same band settings as those presented during the training phase. LAI maps have been created to an arbitrary subset of HyMap flight line using GPR with all bands and GPR with best performing DR method. The advantage of using GPR is that this method provides additional uncertainty estimates. The lower the σ the more confident the retrieval relative to what has been presented during the training phase. Hence, a direct quantitative measure of the mapping performances is provided. All processing was done within ARTMO on a contemporary computer (Windows-64 OS, i7-4790 CPU 3.60 GHz, 16 GB RAM).

Table 3
Range and distribution of input variables used to establish the PROSAIL (PROSPECT4 + SAIL) look-up table.

Model variales		Units	Minimum	Maximum
Leaf variables: PROSPECT-4				
N	Leaf structure index	unitless	1.3	2.5
LCC	Leaf chlorophyll content	[μg/cm^2]	1	80
C_m	Leaf dry matter content	[g/cm^2]	0.002	0.05
C_w	Leaf water content	[cm]	0.002	0.05
Canopy variables: SAIL				
LAI	Leaf area index	[m^2/m^2]	0.1	7
α_{soil}	Soil scaling factor	unitless	0.01	0.01
ALA	Average leaf angle	[°]	0	90
HotS	Hot spot parameter	[m/m]	0.01	0.1
skyl	Diffuse incoming solar radiation	[fraction]	10	10
SZA	Sun zenith angle	[°]	35	35
VZA	View zenith angle	[°]	0	0
RAA	(Sun-sensor) relative azimuth angle	[°]	0	0

4. Results

4.1. PROSAIL results

The PROSAIL simulation dataset that is built of 8 input variables and 2101 output bands is first analyzed. To put the impact of DR methods into the LAI retrieval scheme into perspective, R^2_{CV} validation results (1) without DR, i.e. using all 2101 bands, (2) using the classical PCA and (3) using the best performing DR methods - both for 5 components - are shown in Fig. 2. Overall, training the regression algorithms with a spectral dataset of 2101 bands led to suboptimal results. While KRR and NN performed reasonable with a R^2_{CV} of 0.76 and 0.75, respectively, the other advanced MLRAs led to poorer R^2_{CV} accuracies, between 0.5 and 0.7, and the ordinary least squares LR method completely failed. Alternatively, converting the 2101 bands into 5 PCA components only improved accuracies for LR (R^2_{CV} from 0.05 to 0.5) and GPR (R^2_{CV} from 0.51 to 0.70). For all the other regression models PCA conversion did instead degrade their predictive power, which suggests that a PCA does not always match well with advanced regression algorithms. More remarkable improvements were achieved when converting the spectral dataset into components by the alternative DR methods, particularly by CCA and OPLS. These two top-performing DR methods perform alike and rely on an intermediate clustering step; the R^2_{CV} reached beyond 0.90 for each of the tested regression algorithm. They also perform more robust, as indicated by a generally narrower SD than PCA or when no DR method applied. Thereby, given that LR reached the same accuracies as the advanced MLRAs, suggests that the excellent results are primarily driven by CCA and OPLS.

Performances are inspected in more detail by plotting the scatter plots for three regression algorithms with best results, i.e. (1) LR, because applying DR methods to LR led to most significant improvements; (2) KRR, because this regression algorithm was best performing without DR methods; and (3) GPR, because the combination of DR + GPR led to highest accuracies (Fig. 3). Scatter plots are first shown for the regression algorithms trained with all 2101 bands, second with first applying a PCA, and third with applying the best performing DR method. The 9 scatter plots reveal the role DR methods are playing on the retrieval of LAI. Table 4 provides associated goodness-of-fit statistics and processing time. Since boundary situations are plotted, from worst (LR alone) to best (CCA-GPR) performances, the performances of the other DR and regression combinations fall within these extremes. The following trends can be observed.

First, the LR model trained with 2101 bands caused some extreme outliers, which makes discarding this model for mapping applications. Second, although KRR deals best with processing the full spectra, the model faces difficulties in coping with the well-known saturation effect, i.e. higher LAI cause little spectral variation and therefore tend to be underestimated (Gao et al., 2000). Third, GPR performs poorer than KRR, leading to saturation effect and a broader point cloud around the 1:1-line.

A PCA spectral transformation improved LR and GPR predictions, but saturation effect remains; especially LR failed to deliver LAI predictions above 5; while GPR and KRR delivered only accurate predictions for a low LAI, i.e. until 2, at higher LAI a saturation effect starts to emerge. Conversely, when the spectral data is first reduced through the best performing DR method (i.e., CCA or OPLS) then the regression algorithms deliver substantially more accurate predictions, even for high LAIs, leading to narrowly distributed point clouds along the 1:1-line. This suggests that these two DR methods are able to overcome LAI saturation irrespective of the used regression method. Relative errors fell well below 10%, which is commonly required by the user community.

Another advantage of using these DR methods is the gain in processing time. Although CCA and OPLS run slower than the classical PCA, for GPR the gain in accuracy and processing speed is remarkable compared to no DR method applied, i.e. 41 times faster and a drop in relative errors from 20.3 to 6.8% (see Table 4).

The performances of the DR methods along an increasing number of components from 1 to 10 are subsequently inspected. R^2_{CV} validation results for all DR methods in combination with LR and the advanced GPR are shown in Fig. 4. The trends for LR and GPR are consistent and can be summarized as follows. CCA and OPLS deliver substantially higher accuracies than any of the other DR methods, especially for LR. Remarkably, the use of one component already yields high accuracies, and flattening starts from three components with a R^2_{CV} above 0.9. LR trained with only two OPLS or CCA components even leads to superior performances as opposed to trained with up to 10 components by any other DR method. Their kernel variants, KOPLS and KCCA, are positioned somewhat lower. In case of LR all the other DR methods perform substantially poorer, i.e., on the R^2_{CV} order of 0.5–0.7. For instance, the conventional PCA underperforms until 7 components and only reaches about the same accuracies as PLS, KPLS and KPCA when

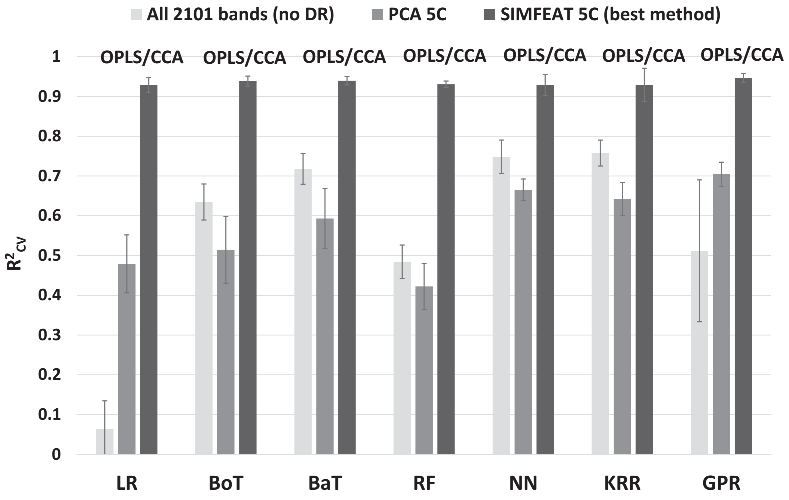


Fig. 2. PROSAIL LAI R^2_{CV} (mean and SD) validation results for directly MLRA, MLRA-PCA 5C (components) and with best performing DR method, given on top.

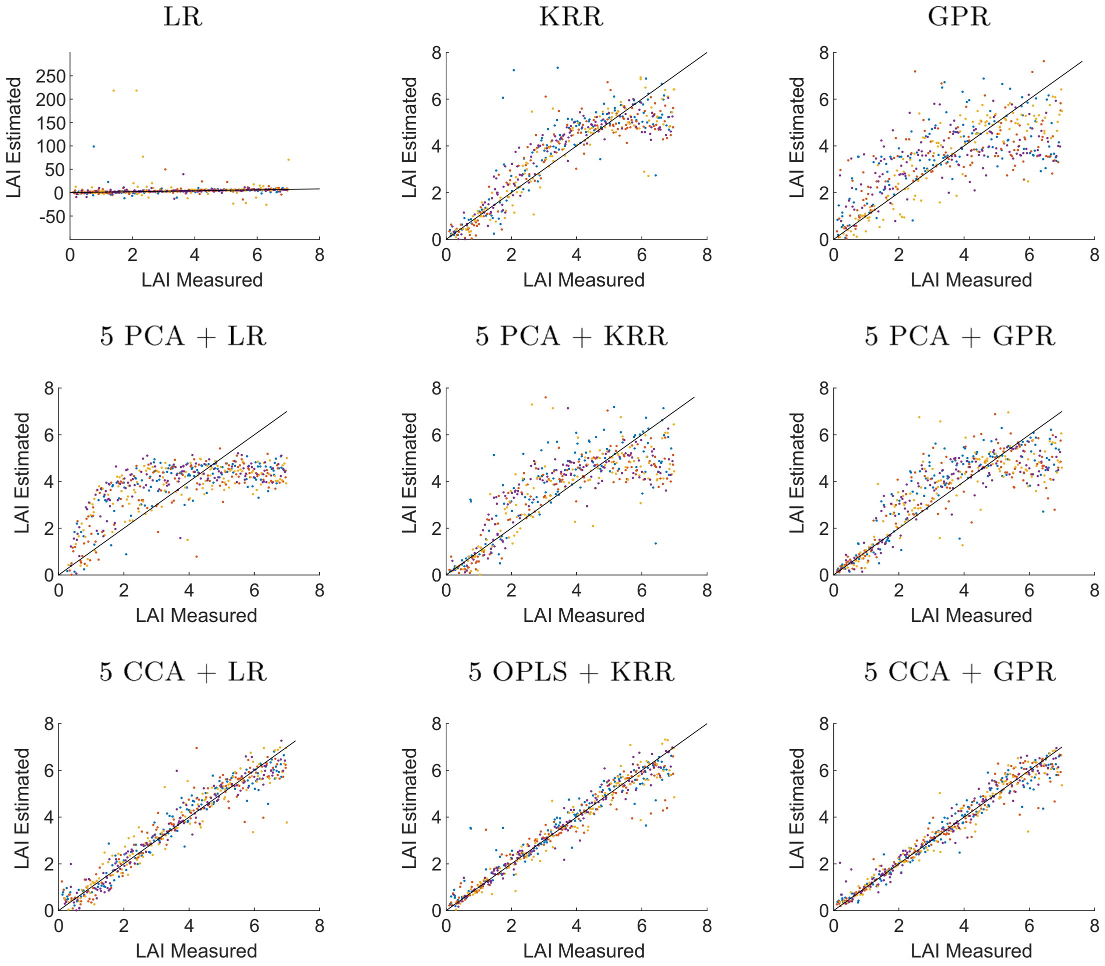


Fig. 3. PROSAIL LAI measured vs estimated scatterplots for LR, KRR and GPR, without DR [top], with 5 PCA [middle] and with best performing 5C DR method [bottom]. The colors represent the 4-*k* subsets. For LR the Y-axis has not been synchronized to enable viewing the full extent of its estimates.

Table 4
Cross-validation statistics and processing time for the results presented in Fig. 3.

MLRA	R^2_{CV} (SD)	RMSE$_{CV}$ (SD)	NRMSE$_{CV}$ (SD) (%)	CPU (SD) (s)
All PROSAIL bands (2010)				
LR	0.06 (0.07)	12.23 (11.7)	179.7 (171.6)	0.4 (0.1)
KRR	0.76 (0.03)	1.00 (0.07)	14.7 (1.0)	7.2 (1.2)
GPR	0.51 (0.18)	1.38 (0.29)	20.3 (4.3)	1077.8 (46.0)
5 PCA				
PCA-LR	0.53 (0.06)	1.37 (0.07)	20.2 (1.1)	0.2 (0.2)
PCA-KRR	0.64 (0.04)	1.21 (0.03)	17.9 (0.5)	1.0 (0.2)
PCA-GPR	0.72 (0.04)	0.81 (0.04)	15.5 (0.7)	3.2 (0.2)
Best performing 5C DR method				
CCA-LR	0.93 (0.02)	0.53 (0.06)	7.8 (2.6)	20.3 (0.9)
OPLS-KRR	0.93 (0.04)	0.52 (0.14)	7.6 (1.9)	22.6 (2.1)
CCA-GPR	0.95 (0.01)	0.43 (0.02)	6.3 (0.2)	25.7 (1.9)

trained by 10 components. In turn, when an advanced, nonlinear regression algorithm such as GPR is used, then PCA or PLS and their kernelized versions KPCA and KPLS behave more alike, meaning that the nonlinear aspect of the kernel DR methods play a less important role. Also, these methods reach the same order of accuracies as KCCA and KOPLS when trained with 10 components.

Overall, both LR and GPR results suggest that most gain in accuracy is achieved due to CCA or OPLS. Thereby, although not shown for the sake of brevity, the same trends were observed for the other MLRAs.

4.2. Assessing the robustness of DR methods to noise

The DR methods are next evaluated on their ability to deal with noisy data. Performances are compared against the earlier noise-free results. R^2_{CV} validation results for LR and GR are shown in Fig. 5; the other regression methods were behaving alike (not shown). The earlier promising CCA and OPLS methods tend to respond most sensitive to the introduction of Gaussian noise. A small injection of 0.1% noise to the spectra already breaks down their superior performances. This especially holds for GPR where

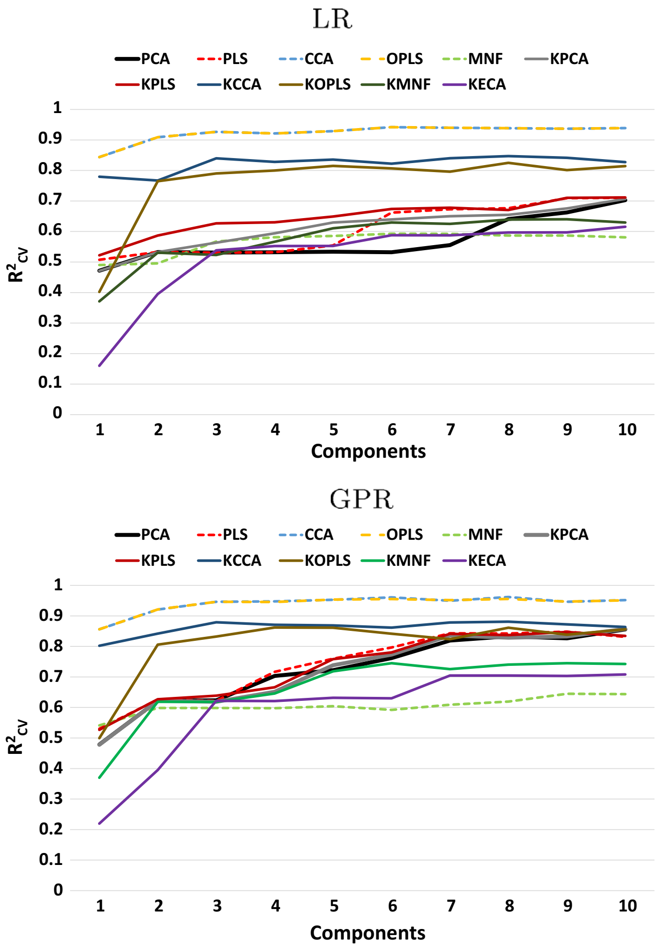


Fig. 4. PROSAIL LAI R^2_{CV} (mean) validation results for each of the DR methods for LR [top] and GPR [bottom] along increasing number of components, from 1 to 10.

relationships rapidly degrade when adding more noise. In turn, the kernel versions of CCA and OPLS (i.e., KCCA and KOPLS) appear to be considerably much more robust to noise and particularly LR maintains excellent results until 5% noise with an R^2_{CV} almost reaching 0.8. GPR, however, does not manage to keep the high accuracies with KCCA and KOPLS and degradation takes place at 5% noise, which suggests tendency of overfitting. Also the mid-range performing DR methods PCA and PLS and the kernel variants KPCA and KPLS tend to cope well with noisy spectral data; the injection of noise had negligible impact. These methods perform quite robust in combination with GPR; R^2_{CV} results maintain above 0.7 and also the SD is kept small. The remaining DR methods perform somewhat poorer (KMNF, KECA) or even failed when data becomes more noisy (MNF + GPR), which makes these methods less attractive for mapping applications. To ascertain whether the number of components play any role in dealing with noisy data, the noise exercise was repeated with 5% noise and 10 components (results not shown). About the same trends were observed as with 5 components, i.e. the R^2_{CV} increased only marginally, in agreement with Fig. 4. This suggests that the conducted noise experiment is valid to derive some general trends from it. The bottom line here is that OPLS and CCA are only excelling in case of noise-free contiguous spectral data such as simulated TOC reflectance data. But when spectral data start to becomes noisy then these relationships degrade rapidly, especially when trained by GPR. This suggests that the degree of noisiness determines the performances of these DR methods, which bears consequences when applying DR methods to real sensor data, as addressed in the next section.

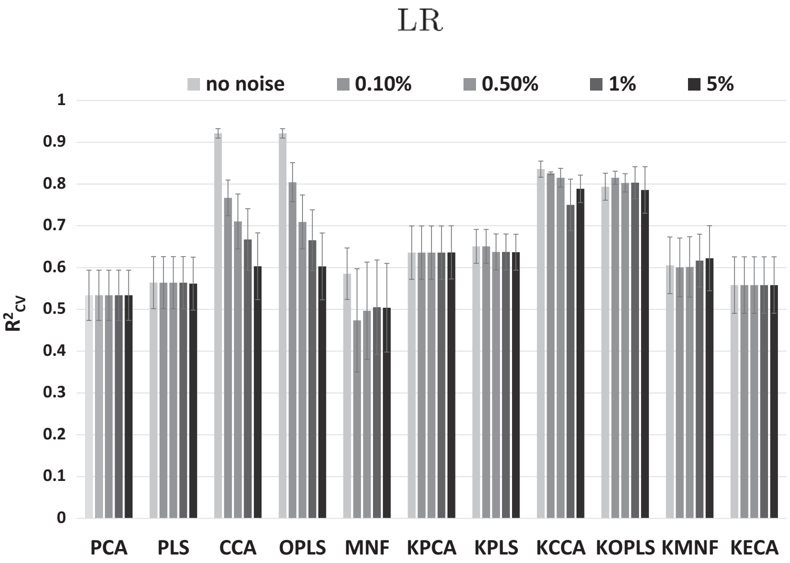
4.3. HyMap results

The analysis has been repeated to an experimental dataset consisting of field LAI measurements and associated airborne hyperspectral measurements as obtained by the HyMap sensor. Similar as before, (1) the MLRA validation R^2_{CV} results without using DR methods, (2) using the classical PCA and (3) with best-performing DR method - the latter two with 5 components - are first shown (Fig. 6). The best performing DR method is displayed on top of the bars, although, as shown in next section, several DR methods perform alike. For LR, KRR and GPR goodness-of-fit statistics and processing speed are provided in Table 5. It leads to the following findings. LR gained mostly from applying a DR step prior to the regression as compared to using all bands. While the PCA improved results from 0.47 (no DR) to 0.92, by using KCCA as DR method R^2_{CV} improved to 0.88. The gain in accuracy by the alternative DR methods is less obvious for the more advanced MLRAs; R^2_{CV} differ only slightly across the three strategies. While bagging trees and random forests benefitted somewhat from a kernel DR method, and PLS is the preferred method for NN, KRR and GPR, overall the gain as compared to PCA is minimal. KRR and GPR yielded even best results when using all bands, although the differences in accuracy as compared to combining with PLS is small: R^2_{CV} 0.94 vs. 0.93, respectively (see also Table 5). When also considering processing speed, then DR methods make the difference. Although the regression algorithms ran fast because of trained by relatively few samples and bands, applying DR methods still caused a substantial acceleration. For instance, PLS-GPR processed model development and validation about 8 times faster than GPR alone.

The performances of the applied DR methods along an increasing number of components are subsequently inspected. R^2_{CV} validation results for all DR methods in combination with LR and GPR are shown in Fig. 7. The main trends can be summarized as follows. None of the tested DR methods act distinctly outstanding but when using LR then KCCA is top performing from 2 components onwards. PCA, PLS and their kernel variants behave alike, just below KCCA, and reach almost the same accuracies when trained with 10 components. When instead using GPR then PLS, PCA and KPCA are top performing and behave alike from 4 components onwards; at 10 components then also KPLS, KCCA, KOPLS reach about the same accuracies. Conversely, CCA and OPLS are only mid-ranging to poorly performing. Hence, these results underline again that, while their kernel variants proved to respond considerably more adaptive, CCA and OPLS face difficulties in coping with more noisy data.

To exemplify the mapping of LAI, an arbitrary subset of HyMap flight line was twice processed using a GPR model: (1) without a DR method, and (2) in combination with 5 components of the best performing DR method, i.e. PLS. The conversion of the HyMap subset to an LAI map using GPR alone took 73 s. When instead using the PLS-GPR model then processing time reduced to only 5 s. Inspection of the LAI maps (Fig. 8, top) reveals the contrast between the irrigated circular parcels with high LAI and the surrounding fallow land. These maps are in agreement with earlier mapping approaches (Verrelst et al., 2012a; Rivera et al., 2014a). Note that LAI mean (μ) estimates vary along the two maps, especially for low LAI on senescent, non-irrigated parcels. For instance, the PLS-GPR model is considerably better adapted to assign close-to-zero values to the fallow lands. In contrast, the LAI map as obtained by GPR alone looks more heterogeneous, which suggests that the model had more difficulty dealing with the spatial variability of the image.

Uncertainty maps are provided in Fig. 8 (bottom). When GPR uses all bands, the uncertainties tend to vary more, with especially high, patchy uncertainties over the harvested or non-irrigated



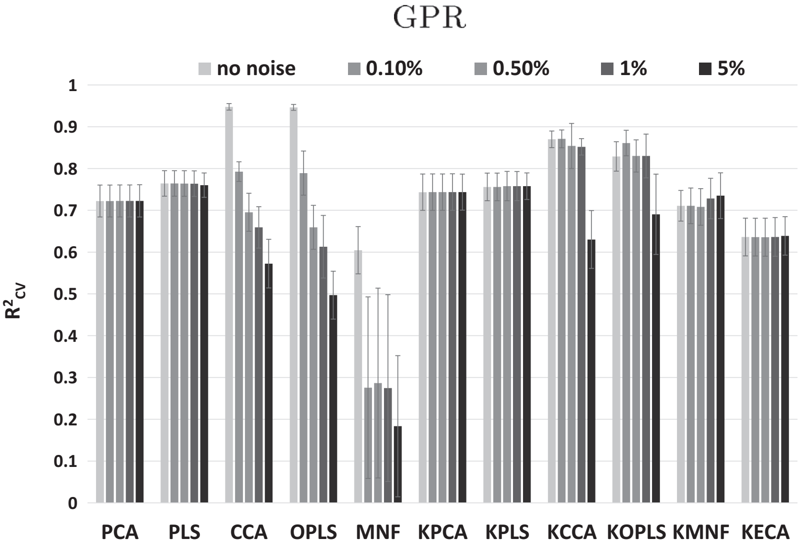


Fig. 5. PROSAIL LAI R^2_{CV} (mean and SD) validation results for each of the DR methods for LR [top] and GPR [bottom] for no noise and increasing Gaussian noise levels.

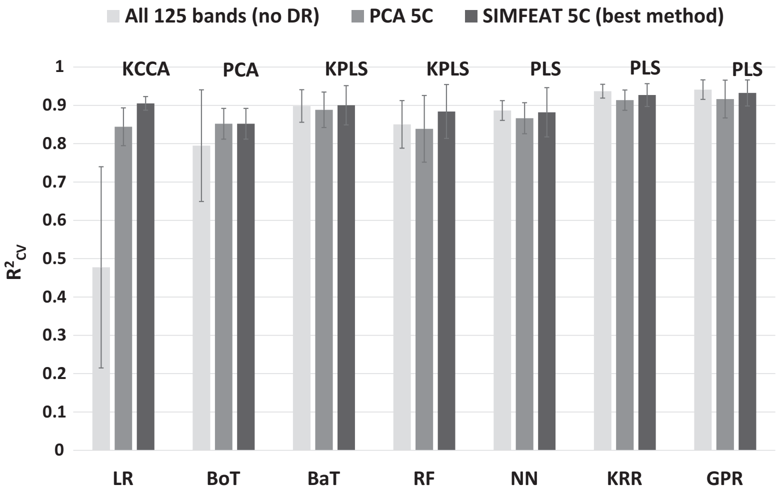


Fig. 6. HyMap LAI R^2_{CV} (mean and SD) validation results for directly MLRA, MLRA-PCA 5C (components) and with best performing DR method, given on top.

Table 5
Cross-validation statistics and processing time for the results presented in Fig. 6.

MLRA	R^2_{CV} (SD)	RMSE$_{CV}$ (SD)	NRMSE$_{CV}$ (SD) (%)	CPU (SD) (s)
All HyMap bands (125)				
LR	0.47 (0.07)	1.41 (0.47)	24.9 (7.3)	0.2 (0.4)
KRR	0.94 (0.02)	0.44 (0.02)	7.8 (0.8)	0.0 (0.0)
GPR	0.94 (0.02)	0.39 (0.06)	7.0 (1.4)	3.9 (0.0)
5 PCA				
PCA-LR	0.84 (0.04)	0.70 (0.04)	12.6 (1.7)	0.0 (0.0)
PCA-KRR	0.91 (0.03)	0.51 (0.04)	9.1 (1.0)	0.1 (0.0)
PCA-GPR	0.92 (0.05)	0.48 (0.07)	8.6 (1.7)	0.5 (0.0)
Best performing 5C DR method				
KCCA-LR	0.92 (0.02)	0.52 (0.06)	9.3 (2.1)	0.7 (0.0)
PLS-KRR	0.93 (0.03)	0.47 (0.08)	8.4 (1.8)	0.1 (0.0)
PLS-GPR	0.93 (0.04)	0.43 (0.03)	7.8 (1.8)	0.5 (0.0)

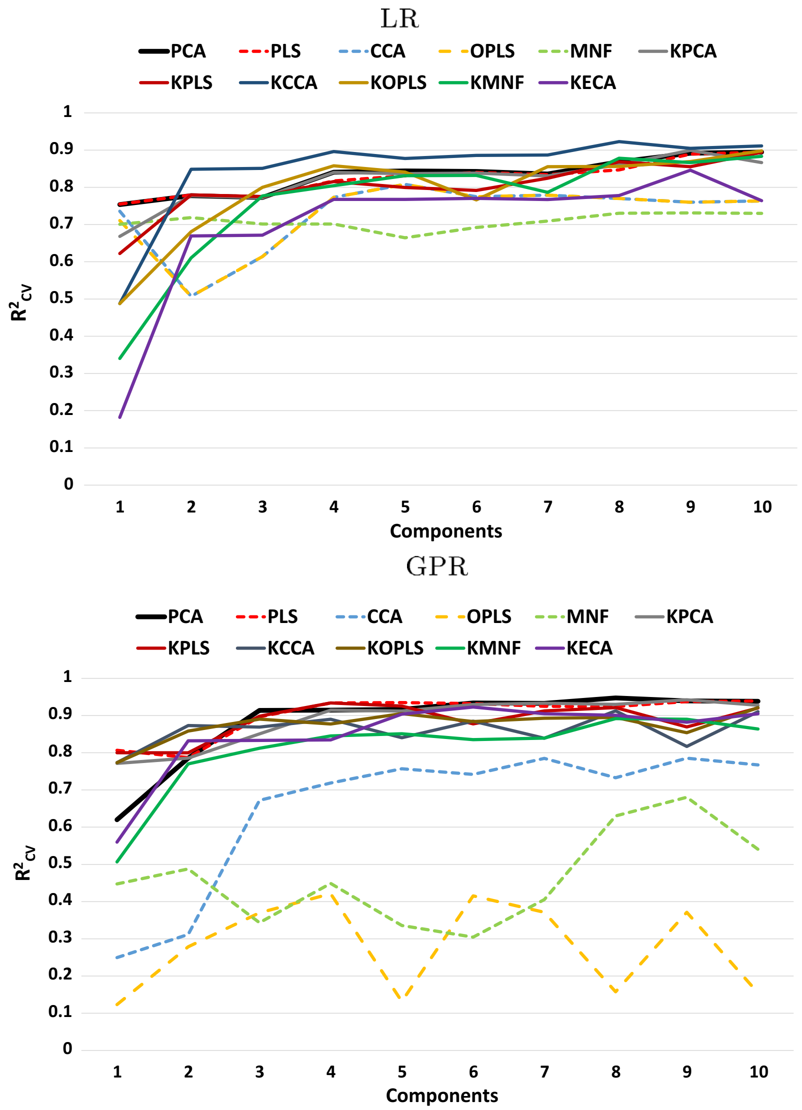


Fig. 7. HyMap R^2_{CV} (mean) validation results for each of the DR methods for LR [top] and GPR [bottom] along increasing number of components, from 1 to 10.

drylands. Low uncertainties are encountered on the green irrigated fields, which can be attributed to the applied sampling design that predominantly focused on crops in vegetative status (Verrelst et al., 2013). Hence, the uncertainty maps of both retrieval algorithms can be compared to derive conclusions about the processing quality of a GPR model. To quantify the extent of reduced uncertainties, the absolute and relative differences between GPR and PLS-GPR maps are mapped in Fig. 9. These maps are dominated by shades of blue, which indicates that for most of the land covers uncertainties are systematically reduced by the PLS-GPR model, mostly on the order of 50%. This thus suggests that training the regression algorithm with components from a DR method not only speeds up image processing but also leads to more certain estimates and thus higher quality maps.

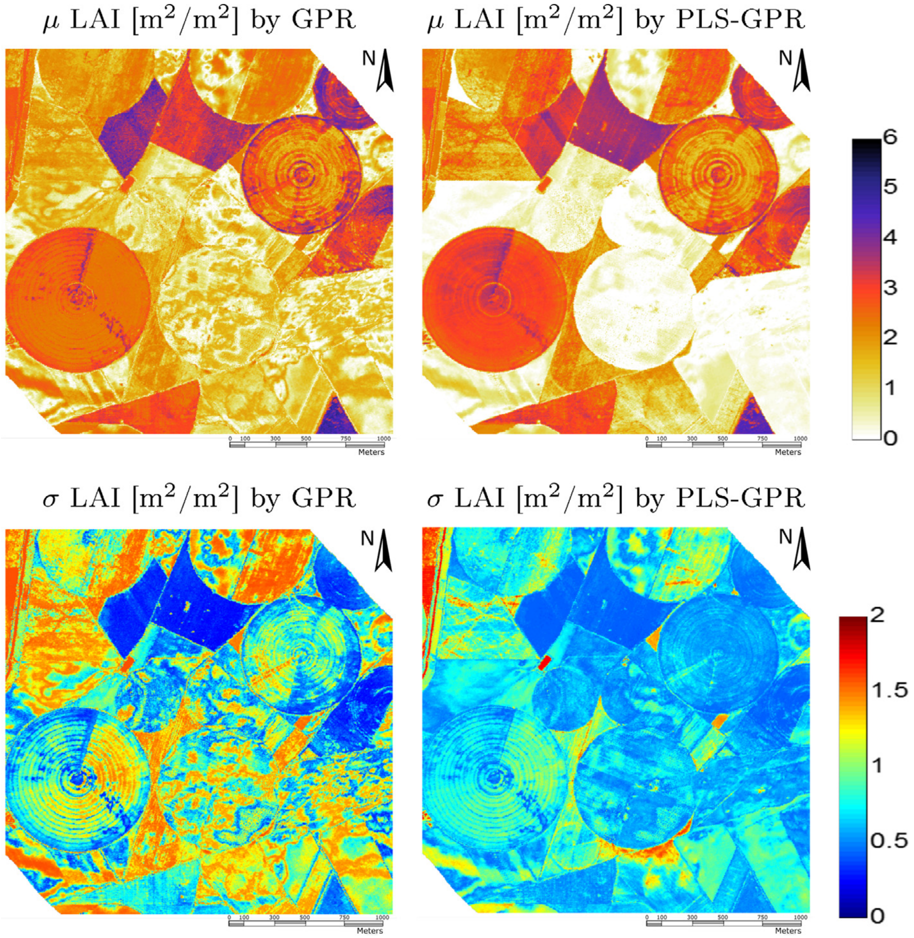


Fig. 8. HyMap LAI map [m^2/m^2] processed by GPR using all 125 bands (top left), LAI map processed by PLS-GPR using 5 components (top right), associated GPR uncertainty estimates (σ), respectively (bottom).

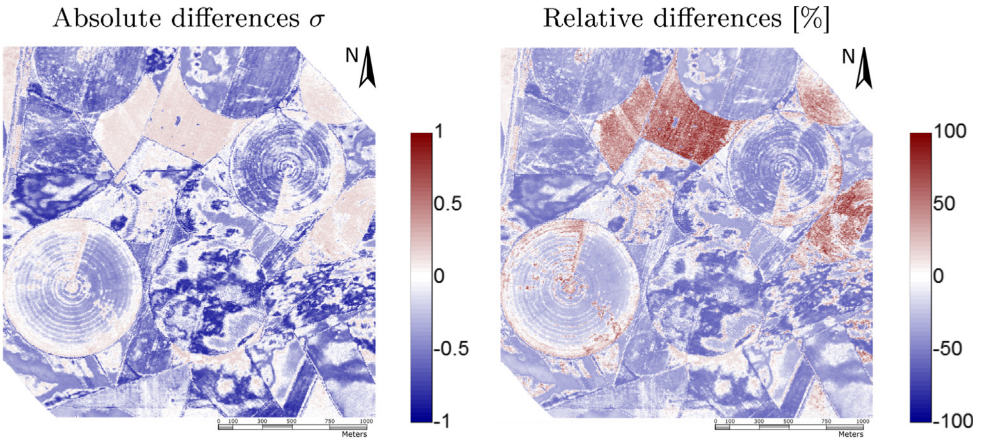


Fig. 9. Differences in SD (σ) between GPR all 125 bands and PLS-GPR in terms of absolute (left) and relative (right) differences.

5. Discussion

The application of MLRAs to hyperspectral data mining and analysis in the area of retrieval algorithms is rapidly gaining interest in the community (Verrelst et al., 2015a). However, the large number of (collinear) bands hampers the development of accurate and fast MRA models, and DR methods that reduce the complexity without loss of information become strictly necessary. In this work, we have demonstrated that simple linear and nonlinear DR methods as brought together into one library (SIMFEAT) can be combined with MLRAs for the quantification of biophysical variables. These methods are implemented in ARTMO's MLRA toolbox as a two-step workflow: first the spectral data is converted into components with a DR method, and second regression is applied over the components. The analysis of DR + MLRA ensembles makes this toolbox powerful in a wide range of mapping applications, especially in view of processing imaging spectroscopy data. To illustrate its utility, we have analyzed the predictive power of DR + MLRA ensembles both to simulated PROSAIL (2101 bands) and experimental HyMap (125 bands) data. The following general findings are briefly discussed.

First, the PROSAIL dataset demonstrated that the classical PCA is not always the best choice when combining with machine learning regression algorithms. In fact, superior accuracies were achieved for the majority of tested MLRAs without a PCA conversion. Nevertheless, most of the alternative DR methods did not perform much better, and also the MLRAs without using DR methods, despite being adaptive, did not lead to excellent accuracies (at best with KRR R^2_{CV} of 0.76). Two reasons can be identified why the MLRAs alone or with PCA faced difficulties with this dataset: (1) the simulated spectral dataset is complex in a sense that is not only driven by LAI, but also by 7 other PROSAIL input variables, which makes that the other variables confound the LAI relationships; (2) the spectral response to LAI is well known to reach a saturation around medium LAI (i.e., LAI of 3) (Gao et al., 2000). The saturation effect was also observed when plotting the scatter plots for models trained with all bands and even when applying a PCA. Remarkably only the DR methods that additionally decomposes the input variables into clusters, i.e. CCA and OPLS, dealt excellently with the multi-dimensional simulated dataset and largely resolved the saturation problem (R^2_{CV} around 0.94). Consequently, first clustering the dataset, i.e. discretizing the output space to k dimensions, and then projecting to a lower dimensional space before proceeding with regression proved to be a successful method in dealing with such simulated data. Although here only a LUT of 500 samples was used, and additional testing with larger LUT sizes is required, this proposed approach opens opportunities towards new generation of hybrid retrieval strategies that are based on advanced machine learning methods, and are applicable to imaging spectroscopy data.

Second, a drawback of CCA and OPLS is that these methods tend to be sensitive to spectral noise. Relationships started to break down already when injecting a tiny bit of noise to the PROSAIL dataset. A solution to make these methods more robust to noise would be to introduce a regularization term (Nielsen et al., 1998), although that goes along with additional tuning. On the other hand, when instead using the kernel version of CCA and OPLS, i.e. KCCA and KOPLS, then these methods responded more robust to noise, making them attractive alternatives to consider in real data applications. When effectively using real (noisy) HyMap data, then not only KCCA and KOPLS but also PCA and PLS and their kernel variants performed more robust than CCA and OPLS, although improvements as compared to no using a DR method are also more modest. PCA and PLS do not consider the separability of classes to generate a lower dimensional representation of original data. That PCA remains an attractive method for hyperspectral data when including sufficient components has been observed before. Martinez and Kak (2001) earlier noted that PCA can outperform other methods when the number of training samples is limited, and also, PCA has less sensitivity to different training datasets.

Noteworthy hereby is that in case of linear regression (LR) it was not PCA or PLS that were best performing (R^2_{CV} around 0.84). This deserves special attention because PCA and PLS in combination with LR are widely applied in remote sensing mapping applications in the form of PCR and PLSR (see Verrelst et al. (2015a) for review). On the other hand, the HyMap results revealed that PLS matched best with neural networks and the kernel machine learning methods KRR and GPR. For instance, PLS in combination with KRR or GPR yielded a R^2_{CV} of 0.93. This suggests that the classical PLSR formulation delivers rather suboptimal results as compared to when combining with non-linear regression algorithms. It also implies that new opportunities are opened to reach more accurate mapping applications by exploiting DR methods in combination with MLRAs. On the other hand, it does not escape our attention that GPR and KRR were more successful in processing HyMap data without using a DR method (R^2_{CV}: 0.93). This can be attributed to the versatility of these advanced methods, but also to the nature of the HyMap dataset. The dataset consists of relatively few samples, i.e. 118, and relatively few bands, i.e. 125. Considering that real data is not free from noise implies that the problem of collinearity is less prevailing to this experimental airborne dataset. Effectively, Rivera et al. (2014a) earlier found that NN, KRR and GPR tend to cope well with experimental airborne and spaceborne hyperspectral datasets. But it goes at a computational cost that can be alleviated by combining with an appropriate DR method. Moreover, these hyperspectral datasets were rather small with at most 125 HyMap bands. When moving towards new-generation imaging spectrometers equiped with a few hunderd spectral bands (e.g., FLEX, ENMAP,HySPIRI, PRISMA) then DR methods become indispensable in their data processing by advanced statistical regression methods.

Another widely used DR method applied to hyperspectral data involves MNF. MNF analysis minimizes the noise fraction, or equivalently, maximizes the signal-to-noise ratio of linear combinations of zero-mean variables (Green et al., 1988). However, MNF was in none of the tested cases evaluated as an attractive candidate to be combined with regression analysis, although it is to be noted that its kernel version seems more promising.

Third, the nonlinear kernel variants of the DR methods triggered mostly improvements but did not always lead to superior results. Their utility largely depended on the nature of the applied regression algorithm. On the one hand, for LR the kernel DR methods substantially improved accuracies. This was most noticeable for LR with KCCA or KOPLS given noisy simulated data or real (noisy) HyMap data. The success of combining a nonlinear DR with LR was to be expected, since LR does nothing more than an ordinary least squares regression. Hence, when combining LR with a nonlinear DR method makes the regression algorithm more adaptive and fast in processing. Conversely, when a nonlinear DR method is combined with an advanced, nonlinear MLRA then improvements in accuracies were less obvious. For instance, in case of HyMap data as processed by GPR then KCCA, KOPLS and KMNF outperformed their non-kernel versions, but accuracies were on the same order as combined with PCA and PLS, especially when trained with enough components (e.g. 10). It suggests that a nonlinear DR method does not provide much added value when already a powerful nonlinear regression model is used. The little to no gain achieved by combining two kernel methods can be explained by the tendency to overfitting and implies extra regularization efforts to alleviate numerical instabilities (Shawe-Taylor and Cristianini, 2004). A second important problem is related to the computational cost. Since $\mathbf{K}_x$ is of size $l \times l$, begin l the number of training samples, the complexity of the methods scales quadratically with l in terms of memory, and cubically with respect to the computation time. Further, the solution of the maximization problem is not sparse, so the feature extraction for new data requires the evaluation of l kernel functions per pattern, becoming computationally expensive for large l. The opposite situation is worth mentioning: when l is small, the extracted features may be useless, especially for high-dimensional data (Abrahamsen and Hansen, 2011). To address these problems, several solutions can be devised: either search for sparse models in which one expresses the solution as a combination of a subset of optimized training data (Arenas-Garcia et al., 2013), or by approximating the kernel functions with randomized versions

(Perez-Suay et al., 2017). These alternatives are being considered to be implemented into the MLRA toolbox.

Finally, the predictive power of DR + MLRA ensembles can also be appreciated when comparing against more conventional mapping methods such as vegetation indices. In a similar paper using the same HyMap dataset, all possible two-band combinations was fitted using various fitting functions (e.g., linear, polynomial, exponential, power). From all those possible combinations, optimized regression models were at best validated with an R^2 of 0.83 (Rivera et al., 2014b). At the same time, an attractive advantage of applying a DR prior to an advanced regression algorithm is the gain in processing speed. The HyMap exercises revealed that processing speed accelerated about 14 times with PLS-GPR as compared to GPR alone. This is not trivial, particularly in view of developing new-generation hybrid algorithms for operational processing of imaging spectroscopy data. The question now arises whether DR + MLRA ensembles can be successfully applied to retrieve other vegetation properties. While LAI is known to have a broad sensitive spectral response, and therefore relevant information can be successfully captured by lower-dimensional components, other biophysical variables may have a more narrow sensitive spectral response. For instance, the absorption region of leaf chlorophyll content (LCC) is restricted to the visible and the red edge, e.g. see Verrelst et al. (2015c). In this respect, for variables that are only sensitive to specific absorption regions, the question arises whether the DR methods will be able to capture the relevant information into their main components. In principle such analysis can be easily done by ARTMO's MLRA toolbox - this is one of the research lines we will explore in future works.

6. Conclusions

We present an evaluation of a proposed statistical biophysical variable retrieval workflow implemented into ARTMO's machine learning toolbox. The approach consists of extracting features (components) from spectral data that are then fed to linear or nonlinear machine learning regression models. We evaluated a library consisting of 11 dimensionality reduction (DR) methods and 8 machine learning regression algorithms (MLRAs). These DR methods enable to reduce the numbers of bands largely while preserving desired intrinsic information of the data. The combination of DR with regression can be powerful for biophysical variable retrieval, as it leads to an accessible feature space of lower dimensionality, which in turn leads to solving simpler, faster convex optimization problems, and eventually results into a more efficient retrieval algorithms. This ensemble approach is especially attractive for processing hyperspectral datasets, typically characterized by a large amount of redundant bands. To demonstrate their predictive power, we applied DR + MLRA ensembles to a PROSAIL simulated dataset consisting of 2101 bands. Training regression algorithms with inputs from CCA or OPLS outperformed any other DR method, or when using directly all bands irrespective of the regression method used. This clearly demonstrates the asset of having a DR step integrated into a hybrid retrieval strategy. However, when shifting towards more real (noisy) sensor data, e.g. as tested here with hyperspectral HyMap data (125 bands), then these DR methods no longer excelled. Instead the kernel version of CCA (i.e., KCCA) led to excellent accuracies when applying linear regression (R^2_{CV} of 0.92 as opposed to 0.47 without a DR method). When combining DR methods with nonlinear MLRAs, then the classical PCA or PLS methods were top performing in terms of accuracy and processing speed. The nonlinear kernelized DR methods hardly led to accuracy improvements, which suggests that combining a linear DR method with a nonlinear regression algorithm would be a first choice to convert hyperspectral data into estimates of biophysical variables. LAI maps with associated uncertainties were generated from an HyMap subset using Gaussian processes regression (GPR) as trained with (1) all bands, and (2) with components coming from PLS. Applying a DR method to a GPR model not only accelerated processing speed, but also systematically reduced mapping uncertainties. In conclusion, when dealing with hyperspectral data we recommend to test ensembles of dimensionality reduction and regression strategies to enable optimizing biophysical variable mapping in terms of accuracy and processing speed.

Acknowledgements

This paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under projects TIN2012-38102-C03-01, ESP2013-48458-C4-1-P, and TIN2015-64210-R, and the European Space Agency under project ESA ESTEC 4000105078/11/NL/AF. Gustau Camps-Valls was supported by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423). We thank the reviewers for their valuable suggestions.

References

Abrahamsen, T., Hansen, L., 2011. A cure for variance inflation in high dimensional kernel principal component analysis. J. Mach. Learn. Res. 12, 2027–2044.

Alonso, L., Moreno, J., 2005. Advances and limitations in a parametric geometric correction of CHRIS/PROBA data. In: Proceedings of the 3rd CHRIS/Proba Workshop, ESA/ESRIN, Frascati, Italy.

Arenas-Garcia, J., Petersen, K., Camps-Valls, G., Hansen, L., 2013. Kernel multivariate analysis framework for supervised subspace learning: a tutorial on linear and kernel multivariate methods. IEEE Signal Process. Mag. 30, 16–29.

Baret, F., Hagolle, O., Geiger, B., Bicheron, P., Miras, B., Huc, M., Berthelot, B., Niño, F., Weiss, M., Samain, O., Roujean, J., Leroy, M., 2007. LAI, fAPAR and fCover CYCLOPES global products derived from VEGETATION. Part 1: Principles of the algorithm. Remote Sens. Environ. 110, 275–286.

Baret, F., Weiss, M., Lacaze, R., Camacho, F., Makhmara, H., Pacholczyk, P., Smets, B., 2013. GEOV1: LAI and FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part1: Principles of development and production. Remote Sens. Environ. 137, 299–309.

Borga, M., Landelius, T., Knutsson, H., 1997. A unified approach to PCA, PLS, MLRA and CCA.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Camps-Valls, G., Gómez-Chova, L., Muñoz-Marí, J., Lázaro-Gredilla, M., Verrelst, J., 2013. simpleR: A simple educational Matlab toolbox for statistical regression. <http://www.uv.es/gcamps/code/simpleR.html>. v2.1.

Crawford, M., Tuia, D., Yang, H., 2013. Active learning: any value for classification of remotely sensed data? Proc. IEEE 101, 593–608. http://dx.doi.org/10.1109/JPROC.2012.2231951.

Damodaran, B.B., Nidamanuri, R.R., 2014. Assessment of the impact of dimensionality reduction methods on information classes and classifiers for hyperspectral image classification by multiple classifier system. Adv. Space Res. 53, 1720–1734.

Delegido, J., Verrelst, J., Meza, C., Rivera, J., Alonso, L., Moreno, J., 2013. A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. Eur. J. Agron. 46, 42–52.

Drusch, M. et al., 2016. The fluorescence explorer mission concept-ESA's Earth explorer 8. IEEE Trans. Geosci. Remote Sens., 1–12

Feilhauer, H., Asner, G.P., Martin, R.E., 2015. Multi-method ensemble selection of spectral bands related to leaf biochemistry. Remote Sens. Environ. 164, 57–65.

Feret, J.B., François, C., Asner, G.P., Gitelson, A.A., Martin, R.E., Bidel, L.P.R., Ustin, S.L., le Maire, G., Jacquemoud, S., 2008. PROSPECT-4 and 5: advances in the leaf optical properties model separating photosynthetic pigments. Remote Sens. Environ. 112, 3030–3043.

Fernández, G., Moreno, J., Gandía, S., Martínez, B., Vuolo, F., Morales, F., 2005. Statistical variability of field measurements of biophysical parameters in SPARC-2003 and SPARC-2004 campaigns. In: Proceedings of the SPARC Workshop.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28, 337–407.

Gao, X., Huete, A.R., Ni, W., Miura, T., 2000. Optical–biophysical relationships of vegetation spectra without background contamination. Remote Sens. Environ. 74, 609–620.

Gómez-Chova, L., Jenssen, R., Camps-Valls, G., 2012. Kernel entropy component analysis for remote sensing image clustering. IEEE Geosci. Remote Sens. Lett. 9, 312–316. http://dx.doi.org/10.1109/LGRS.2011.2167212.

Green, A.A., Berman, M., Switzer, P., Craig, M.D., 1988. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Trans. Geosci. Remote Sens. 26, 65–74. http://dx.doi.org/10.1109/36.3001.

Guanter, L., Alonso, L., Moreno, J., 2005. A method for the surface reflectance retrieval from PROBA/CHRIS data over land: Application to ESA SPARC campaigns. IEEE Trans. Geosci. Remote Sens. 43, 2908–2917.

Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., Kuester, T., Hollstein, A., Rossner, G., Chlebek, C., Straif, C., Fischer, S., Schrader, S., Storch, T., Heiden, U., Mueller, A., Bachmann, M., Muhle, H., Muller, R., Habermeyer, M., Ohndorf, A., Hill, J., Buddenbaum, H., Hostert, P., van der Linden, S., Leitao, P.J., Rabe, A., Doerffer, R., Krasemann, H., Xi, H., Mauser, W., Hank, T., Locherer, M., Rast, M., Staenz, K., Sang, B., 2015. The EnMAP spaceborne imaging spectroscopy mission for earth observation. Remote Sens. 7, 8830. http://dx.doi.org/10.3390/rs70708830.

Hagan, M.T., Menhaj, M.B., 1994. Training feedforward networks with the marquardt algorithm. IEEE Trans. Neural Networks 5, 989–993.

Haykin, S., 1999. Neural Networks – A Comprehensive Foundation. Prentice Hall.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory 14, 55–63.

Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P., Asner, G., François, C., Ustin, S., 2009. PROSPECT + SAIL models: a review of use for vegetation characterization. Remote Sens. Environ. 113, S56–S66.

Jenssen, R., 2010. Kernel entropy component analysis. IEEE Trans. Pattern Anal. Mach. Intell. 31.

Jolliffe, I., 1986. Principal Component Analysis. Springer Verlag, New York.

Labate, D., Ceccherini, M., Cisbani, A., De Cosmo, V., Galeazzi, C., Giunti, L., Melozzi, M., Pieraccini, S., Stagi, M., 2009. The PRISMA payload optomechanical design, a high performance instrument for a new hyperspectral mission. Acta Astronaut. 65, 1429–1436.

Laparra, V., Malo, J., Camps-Valls, G., 2015. Dimensionality reduction via regression in hyperspectral imagery. IEEE J. Sel. Top. Sign. Proces. 9, 1026–1036.

Lee, J.A., Verleysen, M., 2007. Nonlinear dimensionality reduction.

Liu, H., Motoda, H., 1998. Feature extraction, construction and selection: a data mining perspective.

Liu, K., Zhou, Q.B., Wu, W.B., Xia, T., Tang, H.J., 2016. Estimating the crop leaf area index using hyperspectral remote sensing. J. Integr. Agric. 15, 475–491.

Luo, X.Q., Wu, X.J., 2012. Fusing remote sensing images using a statistical model. Appl. Mech. Mater., 263–266

Luo, X.Q., Wu, X.J., Zhang, Z., 2013. Regional and entropy component analysis based remote sensing images fusion. J. Intell. Fuzzy Syst.

MacKay, D.J., 1992. Information-based objective functions for active data selection. Neural Comput. 4, 590–604.

le Maire, G., François, C., Soudani, K., Berveiller, D., Pontailler, J.Y., Bréda, N., Genet, H., Davi, H., Dufrêne, E., 2008. Calibration and validation of hyperspectral indices for the estimation of broadleaved forest leaf chlorophyll content, leaf mass per area, leaf area index and leaf canopy biomass. Remote Sens. Environ. 112, 3846–3864.

Martinez, A., Kak, A., 2001. Pca versus lda. IEEE Trans. Pattern Anal. Mach. Intell. 23, 228–233. http://dx.doi.org/10.1109/34.908974.

McKay, M., Beckman, R., Conover, W., 1979. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Nielsen, A.A., 2011. Kernel maximum autocorrelation factor and minimum noise fraction transformations. IEEE Trans. Image Process. 20, 612–624. http://dx.doi.org/10.1109/TIP.2010.2076296.

Nielsen, F., Hansen, L., Strother, S., 1998. Canonical ridge analysis with ridge parameter optimization. NeuroImage 7.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag. 2, 559–572.

Perez-Suay, A., Amoros, J., Gómez-Chova, L., Laparra, V., noz Marí, M., Camps-Valls, G., 2017. Randomized kernels for large scale earth observation applications. Remote Sens. Environ. 1, 1.

Qin, S., McAvoy, T., 1992. Non-linear pls modelling using neural networks. Comput. Chem. Eng. 23, 395–411.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. The MIT Press, New York.

Rivera, J., Verrelst, J., Alonso, L., Moreno, J., Camps-Valls, G., 2014a. Toward a semiautomatic machine learning retrieval of biophysical parameters. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (4), 1249–1259.

Rivera, J., Verrelst, J., Delegido, J., Veroustraete, F., Moreno, J., 2014b. On the semi-automatic retrieval of biophysical parameters based on spectral index optimization. Remote Sens. 6 (6), 4924–4951.

Rivera, J., Verrelst, J., Leonenko, G., Moreno, J., 2013. Multiple cost functions and regularization options for improved retrieval of leaf chlorophyll content and LAI through inversion of the PROSAIL model. Remote Sens. 5, 3280–3304.

Roberts, D., Quattrochi, D., Hulley, G., Hook, S., Green, R., 2012. Synergies between VSWIR and TIR data for the urban environment: an evaluation of the potential for the Hyperspectral Infrared Imager (HyspIRI) Decadal Survey mission. Remote Sens. Environ. 117, 83–101.

Rosipal, R., 2010. Nonlinear partial least squares: an overview.

Schaepman, M.E., Ustin, S.L., Plaza, A.J., Painter, T.H., Verrelst, J., Liang, S., 2009. Earth system science related imaging spectroscopy-an assessment. Remote Sens. Environ. 113, S123–S137.

Scholkopf, B., Smola, A., MÃller, K.R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10, 1299–1319.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis.

Suykens, J., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9, 293–300.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., Muñoz-Marí, J., 2011. A survey of active learning algorithms for supervised remote sensingimage classification. IEEE J. Sel. Top. Sign. Proces. 4, 606–617.

Uto, K., Kosugi, Y., Saito, G., 2014. Semi-supervised hyperspectral subspace learning based on a generalized eigenvalue problem for regression and dimensionality reduction. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 2583–2599.

Van Der Maaten, L., Postma, E., Van Den Herik, H., 2009. Dimensionality reduction: a comparative review. J. Mach. Learn. Res. 10 (1), 66–71.

Van Wittenberghe, S., Verrelst, J., Rivera, J.P., Alonso, L., Moreno, J., Samson, R., 2014. Gaussian processes retrieval of leaf parameters from a multi-species reflectance, absorbance and fluorescence dataset. J. Photochem. Photobiol., B 134, 37–48.

Verger, A., Baret, F., Weiss, M., 2008. Performances of neural networks for deriving LAI estimates from existing CYCLOPES and MODIS products. Remote Sens. Environ. 112, 2789–2803.

Verhoef, W., 1984. Light scattering by leaf layers with application to canopy reflectance modeling: The SAIL model. Remote Sens. Environ. 16, 125–141.

Verhoef, W., Jia, L., Xiao, Q., Su, Z., 2007. Unified optical-thermal four-stream radiative transfer theory for homogeneous vegetation canopies. IEEE Trans. Geosci. Remote Sens. 45, 1808–1822.

Verrelst, J., Alonso, L., Camps-Valls, G., Delegido, J., Moreno, J., 2012a. Retrieval of vegetation biophysical parameters using gaussian process techniques. IEEE Trans. Geosci. Remote Sens. 50, 1832–1843.

Verrelst, J., Camps-Valls, G., Muñoz Marí, J., Rivera, J., Veroustraete, F., Clevers, J., Moreno, J., 2015a. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - a review. ISPRS J. Photogramm. Remote Sens., 273–290

Verrelst, J., Dethier, S., Rivera, J.P., Munoz-Mari, J., Camps-Valls, G., Moreno, J., 2016a. Active learning methods for efficient hybrid biophysical variable retrieval. IEEE Geosci. Remote Sens. Lett. 13, 1012–1016. http://dx.doi.org/10.1109/LGRS.2016.2560799.

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J., Camps-Valls, G., Moreno, J., 2012b. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. Remote Sens. Environ. 118, 127–139.

Verrelst, J., Rivera, J., Moreno, J., Camps-Valls, G., 2013. Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval. ISPRS J. Photogramm. Remote Sens. 86, 157–167.

Verrelst, J., Rivera, J., Veroustraete, F., Muñoz Marí, J., Clevers, J., Camps-Valls, G., Moreno, J., 2015b. Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods – a comparison. ISPRS J. Photogramm. Remote Sens., 260–272

Verrelst, J., Rivera, J.P., Gitelson, A., Delegido, J., Moreno, J., Camps-Valls, G., 2016b. Spectral band selection for vegetation properties retrieval using gaussian processes regression. Int. J. Appl. Earth Obs. Geoinf. 52, 554–567. http://dx.doi.org/10.1016/j.jag.2016.07.016.

Verrelst, J., Rivera, J.P., van der Tol, C., Magnani, F., Mohammed, G., Moreno, J., 2015c. Global sensitivity analysis of the scope model: what drives simulated canopy-leaving sun-induced fluorescence? Remote Sens. Environ. 166, 8–21.

Verrelst, J., Romijn, E., Kooistra, L., 2012c. Mapping vegetation density in a heterogeneous river floodplain ecosystem using pointable CHRIS/PROBA data. Remote Sens. 4, 2866–2889.

Wold, H., 1966. Non-linear Estimation by Iterative Least Procedures Squares.

Wold, S., Kettaneh-Wold, N., Skagerberg, B., 1989. Nonlinear pls modeling. Chemometr. Intell. Lab. Syst. 7, 53–65.