Int J Appl Earth Obs Geoinformation 88 (2020) 102051

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

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

An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing

Emma Izquierdo-Verdiguier[a,*], Raúl Zurita-Milla[b]

[a] Institute of Geomatics, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria
[b] Faculty of Geo-Information Science & Earth Observation (ITC), University of Twente, Enschede, The Netherlands

ARTICLE INFO

Keywords:
Feature selection
Random forest
Dimensionality reduction
Land cover classification
Biophysical parameter retrieval

ABSTRACT

New Earth observation missions and technologies are delivering large amounts of data. Processing this data requires developing and evaluating novel dimensionality reduction approaches to identify the most informative features for classification and regression tasks. Here we present an exhaustive evaluation of Guided Regularized Random Forest (GRRF), a feature selection method based on Random Forest. GRRF does not require fixing a priori the number of features to be selected or setting a threshold of the feature importance. Moreover, the use of regularization ensures that features selected by GRRF are non-redundant and representative. Our experiments based on various kinds of remote sensing images, show that GRRF selected features provides similar results to those obtained when using all the available features. However, the comparison between GRRF and standard random forest features shows substantial differences: in classification, the mean overall accuracy increases by almost 6% and, in regression, the decrease in RMSE almost reaches 2%. These results demonstrate the potential of GRRF for remote sensing image classification and regression. Especially in the context of increasingly large geodatabases that challenge the application of traditional methods.

1. Introduction

New Earth observation missions and technologies are delivering data with better spatial, spectral and temporal resolutions. At the same time, several agencies and satellite data providers have adopted open data standards and are delivering large amounts of data for free. For instance, the European Space Agency (ESA), in partnership with the European Commission, delivers data from the Copernicus program[1] and the National Aeronautics and Space Administration (NASA) delivers data from its Afternoon Constellation.[2] In addition, new multisource constellations (e.g. optiSAR, UrtheCast[3]) are been actively developed. The large amounts of Earth observation data delivered by current and upcoming missions take the remote sensing field to the big data era. Hence, remote sensing practitioners are bringing cloud computing and other big data technologies to this multidisciplinary field (Aguilar et al., 2018; Izquierdo-Verdiguier et al., 2018; Zurita-Milla et al., 2019). Despite the obvious benefits of using big data technologies, remote sensing data still requires efficient methods to deal with typical issues (Liu et al., 2018) such as noise (Gómez-Chova et al., 2008) and/or redundant information (Dominik, 2017). In this regard, methods that allow us to reduce the noise and the redundancy of the data while keeping the relevant content information, are still vital for the remote sensing community.

Dimensionality reduction is a basic and common preprocessing step to many data-driven modelling problems like image classification and biophysical parameter retrieval. Dimensionality reduction condenses the size of typical remote sensing data problems as well as dealing with the curse of the dimensionality (Bellman, 1961) (also called Hughes phenomenon). In the domain of remote sensing, dimensionality reduction methods are typically classified into feature extraction and feature selection. Both types of methods are essential to build classification or regression models because high dimensional problems often lead to poor results and hamper the process of creating good and reliable prediction maps (Camps-Valls and Bruzzone, 2005; Izquierdo Verdiguier, 2014). And both methods are used to reduce the high collinearity between spectral bands in hyperspectral images (Guyon et al., 2008) and between spatial or spectral feature generated from the image bands (Zurita-Milla et al., 2017). Despite the good results obtained by

[*] Corresponding author.
E-mail addresses: emma.izquierdo@boku.ac.at (E. Izquierdo-Verdiguier), r.zurita-milla@utwente.nl (R. Zurita-Milla).

[1] https://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus.

[2] https://atrain.nasa.gov/.

[3] https://www.urthecast.com/optisar/.

https://doi.org/10.1016/j.jag.2020.102051
Received 5 June 2019; Received in revised form 25 October 2019; Accepted 6 January 2020

feature extraction methods in classification and regression tasks (Izquierdo-Verdiguier et al., 2014), specially using non-linear methods (Izquierdo-Verdiguier et al., 2017a), results are hard to interpret because the physical meaning of the features is lost. In contrast to this, feature selection methods allow an easy interpretation of the most important features for a given model or task (Haury et al., 2011).

Several feature selection methods can be found in literature (Jović et al., 2015). For instance, there are methods based on ranking variables or selecting features that minimize a given criterion (filters methods). Other methods check the performance of the features for a given classifier or regressor (wrapper methods) or select features during the execution of the classification or regression tasks (embedded methods). Considering that wrapper and embedded approaches generally lead to better results than filter-based methods, here we focus on a random forest (RF) method (Breiman, 2001) that is a well-known embedded approach. RF is a simple and a fast way to select “interpretable” features. This, coupled with the fact that remote sensing often provides top results in both classification and regression tasks, explains its pervasive use by the remote sensing community. Moreover, RF selected features are in agreement with existing domain knowledge (e.g. physiological knowledge (Guan et al., 2012)).

RF is an ensemble learning method widely used in both, image classification (Pal, 2005) and biophysical parameter retrieval (Mutanga et al., 2012) tasks. Non-linear and dimensional problems are often addressed by this ensemble method. Furthermore, RF is also a powerful feature selection method because it provides the importance of each feature for the task at hand. As a feature selection method, RF shows an accuracy improvement regarding to filter and wrapper methods (Pal and Foody, 2010). For that reason, RF has become one of the most used methods in remote sensing image classification (Gislason et al., 2004; Rodriguez-Galiano et al., 2012; Xia et al., 2018). For instance, RF was used as a feature selection method and as a classifier to compare the performance of different sensors to map mangrove extent and species (Wang et al., 2018). A recursive and a non-recursive feature elimination processes were used to select the features, which were also used in (Gregorutti et al., 2017) in presence of correlated features. Another application of RF as feature selection method was presented in (Genuer et al., 2010) where the feature importance sensitivity was analyzed versus the characteristics of the data and the RF was used in their proposal feature method. However, the use of RF as a feature selection method either requires fixing a threshold of feature importance or specifying a priori the number of features that will be selected. In addition, the selection of features with high importance does not warranty that this is the best set of features for a given problem. For instance, high dimensional data usually have high correlated features and that has a negative effect on the feature selection (Gregorutti et al., 2017). Different solutions have been proposed to select a relevant subset of features like the Boruta algorithm (Beckschäfer et al., 2014) or an alternative RF implementation, which provides an unbiased variable selection using subsampling without replacement (Strobl et al., 2007).

Methods that regularize RF (Deng and Runger, 2013) have demonstrated their efficiency in reducing model complexity while providing a compact set of features. Regularized RF models, originally proposed and tested for applications in genetic research (Deng, 2013; Deng and Runger, 2013), disregard the features that share information, i.e. features with high collinearity. Thus, regularized feature selection methods do not lead to loss information.

In this paper we present an exhaustive and detailed evaluation of a special kind of regularized random forest for feature selection, namely Guided Regularized RF (GRRF) (Deng and Runger, 2013), in typical remote sensing data analysis tasks. GRRF has previously been applied in remote sensing image classification using hyperspectral data to identify invasive plant species (Mureriwa et al., 2016) and to classify four stages of Maize infection (Dhau et al., 2018). Additionally, GRRF was used to detect infestation in Maize crops (Adam et al., 2017) and to identify smallholder farms (Izquierdo-Verdiguier et al., 2017b). However, none of these studies exhaustively analyzed the best parametrization of GRRF, and neither evaluated nor optimized the number of features. The novel contributions of this paper consist in (1) providing an in-depth analysis of the behavior of GRRF taking into account the different algorithm parameters, (2) optimizing the number of features in an objective fashion and, (3) evaluating the GRRF algorithm by comparing its results with those obtained by a traditional RF trained with as many features as identified by GRRF. All these contributions were done for different images and classification and regression tasks.

2. Review of Random Forest

RF was proposed by Breiman (Breiman, 2001) as a combination of decision trees. This combination reduces the error in classification and regression tasks thanks to the use of bootstrap aggregation or bagging. RF is a supervised and simple (ensemble of decision trees) method, which is fast and robust to the noise of the target data (Kontschieder et al., 2011). The main idea of RF is to reduce the error of the prediction taking into account the decision trees included within the forest and the correlation among their predictions (Chan and Paelinckx, 2008).

Focusing on one tree of the forest, let $\mathbf{P}_i \in \mathbb{R}^{M_i \times N_i}$ where the i defines the ith partition of samples (M_i) and features (N_i). $\mathbf{P}_i$ is randomly selected from the original data ($\mathbf{X} \in \mathbb{R}^{M \times N}$) by generating random samples with replacement (i.e. by bootstrap (Efron and Tibshirani, 1994)). At each node, the feature belonging to the subset N_i are considered candidates to split the available samples (M_i). The Gini Index (GI, see 2.1 for more details) is used to find the best splitting feature and cutoff point. Samples that have higher values than the cutoff point for the selected feature are directed to the right node (ν_R) otherwise, they go to the left node (ν_L). After several splittings, samples have moved from the root node (ν_n) to the terminal nodes, also known as a terminal leaves which supply the predictions of the samples (Fig. 1). The ensemble prediction ($\hat{\mathbf{Y}} \in \mathbb{R}^{M \times 1}$) given by a forest is obtained as a combination of the results of the individuals trees; typically using the majority vote rule for classification or the average for regression problems (Criminisi et al., 2012):

$$\text{Classification: } \hat{Y}_i = \text{mode}_{n=1\cdots N_{trees}} \hat{Y}_n$$

$$\text{Regression: } \hat{Y}_i = \frac{1}{N_{trees}} \sum_{n=1}^{N_{trees}} \hat{Y}_n,$$

where N_{trees} is the total number of trees used in the RF.

Two parameters deserve attention when optimizing a RF: the number of features that will be considered as split candidates (i.e. the size of the N_i subset), and the number of trees in the ensemble (i.e. N_{trees}). The former is often fixed by sqrt(N) for classification or $N/3$ for regression (Liaw and Wiener, 2002), where N is the number of features in $\mathbf{X}$, and the latter is typically set equal to a few hundred of trees because more trees do not necessarily lead to a better performance and

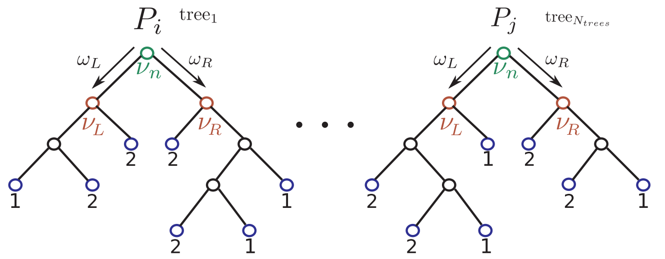

Fig. 1. Schematic representation a RF classifier with N_{trees} trees. At each root node ν_n and internal nodes (ν_L and ν_R), a statistical measure is applied to create homogenous groups from the data partitions ($P_i \cdots P_j$) assigned to each tree. These partitions are recurrent split until reaching the terminal nodes (in blue), which get assigned a label (classes 1 or 2 in this case). The fraction of samples that falls in each node are represented by ω_L and ω_R.

just slow down the processing time. Several criteria are used to optimize these parameters, such as k-fold cross-validation (Stone, 1974).

2.1. Random Forest as feature selection method

RF provides the importance of each feature (Breiman, 2001). This importance is used to identify the most relevant features for a given problem as well as to generate a feature selector method (Saeys et al., 2008). The importance of feature $\mathbf{x}_j$ is determined by:

$$\text{Importance}_j = \frac{1}{N_{\text{trees}}} \sum_{\nu \in S} G(\mathbf{x}_j, \nu), \tag{1}$$

where S is the set of nodes where $\mathbf{x}_j$ is used to split the samples, and $G(\mathbf{x}_j, \nu)$ is so-called the RF gain of $\mathbf{x}_j$. Thus gain is based on impurity measures calculated when the samples are split at each node. Several impurity criteria have been used to split the data and, therefore, to determine the feature importance. Measures like permutation importance (Gregorutti et al., 2017), or alternative implementations of RF like Boruta (Kursa et al., 2010) or subsample without replacement (Strobl et al., 2007), were developed to improve the selection of features when they are correlated. However, GRRF is based on regularization and therefore, it uses the most common criteria, which for classification problems is the GI function (Breiman et al., 1984). The GI, which is simple and fast to compute (Nembrini et al., 2018), minimizes the probability of misclassification by $\text{GI} = 1 - \sum_{i=1}^{n_c} (p_i)^2$, where n_c is the number of classes and p_i is the probability of class i. So, the function G in Eq. (1) is calculated by means of the following equation:

$$G(\mathbf{x}_j, \nu) = \text{GI}(\mathbf{x}_j, \nu) - \omega_R \text{GI}(\mathbf{x}_j, \nu_R) - \omega_L \text{GI}(\mathbf{x}_j, \nu_L), \tag{2}$$

where ω_R and ω_L are the fractions of the samples that fall in each node.

The split criteria in regression is often the measured by Residual Sum Squares (RSS): $\text{RSS} = \sum_{i=1}^{N_i} (y_i - \hat{y}_i)^2$, which is equivalent to look for the split that maximizes the sum squares between-groups in an analysis of variance (Therneau and Atkinson, 1997). In this case, G is obtained as follows:

$$G(\mathbf{x}_j, \nu) = \text{RSS}(\text{RSS}_L + \text{RSS}_R), \tag{3}$$

where RSS_R and RSS_L are the RSS in right and left nodes, respectively.

2.2. Regularized Random Forest

Regularized RF (RRF) provides high quality feature subsets as it was demonstrated in (Deng and Runger, 2012), and it leads to a reduction in the number of features selected for classification and regression problems.

Let be F the selected subset of features (initially empty) and $\mathbf{x}_j$ each feature, the gain of the RRF is calculated by:

$$G_{\text{RRF}}(\mathbf{x}_j, \nu) = \begin{cases} G(\mathbf{x}_j, \nu) & \text{if } j \in F \\ \lambda G(\mathbf{x}_j, \nu) & \text{if } j \notin F, \end{cases} \tag{4}$$

where G is the gain (Eqs. (2) and (3)), F is the subset of features selected to split the samples in previous nodes and $\lambda \in [0, 1]$ is a penalty factor for the features not selected in previous nodes. It is worth noting that a feature should have high importance value to be selected since its gain is penalized. However, if a feature is already selected, then its gain is equal to that of the standard RF. Taking into account that the irrelevant features have very low important value (Louppe et al., 2013), the features selected by RRF are non-redundant features because the features whose G_{RRF} is equal to zero are not included in the selected group. The penalty factor (λ) is equal for all features and when equal to one, RRF behaves like a standard RF.

RRF uses a sequential approach (i.e. it goes through all the nodes of a tree and through the features selected in previous trees). See Algorithm 1 for further details. Notice that some of the RRF selected features can have a representativeness problem. This problem happens because the number of distinct values of the G function is limited and several features can have the same gain when the number of samples in the node is small (Deng and Runger, 2013). The GRRF (c.f. Section 3) can be used to avoid this problem because it uses a second regularization of the gain of the features.

Algorithm 1. RRF feature selection process.

```
Require: $N_{\text{trees}}$: num. of trees, $\lambda$: penalty factor, $F_1$: set of features indices and $F$:
    subset of selected feature indices.
Ensure: RF: random forest model
  Train the RF prediction model to obtain the feature importance.
  Initialize $F = \{\}$ and a threshold gain ($G^* = 0$).
  Select samples and features.
  for $t = 1$ to $N_{trees}$ do
    $\nu \longleftarrow$ numberofthenodesinthetree
    for $n = 1$ to $\nu$ do
      while $F_1 \neq \emptyset$ do
        $j \longleftarrow$ indexofselectedfeaturefrom $F_1$
        Calculate the $G_{\text{RRF}}(\mathbf{x}_j, \nu)$ (Eq. (4))
        if $G_{\text{RRF}}(\mathbf{x}_j, \nu) > G^*$ then
          $\{F, j\} \longleftarrow F$
          $G^* \longleftarrow G_{\text{RRF}}(\mathbf{x}_j, \nu)$
        end if
        Delete $j$ from $F_1$
      end while
    end for
  end for
  $N_f \longleftarrow$ lengthof $F$
```

Algorithm 2. GRRF feature selection process.

```
Require: $N_{\text{trees}}$: num. of trees, $\lambda$: penalty factor, $\gamma$: weight of normalize feature
    importance, $F_1$: set of feature indices and $F$: subset of selected feature indices.
Ensure: RF: random forest model
  Train the RF prediction model to obtain the feature importance.
  Initialize $F = \{\}$ and a threshold gain ($G^* = 0$).
  Select samples and features.
  for $t = 1$ to $N_{trees}$ do
    $\nu \longleftarrow$ numberofthenodesinthetree
    for $n = 1$ to $\nu$ do
      while $F_1 \neq \emptyset$ do
        $j \longleftarrow$ indexofselectedfeaturefrom $F_1$
        Calculate the regularization parameter ($\alpha_j$) using eq. (6).
        Calculate the $G_{\text{GRRF}}(\mathbf{x}_j, \nu)$ (eq. (5))
        if $G_{\text{GRRF}}(\mathbf{x}_j, \nu) > G^*$ then
          $\{F, j\} \longleftarrow F$
          $G^* \longleftarrow G_{\text{GRRF}}(\mathbf{x}_j, \nu)$
        end if
        Delete $j$ from $F_1$
      end while
    end for
  end for
  $N_f \longleftarrow$ lengthof $F$
```

3. Guided Regularized Random Forest

The Guided Regularized Random Forest (GRRF) consists of a RRF where the regularization is steered by the feature importance of the standard RF. In GRRF, the calculation of the gain (and therefore the feature importance) considers information from all the nodes instead of only relying on information from a single node (Eq. (1)). Because of this design, GRRF uses a specific regularization parameter for each feature.

The regularization parameter preserves the RF gain (Eq. (2) for classification or Eq. (3) for regression) of the features selected in previous nodes and punishes the gain of new features. The GRRF gain is defined as:

$$G_{\text{GRRF}}(\mathbf{x}_j, \nu) = \begin{cases} G(\mathbf{x}_j, \nu) & \text{if } j \in F \\ \alpha_j G(\mathbf{x}_j, \nu) & \text{if } j \notin F, \end{cases} \tag{5}$$

where α_j is the regularization parameter of each feature, which depends of the normalized RF feature importance provided by the standard RF:

$$\alpha_j = (1-\gamma)\cdot\lambda + \gamma\frac{\text{importance}_j}{\max_{j=1,\cdots,N_i}(\text{importance}_j)}. \tag{6}$$

Note that, $\lambda \in [0, 1]$ is a penalty factor, $\gamma \in [0, 1]$ is the weight of the normalized feature importance and that pairwise combination $(\lambda, \gamma) = (0, 0)$ does not make sense. The GRRF gain is equal to the RRF one when γ is equal to zero and it is equal to the gain of a standard RF when λ is equal to one and γ is equal to zero. Note also that both RRF and GRRF are just feature selection methods that should not be used in prediction tasks because their trees are grown in a sequential manner oriented towards the identification of the most important features (and not to improve the prediction, like boosted trees do). This training approach leads to a high variance of the predictions (Deng and Runger, 2013). Thus, features selected must be used as input in standard classification and regression methods (to test and evaluate their value).

The GRRF feature selection process is summarized in the pseudo-code shown in Algorithm 2. The processing has implemented in Python 3.6 and used the GRRF toolbox (Deng, 2013).

Summarizing, RF identifies important features based on their gain in all nodes of the trees. However, its use as a feature selection method requires either fixing the number of features to select or applying a threshold of feature importance. RRF selects non-redundant features that, in some cases, suffer from a lack of representativeness. Furthermore, the penalization linked to the regularization is constant for all the features. GRRF uses a double regularization based on the RF feature importance and on penalizing each feature individually. This is so called guided regularization generates a subset of non-redundant and representative features.

4. Data

This section describes the data used in the classification and regression experiments designed to evaluate the proposed feature selection technique. We tackle the classification task using two types of data: first, a temporal series of very high spatial resolution multi-spectral images. Second, various hyperspectral images with different spatial resolution, number of bands (features) and classes. The regression task deals with the retrieval of biophysical parameters. The task consists on predicting chlorophyll (Chl), leaf area index (LAI) and fraction cover (fCover) from a hyperspectral image. The classification and regression experiments are designed to evaluate how GRRF copes with different number of classes (in classification), number of features and spatial resolution.

4.1. Classification databases

4.1.1. Multispectral temporal image series (WorldView-2)

The multispectral images used in this work were acquired by *WorldView-2*.[4] Seven acquisition were used in this case study from May to November of 2014 (Fig. 2). These acquisitions cover the main crop season in the study area (Vrieling et al., 2011). Each multispectral image consists of eight spectral bands and with a spatial resolution of 2 m covering 10 km by 10 km near Sukumba, Koutiala district, Mali. All of them were preprocessed using the satellite image workflow (Stratoulias et al., 2017) developed in the STARS project.[5] With this workflow the images were mosaicked, orthorectified, co-registered, and trees and clouds were automatically masked out.

In addition to the $56(7 \times 8)$ bands, the database was extended by spectral and spatial features. Vegetation indices such as Normalized Difference Vegetation Index (NDVI) (Tucker, 1979), Soil Adjusted Vegetation Index (SAVI) (Huete, 1988), Transformed Chlorophyll Absorption Reflectance Index (TCARI) (Haboudane et al., 2002), Modified Soil-Adjusted Vegetation Index (MSAVI2) (Qi et al., 1994), Enhanced vegetation index (EVI) (Huete et al., 2002) and Green Vegetation Index (GLI) (Yamamoto et al., 2005) were obtained. Furthermore, all pairwise band combinations between bands 2 and 8 were used to calculate differences, ratios and normalized differences. The spatial features were based on the Local Binary Pattern (LBP) (Pietikäinen, 2010) of the 56 spectral bands and on textural metrics derived from the Gray Level Co-occurrence Matrix (GLCM) (Haralick et al., 1973; Conners et al., 1984). A total of 17 textures were calculated from the bands, vegetation index and LBP features. After all these calculations, the dimensionality increased from 56 to 10572 features.

The ground truth is composed of 45 farm field polygons, which were delineated during field work. The fields were divided in four sub-polygons of approximately the same size. Two polygons were used to choose the training samples and the other two, the test samples to ensure independent both subsets. A total of 1500 pixels were extracted for the training and validation sets and 990 samples were used to create the test set. Five crop classes of interest were identified in the farm fields: Maize, Millet, Peanut, Sorghum and Cotton.

4.1.2. Hyperspectral images (Indian Pines, Pavia and Salinas)

Three classical hyperspectral images in remote sensing image classification were used to evaluate the potential of GRRF to reduce dimensionality without losing valuable information.

The first image is a scene over Indian Pines site in North-western Indiana. This image was acquired by the AVIRIS sensor and consists of 224 spectral bands in a range of [400, 2500] nm and of 145×145 pixels. Note that, the water absorption bands were removed (24 bands). The image covers an agricultural area and its ground truth contains 16 classes.

The second image was acquired by the DAIS7915 sensor over Pavia (Italy). The image reveals a dense residential area at 5 m spatial resolution with nine classes: Water, Trees, Asphalt, Parking, Bitumen, Brick roofs, Meadows, Bare soil and Shadows. The image consists of 400×400 pixels and has a spectral range from 500 to 1760 nm splitting into 40 bands (skipping the thermal and middle infrared range (Graña and Duro, 2008)).

The last hyperspectral image used in the classification experiments covers an agricultural area of California (USA) and was acquired by AVIRIS over the Salinas Valley. The size of the image is 217×512 with 204 spectral bands which cover a spectral range [400, 2450] nm. As with Indian Pines, the water absorption bands were removed (in this case, 20 bands). The ground truth contains 16 classes: two kinds of Brocoli, three types of Fallow, Stubble, Celery, Grapes, Soil vineyard, Corn, two vineyard and four Lettuce classes.

The three images are available in http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.

4.2. Biophysical parameter retrieval data

We handle the prediction of three biophysical parameters by mean of the multi-spectral image to show the efficiency of GRRF in regression tasks. The database consists of image data and *in situ* measurements. Note that, the difficulty of the scenario since the target data is lacking and/or their relations, between satellite derived data and the site visit data is believed nonlinear.

This database was obtained in the SPectra bARrax Campaign (SPARC) in Barrax, Spain.[6] A hyperspectral image was collected in 2003 by the CHRIS/PROBA spaceborne sensor. The data provided have 62 bands, although 7 bands were removed to avoid noise problems. The band range covers the visible and near-infrared (NIR) region (400–1000 nm) at a spatial resolution of 34 m. The image selected for this experiment was those acquired from the nadir view sharing similar

[4] http://www.satimagingcorp.com/satellite-sensors/worldview-2/.

[5] http://www.stars-project.org/en/.

[6] https://www.uv.es/leo/sparc/.

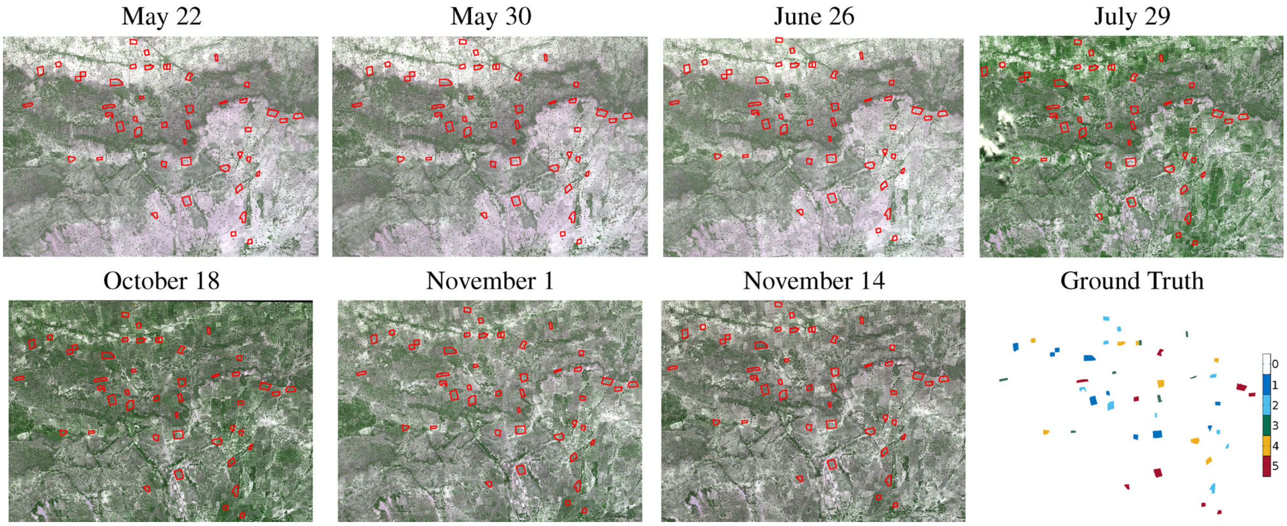

Fig. 2. Farm field polygons (red) overlapping the time series of RGB composites and the available ground truth (class 0: no data, 1: Maize, 2: Millet, 3: Peanut, 4: Sorghum and 5: Cotton).

Table 1
Summary of classification databases (N: number of features, n_c: number of classes, N_{train}: number of training samples, N_{val}: number of validation samples, and N_{test}: Number of test samples.) used in classification.

Image	N	n_c	N_{train}	N_{val}	N_{test}
WorldView-2	10572	5	500	250	990
Indian Pines	200	16	1323	640	8286
Pavia	40	9	900	450	13224
Salinas	204	16	1600	800	50230

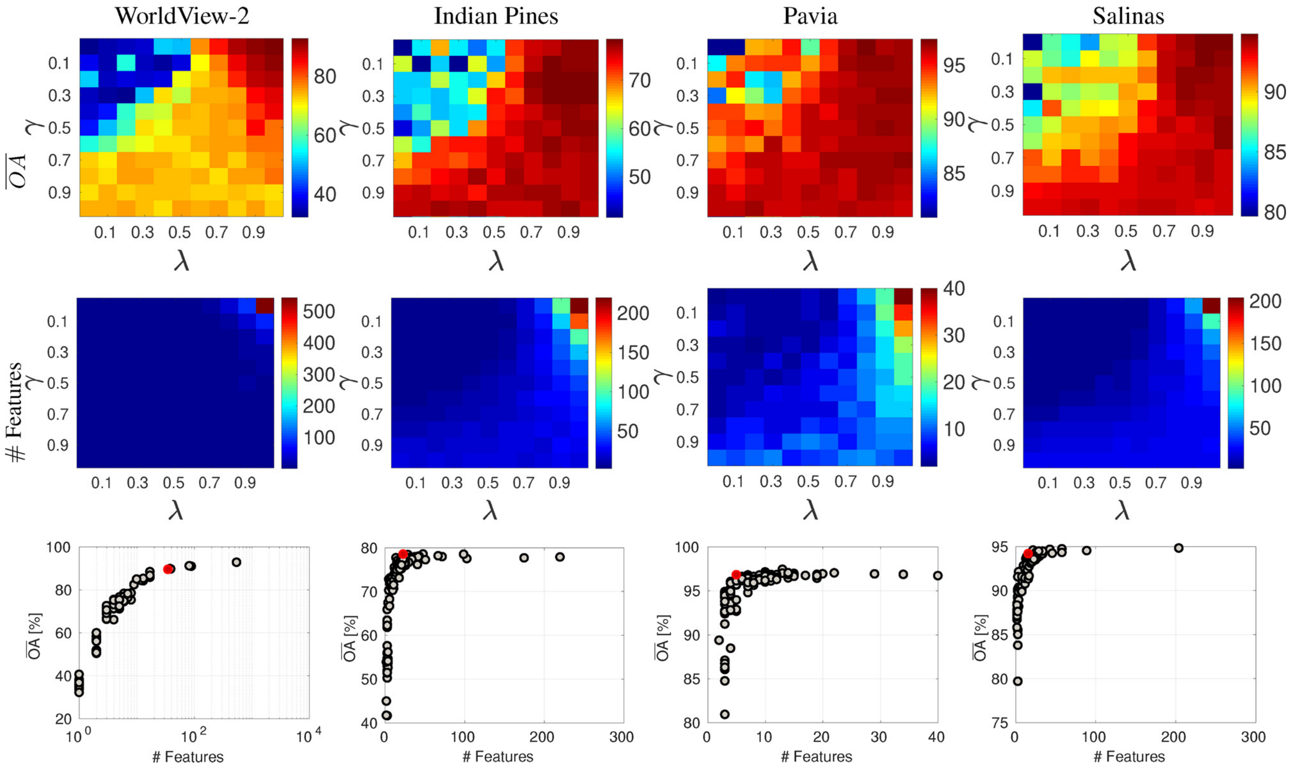

Fig. 3. Overall accuracy (%) (Top) and number of features selected by GRRF (Middle) for different γ and λ values. Overall accuracy versus the number of selected features for all classification databases (Bottom). The red points indicate the optimal number of features and provide γ^* and λ^*.

observation configuration in order to minimize angular and atmospheric effects. The image was geometrically and atmospherically corrected using the official CHRIS/PROBA Toolbox for BEAM (Alonso et al., 2009). Simultaneously to the acquisition, ground data was collected in the test area. Barrax is an agricultural research facility characterized by a flat landscape and large uniform land-use units of irrigated and dry lands and has an extension of 5 × 10 km. The vegetation biophysical parameters were measured among different crops. The Chl was measured with a calibrated Minolta CCM-200, the LAI was derived from canopy measurements made with a LiCor LAI-2000 and the fCover was derived from hemispherical photographs. All parameters present standard errors between 3% and 10%. The field-measured values of Chl between 2 and 55 $\mu g/cm^2$, LAI vary between 0.4 and 6.3, and fCover

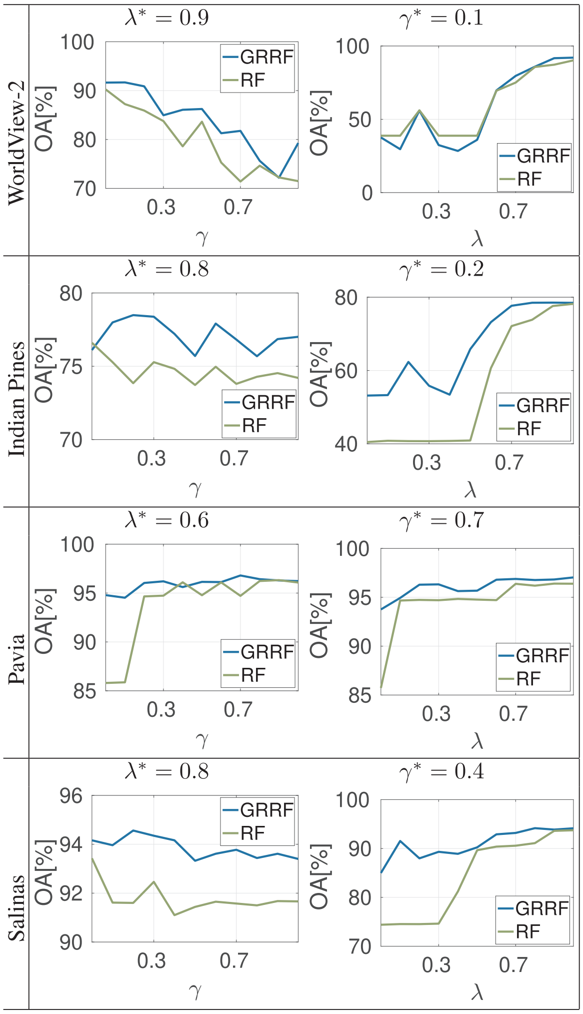

Fig. 4. Comparison of $O\bar{A}$ using n GRRF features and the n top features of RF sorted by their importance for all classification databases. n is fixed to Pareto optimization.

between 0 and 1. A total of 135 measurements of Chl, LAI, and fCover were extracted from the Barrax database and their associated 55 CHRIS reflectance channels form the database.

5. Experiments and results

This section presents the classification and regression experiments and their results. All the experiments require completing the following three steps:

1. *Parameter optimization*: The training set was used to obtain the feature importance from a RF model. Then a range of λ and γ values was used within GRRF to obtain various sets of selected features. After that, a RF model was trained, and the parameters that optimize the model were selected as the optimum ones (λ^* and γ^*). Note that, after this step the subset of features is still unknown due to the randomness of the RF.
2. *GRRF vs. standard RF features*: Once λ^* and γ^* were fixed, GRRF was applied to different partitions of the training data. The subset that optimizes the model was selected to evaluate the added value of the N_f selected features by GRRF, their performance was compared against that obtained by the top N_f features provided by standard RF.
3. *Model assessment*: The GRRF selection was evaluated using two well-known methods in remote sensing field: RF and SVM. Both methods were used to compare the results obtained with the features selected by GRRF and RF, and when using all available features. To measure the quality of the selected features, we report the Overall Accuracy (OA) and Cohen's kappa (κ) (Cohen, 1960) for classification problems. For regression, we report the root mean square error (RMSE) and the mean absolute error (MAE) to evaluated the precision of the

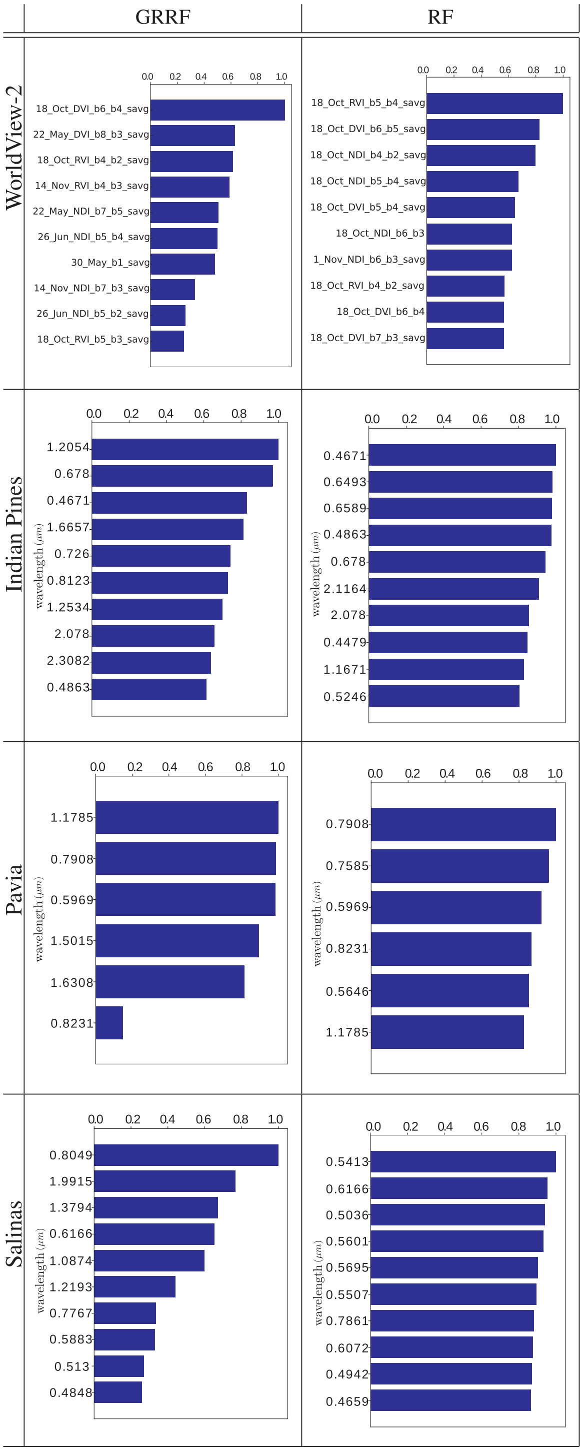

Fig. 5. Top 10 features selected by GRRF and RF together with their normalized feature importance. Notation WorldView-2: *date_feature_texture*: date of the acquisition, spatial or spectral feature and type of texture. Spectral or spatial feature: bx: spectral band of the image, NDI: normalize difference index, DVI: difference vegetation index, RVI: ratio vegetation index. Texture feature: svag: sum average.

Table 2
Mean overall accuracy, κ index and, percentage reduction in overall accuracy with respect to the best $\bar{OA}$.

WorldView-2 Feat. selection	# Feat.	RF $\bar{OA}$	$\bar{\kappa}$	%	SVM $\bar{OA}$	$\bar{\kappa}$	%	Pavia Feat. selection	# Feat.	RF $\bar{OA}$	$\bar{\kappa}$	%	SVM $\bar{OA}$	$\bar{\kappa}$	%
None	10572	87.15	0.84	1.36	90.09	0.88	–	None	40	96.28	0.95	1.31	97.70	0.97	0.07
GRRF	35	88.35	0.85	–	87.75	0.85	2.59	GRRF	7	97.56	0.97	–	97.56	0.96	–
RF	35	83.94	0.80	4.99	88.98	0.86	1.23	RF	7	96.09	0.95	1.51	97.17	0.97	0.68
Indian Pines Feat. selection	# Feat.	RF $\bar{OA}$	$\bar{\kappa}$	%	SVM $\bar{OA}$	$\bar{\kappa}$	%	Salinas feat. selection	#. Feat.	RF $\bar{OA}$	$\bar{\kappa}$	%	SVM $\bar{OA}$	$\bar{\kappa}$	%
None	200	70.28	0.66	0.21	70.16	0.66	6.89	none	204	87.56	0.86	–	89.09	0.88	–
GRRF	27	70.43	0.66	–	75.35	0.72	–	GRRF	16	86.08	0.84	1.69	87.79	0.86	1.45
RF	27	68.62	0.64	2.57	72.62	0.69	3.62	RF	16	83.86	0.82	4.22	85.06	0.83	4.52

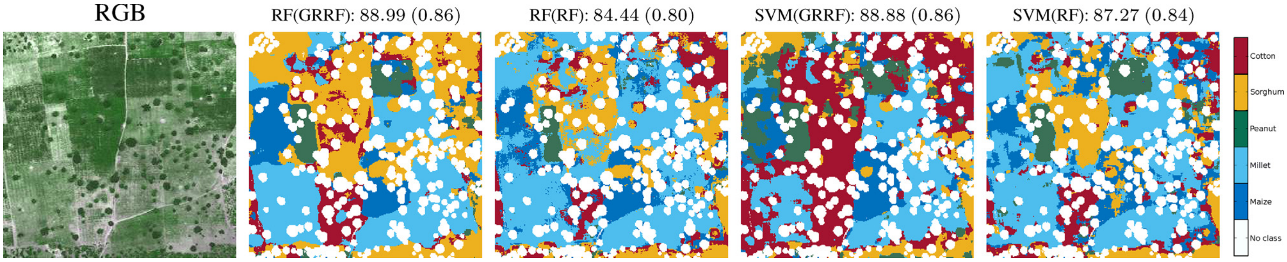

Fig. 6. (Left to right) RGB composite of a subset of the study area and the corresponding classification maps for a combination of classifier and feature selection method [Notation: classifier(feature selection method): OA (κ)]. In all the cases, the top 35 features were used (c.f. Table 2) and all pixels in the image are classified but the trees and clouds are masked (No class).

Table 3
Confusion matrix of the best classifiers for WorldView-2. [Notation: GT: Ground Truth, Pred.: Predictions, Ma: Maize, Mi: Millet, P: Peanut, S: Sorghum, C: Cotton, UA: User's Accuracy, Fsc: Fscore and, PA: Producer Accuracy].

	Classifier	RF							SVM						
		Pred.													
Sel.	GT	Ma	Mi	P	S	C	UA	Fsc	Ma	Mi	P	S	C	UA	Fsc
GRRF	Ma	167	6	3	4	3	0.92	0.91	159	8	1	9	6	0.88	0.86
	Mi	3	206	10	12	6	0.87	0.86	10	207	13	5	2	0.87	0.87
	P	0	0	108	10	2	0.86	0.90	0	0	113	6	1	0.86	0.94
	S	1	8	0	226	1	0.90	0.95	2	17	2	213	2	0.89	0.90
	C	6	14	8	12	174	0.87	0.81	4	2	13	7	188	0.91	0.87
PA		0.94	0.88	0.83	0.85	0.93			0.90	0.88	0.79	0.88	0.94		
RF	Ma	158	14	2	7	2	0.85	0.86	162	11	2	7	1	0.89	0.88
	Mi	7	189	11	21	9	0.79	0.79	12	199	14	10	2	0.86	0.83
	P	7	5	103	7	1	0.81	0.83	2	7	104	6	1	0.80	0.86
	S	4	11	4	208	9	0.86	0.88	2	4	7	212	11	0.87	0.89
	C	9	17	9	1	178	0.86	0.83	0	4	11	12	187	0.89	0.87
PA		0.85	0.80	0.79	0.85	0.89			0.91	0.88	0.75	0.85	0.92		

models, the mean error (ME) to determine the bias and the Pearson's correlation (R) to measure the goodness of the model fit.

5.1. Feature selection for classification

5.1.1. Training, validation and test data

The available databases were split into three subsets: training, validation and test. Training and validation sets were used to optimize and train the classifiers, and the test set was used to check the quality of the models. For WorldView-2, 100 and 50 samples per class were respectively used to train and validate the hyperparameters. Both sets were randomly selected from the available training samples (see Section 4.1.1). For the hyperspectral images, 100 samples per class were selected for training, 50 samples per class for validating and the rest of pixels of the database were used for testing the classifiers (see Table 1). Note that there are 4 classes in Indian Pines with few samples per class. In these cases two-thirds of the samples were randomly selected for training and the rest were split into validation (two-thirds of the remaining pixels) and test (one-thirds of the remaining pixels).

5.1.2. Parameter optimization

As mentioned in Section 3, GRRF requires training a standard RF to get the importance of all the features. Here we use a standard RF with 500 trees. The square root of the number of features is used to fix the number of features for each tree. A range of $\lambda \in [0, 1]$ and $\gamma \in [0, 1]$ both in steps of 0.1 was used to parametrize the GRRF and select the most important features. These features were evaluated using ten runs of a standard RF classifier, which provided ten OA values for each pairwise combination of (λ, γ). However, different (λ, γ) combinations provide different $\bar{OA}$ with the same number of features selected. A Pareto optimality (Box and Meyer, 1986) is used to find the best values GRRF parameters (λ^* and γ^*). For this, we use all combinations whose $\bar{OA}$ is higher than 98% of the maximum $\bar{OA}$. If more than one Pareto point is obtained, we select the λ^* and γ^* that minimize the number of

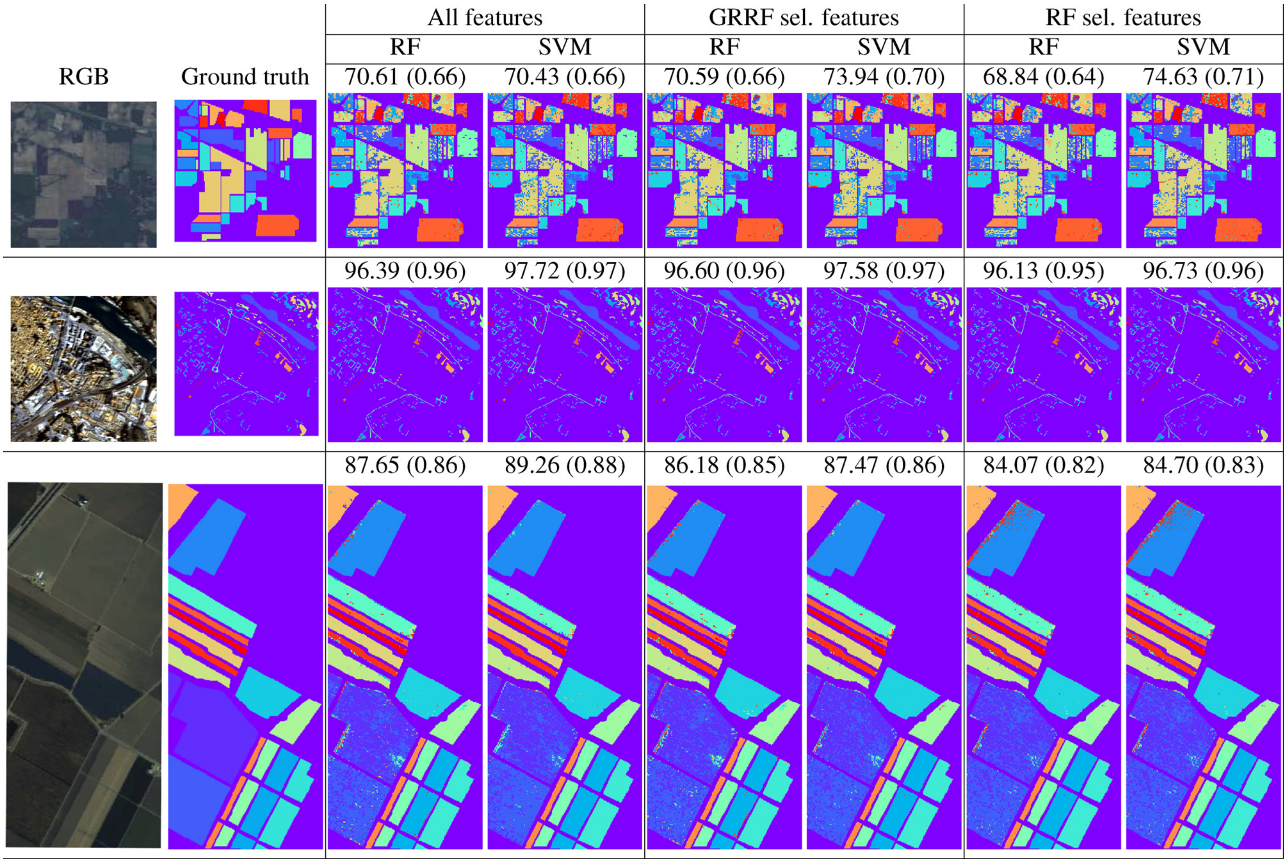

Fig. 7. (Left to right) RGB composite, ground truth, and six classification maps for the (top) Indian Pines image, (middle) Pavia image, and (bottom) Salinas image using 22, 8 and 18 features, respectively.

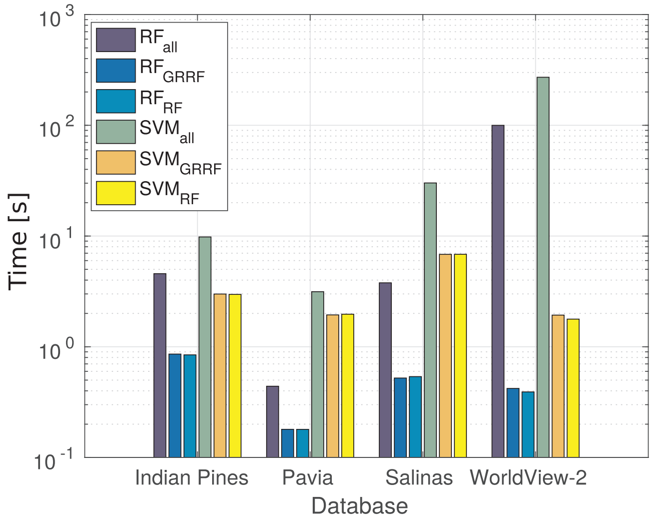

Fig. 8. Computational time per classifier and database using all features and selected features by GRRF and RF.

selected features.

The ŌA surfaces obtained for the different values of λ and γ are represented in Fig. 3 [Top] for the four classification databases. All surfaces show two patterns: low ŌA values for small λ and γ values and high ŌA values for large λ and small γ values. These patterns can be explained by the reduction in data dimensionality after applying GRRF (Fig. 3 [Middle]). This feature selection method extremely reduces the dimensionality of the data when both parameters are small whereas there is not reduction when $\lambda = 1$ and $\gamma = 0$.

For the WorldView-2 database, which has 10572 features (Table 1), the number of features is reduced to almost 600 when $\lambda = 1$ and $\gamma = 0$. This is because this database has features with a RF gain of to zero and that they do not fulfil the condition $G_{\text{GRRF}}(\mathbf{x}_j, \mathbf{v}) > G^*$ (see Algorithm 2). For the same reason, the case $\lambda = 0$ and $\gamma = 0$ is not calculated.

In general, both GRRF parameters yield high values of ŌA in all databases. However, the WorldView-2 database has a concentration of ŌA values smaller than 40% for λ and γ values below 0.6. Those values correspond to the lowest number of selected features (Fig. 3 [middle]). At this point, it is worth mentioning that small λ and γ values are always associated to a very number of features (typically less 5).

The bottom row of Fig. 3 shows the changes of ŌAs versus the number of features selected. As expected the ŌA saturates when the number of selected features is larger than the optimal one. This saturation point corresponds with the Pareto optimum.

5.1.3. GRRF vs. standard RF features

As outlined in Section 5.1.2, once the GRRF parameters are fixed, the best subset of features is selected by training ten RF classification models with different partitions of the training data. In this context, the best set of features is the one that leads to the model with the highest ŌA. These optimal features are compared with the same number of features extracted from a standard RF, and sorted by their RF importance.

Fig. 4 shows the optimal GRRF parameters (λ^* and γ^*) for each database as well as the relationship between the ŌA obtained with both subset of features and the values of γ (and λ) when fixing $\lambda = \lambda^*$ (and $\gamma = \gamma^*$). In all cases, the λ^* values are close to one and the γ^* are close to zero (except Pavia image). This confirms previous results where λ is typically fixed to one so that only γ needs to be optimized (Deng and Runger, 2013). Fig. 4 also shows that differences in ŌA between both

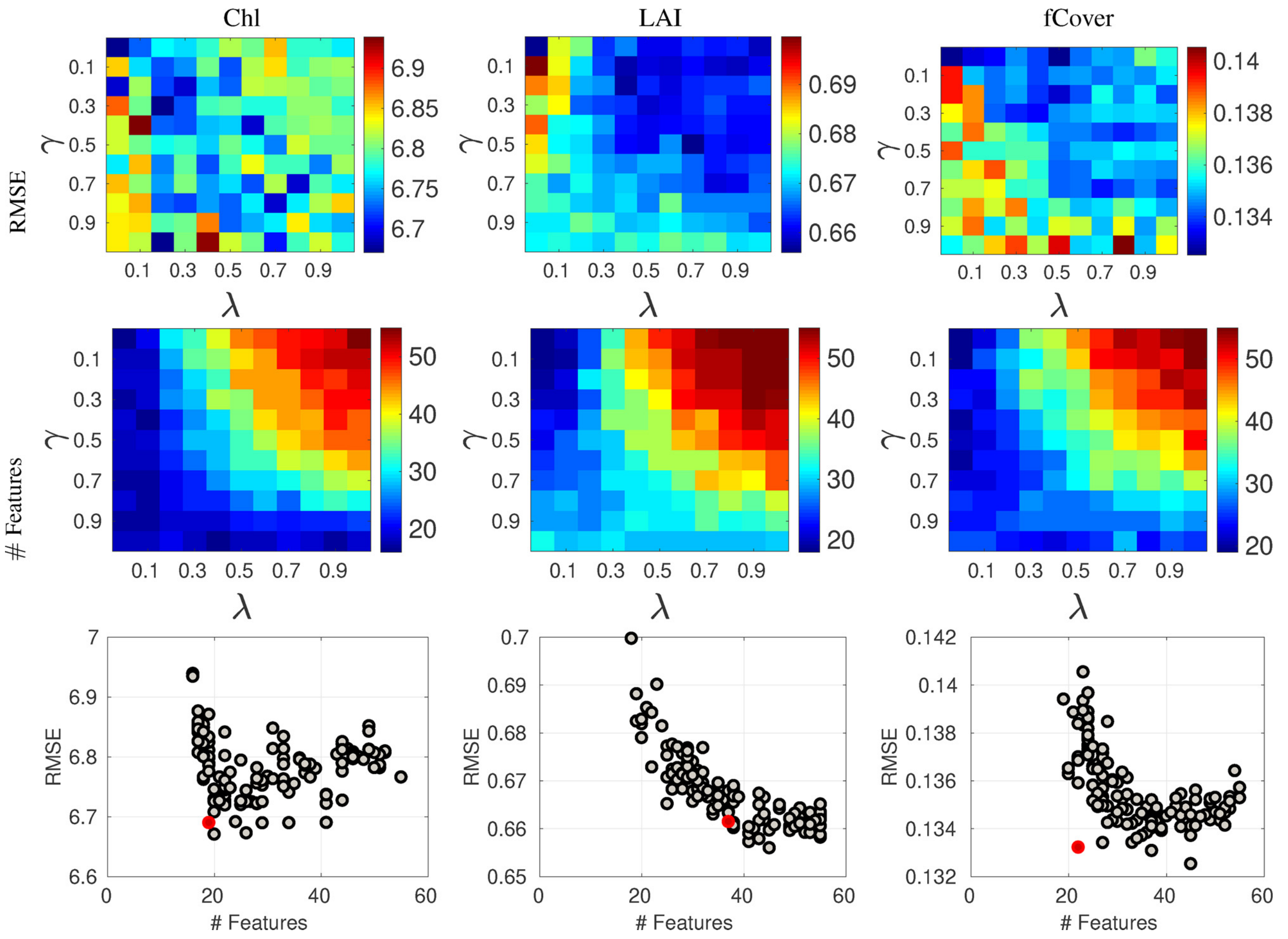

Fig. 9. RMSE comparison for a range of γ and λ values for different databases [Top]. Number of features selected by GRRF with different γ and λ values [Middle]. RMSE versus the number of features selected by GRRF [Bottom]. The red points indicate the optimal number of features and provide γ^* and λ^*.

subsets of features are small when $\lambda = 1$ for γ^* and $\gamma = 0$ for λ^* because almost all features are selected. However, when the number of features is smaller, the GRRF selected features often lead to much better $\overline{\text{OA}}$. Differences in $\overline{\text{OA}}$ are very small for all λ values when $\gamma^* = 0.1$ in the WorldView-2 database whereas they are particularly visible in the Pavia database. The $\overline{\text{OA}}$ is a non-monotonic function because the number of selected features are different for different sets of GRRF parameters. However, $\overline{\text{OA}}$ tends to increase as λ increases and decrease when γ increases.

GRRF, as RF-based method, allows visualizing the importance of the selected features. This helps to hypothesize over possible physical meaning and allows the creation of more transparent and interpretable models. The top 10 features selected by both GRRF and RF are shown in Fig. 5, together with their normalized importance. Notice that GRRF only found six important features in the Pavia database. In all the cases, the decrease in feature importance from RF is smoother than the one shown for GRRF. This may be caused by GRRF is more strict in selecting uncorrelated features, i.e. not selected the redundant and non-representative features.

For WorldView-2, the GRRF selected features come from five out of the seven acquisitions included in this database whereas the RF selected features focus on a single acquisition. We observe a prevalence for textures, specifically for vegetation index textures in both subsets of selected features. The *sum average* is the most important GLCM feature in both subsets of features. For the hyperspectral databases, GRRF selected high wavelength bands (higher than 1.5 µm) in all the cases, unlike the RF selected features, which only contain a few high wavelength bands for Pavia and Indian Pines cases. Regarding agricultural classifications (Indian Pines and Salinas), GRRF focuses on the red and NIR range contrarily to RF whose selection spreads over the visible range and includes bands in the green and blue range of the electromagnetic spectrum.

5.1.4. *Classification assessment*

Standard RF and SVM classification models were created using all the features and the GRRF and RF subsets to thoroughly evaluate the proposed GRRF features selection method. We used a 5 k-fold cross-validation to optimize the SVM model's complexity (C) and the Radial Basis Function (RBF) kernel length-scale (σ) parameters with the standardized data. C is optimized in a logarithmic range from 0.1 to 1000 in 10 steps and, $\sigma \in [0.5, 30]\times$ the mean Euclidean distance among labeled training data. In both cases, we sampled ten equally spaced values from the given ranges. Finally, the best models are used to create classification maps.

Table 2 shows the $\overline{\text{OA}}$ and $\bar{\kappa}$ as well as the percentage of reduction in $\overline{\text{OA}}$ with respect to the classification model with maximum $\overline{\text{OA}}$. GRRF selected features provide good results. In all the cases and for all the classifiers, the reduction in $\overline{\text{OA}}$ is less than 3%. This indicates that GRRF features can be used with various kinds of classifiers. Additionally, GRRF is not affected by the dimensionality of the databases, unlike RF selection. The features selected by RF are less informative as shown by the reduction in $\overline{\text{OA}}$, which ranges from 1 to almost 5%.

Considering the type of classifier, SVM always gets better results than RF. This indicates that the SVM classifier is less dependent on the number of features than RF.

Fig. 6 shows the RGB composite and the classifications maps of a

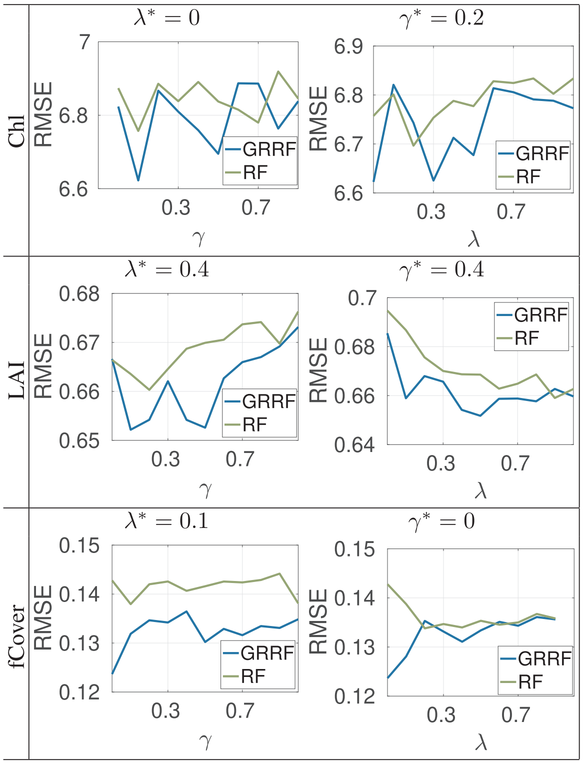

Fig. 10. Comparison of $R\bar{M}SE$ using n GRRF features and the n top features of RF sorted by their importance for Chl, LAI and fCover variables. n is fixed to Pareto optimization.

subset of the study area of the WorldView-2 database using the two feature selection methods and the two classifiers. The creation of the classification maps with all features is not computationally possible due to the very high dimensionality of the database (see Table 2). This subset was selected to visualize the results because it is large enough to contain all the classes of interest and small enough to be displayed with good resolution. The four classification maps show the smallholder farms in a clear way, and they can be recognized in the RGB composite. However, differences among the classification results are clearly visible. The GRRF-based maps are less noisy than the maps based on RF selected features. This confirmed by the higher OA and κ index. Regarding the classifiers, RF mixes less crops within the fields although the confusion matrices (Table 3) obtained from the test samples show that both classifiers have a similar performance (SVM being slightly better) as confirmed by the different statistics (Producer accuracy, User's accuracy and F-score).

The classification maps of the hyperspectal images are shown in Fig. 7 together with their OA and the κ values. In this case, maps were made using all features and the best subsets of GRRF and RF selected features. Results confirm Table 2 and the predominance of GRRF-based features over the standard feature selection offered by the RF classifier. An advantage of feature selection is the reduction in computational time of prediction the models. Fig. 8 show the processing time for both classifiers using different number of features (all, GRRF and RF sets). As expected, the computational time with all features is the highest for all databases and the times using GRRF and RF features are equal. We highlight the short time required to classify the WorldView-2 database using GRRF features. The proposed feature selection method coupled with the RF classifier outperforms the accuracy of non-feature selection (see Table 2) and its classification time is around 250 times faster.

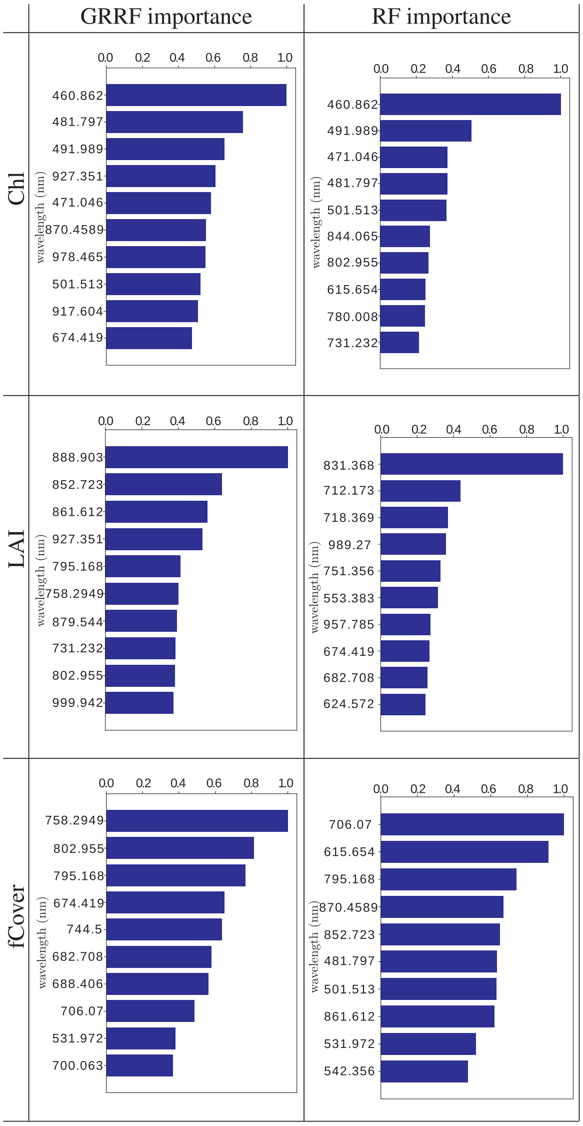

Fig. 11. Top 10 feature selected by GRRF [left] and RF [right] for Chl, LAI and fCover variables.

5.2. Feature selection for regression

5.2.1. Training, validation and test data

Unlike classification, the database has a limited number of samples that can be used for training, validating and testing the models. For that reason, we used 75 samples to train the model, 25 to validate it and the rest of the labeled samples were used to independently test the regression models.

5.2.2. Parameter optimization

Like for classification, we first trained a standard RF model to get the importance of all the features. In this case, we fixed the number of trees to 500 and the number of features available to each tree to one-third of the total number of features (N). The general process to optimize the GRRF parameters is identical to the one used in the classification problems (Section 5.1.2). The only difference is that here we use all pairwise combinations of λ and γ that yield a R$\bar{M}$SE that is up to 1% worse than the minimum R$\bar{M}$SE.

Fig. 9 shows the R$\bar{M}$SE (top) and the number of features selected (middle) for all the evaluated combinations of λ and γ for the three biophysical parameters. As in the case of classification, the best values of R$\bar{M}$SE are for high λs and small γs. Nevertheless, the regression

Table 4
Regression results for all the databases using all the dimensions and the selected features by GRRF and RF. [The number of selected feature is 17 for Chl, 32 for LAI and 21 for fCover].

Chl	RF			SVR		
	All	GRRF	RF	All	GRRF	RF
RMSE	6.568	6.402	6.649	4.568	4.271	6.032
MAE	3.794	3.910	3.867	2.615	2.557	3.551
ME	0.418	0.238	0.338	0.768	0.446	0.501
R	0.895	0.900	0.892	0.901	0.914	0.817
LAI	RF			SVR		
	All	GRRF	RF	All	GRRF	RF
RMSE	0.622	0.624	0.605	0.559	0.573	0.546
MAE	0.479	0.479	0.460	0.430	0.441	0.403
ME	−0.098	−0.105	−0.090	−0.156	−0.180	−0.098
R	0.892	0.891	0.896	0.824	0.816	0.832
fCover	RF			SVR		
	All	GRRF	RF	All	GRRF	RF
RMSE	0.138	0.137	0.136	0.133	0.141	0.142
MAE	0.094	0.095	0.096	0.104	0.111	0.107
ME	0.007	0.004	0.005	0.006	0.004	−0.011
R	0.900	0.902	0.902	0.812	0.786	0.785

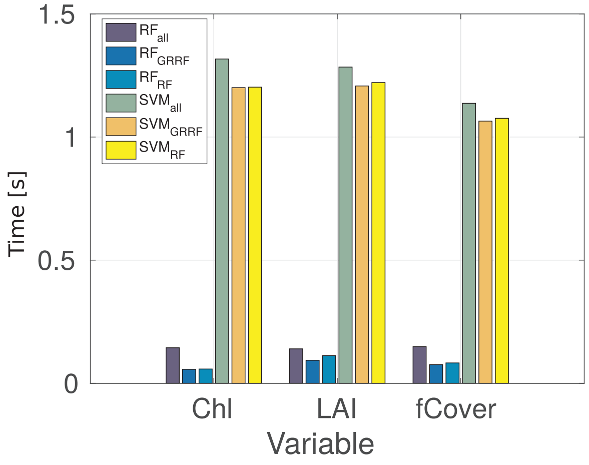

Fig. 13. Computational time per predition model and variable using all features and selected features by GRRF and RF.

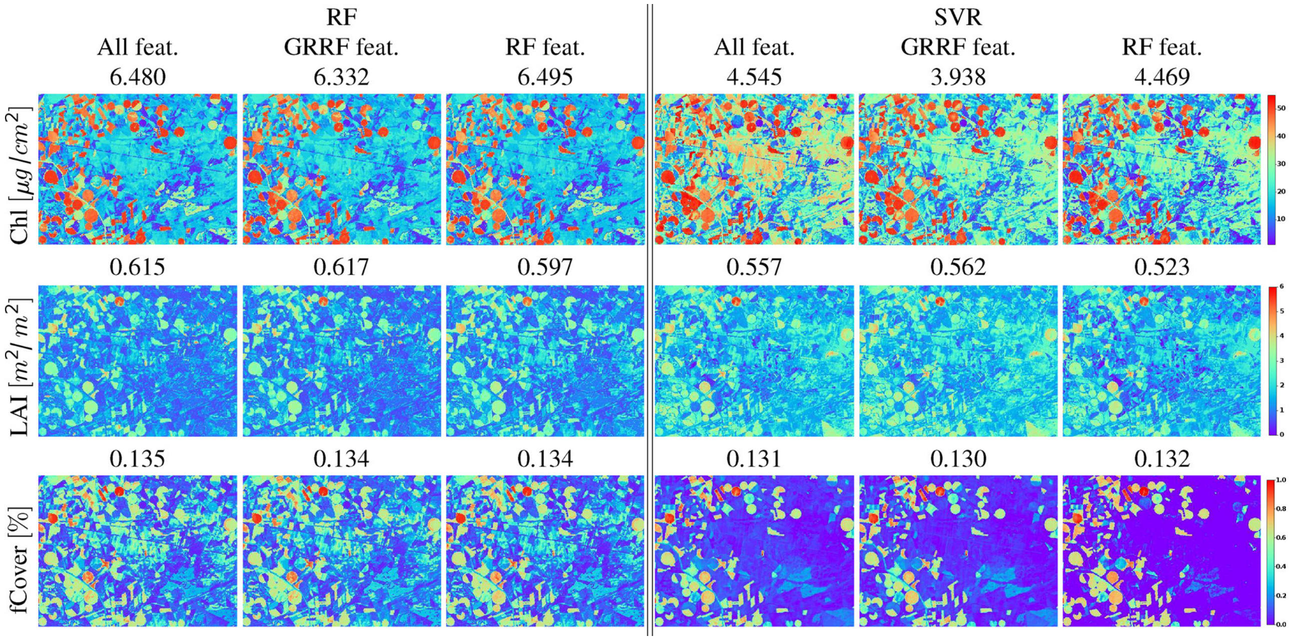

Fig. 12. Predictions maps and their RMSE values for (top) chlorophyll, (middle) LAI, and (bottom) fCover for CHRIS/PROBA using all features (55), and 17, 32 and 21 features, respectively for each biophysical parameter.

results show some notable differences with respect to the classification ones:

1. The errors are less concentrated than the misclassifications. Both high and low RMSE are found across a wide range of λ and γ values.
2. A smaller range of λ and γ values are associated with a significant reduction in the number of features.
3. Several combinations of (λ, γ) yield the same number of selected features (Fig. 9 [bottom]).

These differences show that regression problems are more sensitive to the values of the GRRF parameters.

5.2.3. GRRF vs. standartd RF features

Following the same process chain used in the classification problems, ten partitions of the training data were used to train ten GRRF models with the optimal set of parameters found in the previous step (λ*, γ*). The corresponding sets of selected features are evaluated through the $\overline{\mathrm{RMSE}}$ obtained from standard RF models. The set that yields the minimum $\overline{\mathrm{RMSE}}$ is selected as the best one. The results of this best set of features are compared against the results obtained with the same number of features selected according to their standard feature importance.

Fig. 10 shows the optimum parameters and the relationship between $\overline{\mathrm{RMSE}}$ and λ and γ. In general, GRRF features improve the results compared to RF features. We highlight two key points: (1) unlike classification, where optimum values were typically high λ and low γ, the optimum values for regression correspond to low λ and γ values. And (2) the differences between the feature selection methods are smaller in regression than in classification.

These two points show the importance of properly optimizing the

GRRF parameters for regression problems. Moreover, our regression results indicate that it is important to optimize both parameters (contrarily to what was shown in classification).

The selected features (Fig. 11) mainly focus on the NIR and red range for both selection feature methods. Yet, both methods also selected a few features in the blue range to predict Chl (see variable definition Section 4.2). Compared with the classification case (Fig. 5), the importance of the selected features decreases in a similar fashion for both GRRF and RF.

5.2.4. Biophysical parameter retrieval assessment

Regression results were obtained using standard RF and a Support Vector Regression (SVR) models. These models were trained using 10 partitions of the training samples and tested with the remained samples. The parameters of the SVR were fixed by a 5-fold cross-validation of the model's complexity (C was defined in logarithmic range from 0.1 to 1000 in 10 steps), RBF kernel length-scale ($\sigma \in [0.5, 30]\times$ the mean Euclidean distance of the training data) and the deviation of the predictions from the targets ($\varepsilon \in [0.1, 0.5]$) with the data standardized.

Table 4 shows the statistical metrics derived from the models when using all the features as well as the features selected by GRRF and RF. These latter features sets yielded low errors for both the RF and SVR models. Focusing on the type of prediction, RF results excel in the prediction of Chl, where the use of GRRF features results in up to 2% less error than using all the available features. However, the SVR models are much better with a reduction in the error about 6.5%. Despite these percentages, the statistical metrics are good to very good for all the three variables. Note that, both regression models tend to overestimate the results for the three cases (RF and GRRF selection and all features) for predicting LAI whereas underestimate for the threes cases for predicting Chl. Fig. 12 presents the prediction maps for the three variables included in the CHRIS/PROBA database together with their RMSE. For comparison purposes these maps were prepared using the three cases: all features and the sets of selected features (GRRF and RF). In all the cases, the RF and SVR models lead to low RMSEs. Yet, the SVR models always outperform the RF ones.

Last but not least, Fig. 13 shows the computational time of the 6 prediction models for each variable. The models trained with the selected features are much faster than the models based on all the features. This is particularly important considering the small differences in the prediction results (see Table 4). Like in the classification case, the GRRF and RF models are equally fast but it is important to remember that GRRF does not require a priori knowledge to fix the number of important features.

6. Discussion and conclusions

Efficient data dimensionality reduction methods are of great importance in this new Earth observation era characterized by an ever-increasing access to big geodatabases. In this work, we evaluate a novel feature selection method based on random forests, namely guided regularized random forest or GRRF. Considering that the literature on the feature selection methods is vast, that random forest based methods are the most popular ones amongst the remote sensing literature (Belgiu and Dragut, 2016; Hariharan et al., 2018) and, that embedded methods outperform other feature selection approaches (Pal and Foody, 2010), here we only compare GRRF against the standard use of random forest as feature selector. Our experimental results show that GRRF efficiently identifies the most important features for various classification and regression tasks after optimizing its two parameters. Our experiments also show that this optimization is more critical for regression than for classification tasks, and that the features selected by GRRF are in a different spectral range than the ones selected by the standard RF. Moreover, the use of GRRF leads to a reduction of the data dimensionality of about 80% for the selected classification problems (with only 2.5% decrease in overall accuracies) and of about 60% for the selected regression problem (with virtually no difference in the regression errors). Last but not least, our experimental results show that the GRRF can be successfully in conjunction with the two most popular and robust machine learning methods (i.e. RF and SVM/SVR). Despite the fact that these methods can deal with high dimensional problems, the use of GRRF selected features lead to better results in some of our experiments. Therefore, GRRF offers new possibilities to simplify the analysis of large amounts of Earth Observation data while allowing a deeper analysis of the selected features for classification and regression tasks.

Acknowledgements

The authors would like to acknowledge Prof. Paolo Gamba from the University of Pavia (Italy) for kindly providing the DAIS7915 data over Pavia, Italy. We also wish to express our gratitude to all the STARS project members and, in particular, to the ICRISAT-led team for organizing and collecting the required field data in Mali, and to the STARS ITC team for pre-processing the WorldView-2 images. This research was partially supported by the Bill and Melinda Gates Foundation via the STARS Grant Agreement (1094229-2014) and the grant APOSTD/2017/099 Generalitat Valenciana (Spain).

References

Izquierdo-Verdiguier, E., Zurita-Milla, R., Ault, T.R., Schwartz, M.D., 2018. Development and analysis of spring plant phenology products: 36 years of 1-km grids over the conterminous US. Agric. For. Meteorol. 262, 34–41.

Gómez-Chova, L., Alonso, L., Guanter, L., Camps-Valls, G., Calpe, J., Moreno, J., 2008. Correction of systematic spatial noise in push-broom hyperspectral sensors: application to CHRIS/PROBA images. Appl. Opt. 47 (28), F46–F60.

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press.

Camps-Valls, G., Bruzzone, L., 2005. Kernel-based methods for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 43 (6), 1351–1362.

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A., 2008. Feature Extraction: Foundations and Applications, vol. 207. Springer.

Zurita-Milla, R., Goncalves, R., Izquierdo-Verdiguier, E., Ostermann, F., 2019. Exploring spring onset at continental scales: mapping phenoregions and correlating temperature and satellite-based phenometrics. IEEE Trans. Big Data. https://doi.org/10.1109/TBDATA.2019.2926292. 1–1.

Zurita-Milla, R., Izquierdo-Verdiguier, E., de By, R.A., 2017. Identifying crops in smallholder farms using time series of WorldView-2 images. In: 2017 9th Inter. Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp). IEEE. pp. 1–3.

Izquierdo Verdiguier, E., 2014. Kernel feature extraction methods for remote sensing data analysis. University of Valencia.

Izquierdo-Verdiguier, E., Gómez-Chova, L., Bruzzone, L., Camps-Valls, G., 2014. Semisupervised kernel feature extraction for remote sensing image analysis. IEEE Trans. Geosci. Remote Sens. 52 (9), 5567–5578.

Izquierdo-Verdiguier, E., Laparra, V., Jenssen, R., Gómez-Chova, L., Camps-Valls, G., 2017a. Optimized kernel entropy components. IEEE Trans. Neural Netw. Learn. Syst. 28 (6), 1466–1472.

Haury, A.C., Gestraud, P., Vert, J.P., 2011. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLOS ONE 6 (12), 1–12.

Jović, A., Brkić, K., Bogunović, N., 2015. A review of feature selection methods with applications. 2015 38th Int. Convention on Information and Communication Technology, Electronics and Microelectr. (MIPRO) 1200–1205.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Guan, H., Yu, J., Li, J., Luo, L., 2012. Random forests-based feature selection for land-use classification using Lidar data and orthoimagery. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. 39, B7.

Pal, M., 2005. Random forest classifier for remote sensing classification, Inter. J. Remote Sens. 26 (1), 217–222.

Mureriwa, N., Adam, E., Sahu, A., Tesfamichael, S., 2016. Examining the spectral separability of prosopis glandulosa from co-existent species using field spectral measurement and guided regularized random forest. Remote Sens. 8 (2).

Mutanga, O., Adam, E., Cho, M.A., 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int. J. Appl. Earth Observ. Geoinf. 18, 399–406.

Pal, M., Foody, G.M., 2010. Feature selection for classification of hyperspectral data by SVM. IEEE Trans. Geosci. Remote Sens. 48 (5), 2297–2307.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2004. Random forest classification of multisource remote sensing and geographic data. IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium, vol. 2 1049–1052.

Rodriguez-Galiano, V., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P., Jeganathan, C., 2012. Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens. Environ. 121, 93–107.

Xia, J., Ghamisi, P., Yokoya, N., Iwasaki, A., 2018. Random forest ensembles and extended multiextinction profiles for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 56 (1), 202–216.

Wang, D., Wan, B., Qiu, P., Su, Y., Guo, Q., Wang, R., Sun, F., Wu, X., 2018. Evaluating the performance of sentinel-2, landsat 8 and pléiades-1 in mapping mangrove extent and species. Remote Sens. 10 (9), 1468.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Stat. Comput. 27 (3), 659–678.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31 (14), 2225–2236.

Beckschäfer, P., Fehrmann, L., Harrison, R.D., Xu, J., Kleinn, C., 2014. Mapping leaf area index in subtropical upland ecosystems using rapideye imagery and the randomforest algorithm. IForest-Biogeosci. Forest. 7 (1), 1.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 8 (1), 25.

Deng, H., Runger, G., 2013. Gene selection with guided regularized random forest. Pattern Recognit. 46 (12), 3483–3489.

Dhau, I., Adam, E., Mutanga, O., Ayisi, K.K., 2018. Detecting the severity of maize streak virus infestations in maize crop using in situ hyperspectral data. Trans. R. Soc. S. Afr. 73 (1), 8–15.

Adam, E., Deng, H., Odindi, J., Abdel-Rahman, E.M., Mutanga, O., 2017. Detecting the early stage of phaeosphaeria leaf spot infestations in maize crop using in situ hyperspectral data and guided regularized random forest algorithm. J. Spectrosc. 2017, 8.

Izquierdo-Verdiguier, E., Zurita-Milla, R., de By, R.A., 2017b. On the use of guided regularized random forests to identify crops in smallholder farm fields. 2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp). pp. 1–3.

Kontschieder, P., Bulò, S.R., Bischof, H., Pelillo, M., 2011. Structured class-labels in random forests for semantic image labelling. 2011 Inter. Conf. on Computer Vision 2190–2197.

Chan, J.C.-W., Paelinckx, D., 2008. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sens. Environ. 112 (6), 2999–3011.

Dominik, W.A., 2017. Exploiting the redundancy of multiple overlapping aerial images for dense image matching based digital surface model generation. Remote Sens. 9 (5).

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends® Comput. Graph. Vis. 7 (2-3), 81–227.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc.: Ser. B (Methodol.) 36 (2), 111–133.

Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, pp. 313–325.

Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta-a system for feature selection. Fundam. Inform. 101 (4), 271–285.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Nembrini, S., K&rdquo;onig, I.R., Wright, M.N., 2018. The revival of the Gini importance? Bioinformatics 34 (21), 3711–3718.

Therneau, T.M., Atkinson, E.J., 1997. An Introduction to Recursive Partitioning Using the RPART Routines. Mayo Foundation Tech. rep., Tech. rep.

Deng, H., 2013. Guided random forest in the RRF package. CoRR abs/1306.0237.

Deng, H., Runger, G., 2012. Feature selection via regularized trees. The 2012 International Joint Conference on Neural Networks (IJCNN) 1–8.

Liu, P., Di, L., Du, Q., Wang, L., 2018. Remote sensing big data: theory, methods and applications. Remote Sens. 10 (5).

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees. Advances in Neural Information Processing Systems 431–439.

Vrieling, A., de Beurs, K.M., Brown, M.E., 2011. Variability of African farming systems from phenological analysis of NDVI time series. Clim. Change 109 (3-4), 455–477.

Stratoulias, D., Tolpekin, V., de By, R., Zurita-Milla, R., Retsios, V., Bijker, W., Hasan, M., Vermote, E., 2017. A workflow for automated satellite image processing: from raw VHSR data to object-based spectral information for smallholder agriculture. Remote Sens. 9 (10), 1048.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8 (2), 127–150.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sens. Environ. 25 (3), 295–309.

Haboudane, D., Miller, J.R., Tremblay, N., Zarco-Tejada, P.J., Dextraze, L., 2002. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. Remote Sens. Environ. 81 (2-3), 416–426.

Pietikäinen, M., 2010. Local binary patterns. Scholarpedia 5 (3), 9775.

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sens. Environ. 48 (2), 119–126.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83 (1-2), 195–213.

Yamamoto, H., Hashimoto, T., Seki, M., Yuda, N., Mitomi, Y., Yoshioka, H., Honda, Y., Igarashi, T., 2005. Evaluation of GLI Reflectance and Vegetation Indices with MODIS Products. JAXA Chofu (Japan) Advance Space Tech. Research Group Tech. rep.

Haralick, R.M., Shanmugam, K., et al., 1973. Textural features for image classification. IEEE Trans. Syst., Man, Cybern. (6), 610–621.

Conners, R.W., Trivedi, M.M., Harlow, C.A., 1984. Segmentation of a high-resolution urban scene using texture operators. Comput. Vis., Graph., Image Process. 25 (3), 273–310.

Graña, M., Duro, R.J., 2008. Computational Intelligence for Remote Sensing, vol. 133. Springer.

Aguilar, R., Zurita-milla, R., Izquierdo-verdiguier, E., De By, R.A., 2018. A cloud-based multi-temporal ensemble classifier to map smallholder farming systems. Remote Sens. 10 (5).

Alonso, L., Gómez-Chova, L., Moreno, J., Guanter, L., Brockmann, C., Fomferra, N., Quast, R., Regner, P., 2009. CHRIS/PROBA toolbox for hyperspectral and multiangular data exploitations. IEEE Geosc. Rem. Sens. Symp. (IGARSS), vol. II 202–205.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Box, G.E.P., Meyer, R.D., 1986. An analysis for unreplicated fractional factorials. Technometrics 28 (1), 11–18.

Belgiu, M., Dragut, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogr. Remote Sens. 114, 24–31.

Hariharan, S., Mandal, D., Tirodkar, S., Kumar, V., Bhattacharya, A., Lopez-Sanchez, J.M., 2018. A novel phenology based feature subset selection technique using random forest for multitemporal polsar crop classification. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 11 (11), 4244–4258.