

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800

Open access books available

122,000

International authors and editors

135M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Spectral Optimization of Airborne Multispectral Camera for Land Cover Classification: Automatic Feature Selection and Spectral Band Clustering

*Arnaud Le Bris, Nesrine Chehata, Xavier Briottet  
and Nicolas Paparoditis*

## Abstract

Hyperspectral imagery consists of hundreds of contiguous spectral bands. However, most of them are redundant. Thus a subset of well-chosen bands is generally sufficient for a specific problem, enabling to design adapted superspectral sensors dedicated to specific land cover classification. Related both to feature selection and extraction, spectral optimization identifies the most relevant band subset for specific applications, involving a band subset relevance score as well as a method to optimize it. This study first focuses on the choice of such relevance score. Several criteria are compared through both quantitative and qualitative analyses. To have a fair comparison, all tested criteria are compared to classic hyperspectral data sets using the same optimization heuristics: an incremental one to assess the impact of the number of selected bands and a stochastic one to obtain several possible good band subsets and to derive band importance measures out of intermediate good band subsets. Last, a specific approach is proposed to cope with the optimization of bandwidth. It consists in building a hierarchy of groups of adjacent bands, according to a score to decide which adjacent bands must be merged, before band selection is performed at the different levels of this hierarchy.

**Keywords:** hyperspectral, classification, band selection, spectral optimization, land cover

## 1. Introduction

High-dimensional remote sensing imagery, such as hyperspectral (HS) imagery, generates huge data volumes, consisting of hundreds of contiguous spectral bands. Several difficulties are caused by this high dimensionality. First, the Hughes phenomenon [1] can occur when classifying such data, even though modern classifiers such as support vector machines (SVM) and random forests (RF) are less sensitive to it [2, 3] except when very few training data are available [4]. Second, important computing times are required to process such high-dimensional data. Third, storing

data requires huge volumes. Last, displaying high-dimensional imagery can be necessary, while human vision is limited to three colours [5, 6].

Hyperspectral data consist of hundreds of contiguous spectral bands, but most of these adjacent bands are highly correlated to each other. Thus a subset of well-chosen bands is generally sufficient for a specific problem. This enables to design adapted superspectral sensors dedicated to such specific land cover classification. **Spectral optimization** (SO) or optimal band extraction (BE) consists in identifying the most relevant spectral band subsets for such specific applications. Spectral optimization is a specific dimensionality reduction (DR). DR aims at reducing data volume minimizing the loss of useful information and especially of class separability. Dimensionality reduction techniques can be separated into **feature extraction** (FE) and **feature selection** (FS) categories.

FE consists in reformulating and summing up original information. Principal component analysis (PCA), minimum noise fraction (MNF), independent component analysis (ICA) and linear discriminant analysis (LDA) are examples of state-of-the-art feature extraction techniques. On the opposite, FS selects the most relevant features for a problem. When applied to HS data, it is named band selection (BS) and compared to FE; it enables to keep the physical meaning of the selected bands. For instance, in spectroscopy, FS has sometimes been performed by specialists identifying specific absorption bands or spectrum behaviour corresponding to a material, and this knowledge has then been used in expert systems (e.g. [7] for specific minerals, [8] for asbestos, [9] for asphalt or [10] for urban materials). At the end, SO is at the interplay between FS and FE as it aims at optimizing both band positions (FS) along the spectrum and width (FE).

This study aims at defining a SO strategy to design superspectral sensors dedicated to specific land cover classification problems. SO and FS are optimization problems involving both a **metric** (that is to say a score measuring the relevance of band subsets) to optimize and an **optimization** strategy. This study first focuses on the choice of a FS relevance score suitable for generic optimization heuristics. Both classification performance and selection stability will be considered. As an intermediate result, band importance profiles are considered providing hints about the relevance of the different parts of the spectrum. Once the FS criterion chosen, this chapter copes with the optimization of bandwidth, applying FS within a hierarchy of groups of adjacent bands.

## 2. FS: requirements and state of the art

In the state of the art, FS is often a first step in a specific classification workflow, while the context of this work is the design of superspectral sensors dedicated to specific land cover classification problems. Thus the selected band subset must be as efficient as possible **for most classifiers and not only for the used FS criteria**. Thus, their ability to discriminate between classes using selected feature subsets (that is to say their classification performance) independently from any classifier has to be considered to assess the FS criteria quality. Furthermore, the **stability** of the proposed solutions has also to be considered. Last but not least, in this sensor design context, constraints about the **maximum** number of bands to select exist. To sum it up, a good FS criterion for sensor design has to **be parsimonious, making it possible to select stable band subsets discriminant for most classifiers**. Thus, for a fair analysis, FS criteria must be compared for a same selected band subset size, and results must be evaluated according to different classifiers. Besides, computing time was not considered as an important criterion in this specific context of sensor design, where FS is not a preprocessing in a classification workflow.

Thus, this study focuses on the comparison of several FS **criteria** (presented in Section 2.1) for **supervised** classification problems (that is to say when classes and their ground truth are taken into account). To have a fair comparison, all these criteria will be optimized using the **same generic optimization** algorithms. It was here decided to use such generic optimization heuristics, in the context of sensor design, since such methods enable to easily control the number of bands to select and to add additional constraint within the band extraction process as in the second part of the study. The use of generic optimization methods necessarily excludes of the comparison feature ranking criteria (such as ReliefF [11, 12]) and FS methods where the score and the optimization method are strongly related to, for instance, SVM-RFE [13]. All criteria will be tested on several classic hyperspectral data sets.

## 2.1 FS: state of the art

Even though hybrid approaches involving several criteria exist [14, 15], FS methods and criteria are often differentiated between ‘filter’ (independent from any classifier), ‘wrapper’ (related to the classification performance of a classifier) and ‘embedded’ (related to the quality of classification models estimated by a classifier, but not directly to classification accuracy). It is also possible to distinguish supervised and unsupervised ones, especially for filters, that is to say whether a notion of classes is taken into account or not. All approaches mentioned below are summed up in **Tables 1** and **2**. Nevertheless, it must be kept in mind that hybrid approaches involving several criteria belonging to these different FS criteria categories often exist, as, for instance, in [14] or [15], where features are selected based on a wrapper method, respectively, guided or associated with filter criteria (mutual information between selected bands and between the ground truth).

### 2.1.1 Filter

Filter methods compute relevance scores independently from any classifier. Some filter methods are ranking approaches: features are ranked according to an individual score of importance. Such individual feature scores can be supervised or

Filters	Wrapper
<b>Unsupervised :</b>	<b>Classification quality rates reached by :</b>
<b>Individual importance scores :</b>	<u>SVM classifier</u> [23, 42]
PCA or ICA or between band correlation [14, 21]	[65, 70]
[15, 30]	<u>Maximum likelihood classifier</u> [25, 67]
<b>Feature subset scores :</b>	<u>Random Forests classifier</u> [20]
<u>Mutual information</u> [41, 44]	Target detector [46]
<u>Linear combination of bands</u> [22]	
<b>Other :</b>	<b>Embedded</b>
SVC (Support Vector Clustering) [9]	<b>Regularization models :</b>
Method based on between band correlation [15]	L1-SVM [69]
	Lasso [59, 61]
	Other regularization models [43, 62]
<b>Supervised :</b>	<b>Built-in feature selection :</b>
<b>Individual importance scores :</b>	Random Forests classifier [6]
ReliefF [37, 38]	Decision Tree classifier [7]
Other scores based on correlation or mutual information [29]	
	<b>Individual feature importance scores :</b>
<b>Feature subset scores :</b>	Importance measures from Random Forests [6]
<b>Separability measures :</b>	Multiple kernel SVM [60]
<u>Separability measures : Bhattacharyya</u> [8, 32]	
<u>or Jeffries-Matusita distances</u> [18, 55]	<b>Pruning methods :</b>
[67]	SVM-RFE [28]
<u>Minimum Estimated Abundance Covariance</u> [64]	
<b>Correlation to ground truth :</b>	<b>Generalization performance :</b>
<u>Mutual information</u> [2, 23, 27]	SVM margin [26, 49]
[11, 56]	<u>Out-Of-Bag error of a Random Forests classifier</u> [26, 49]
	<b>Other :</b>
	Unmixing process [66]

**Table 1.**

*State of the art of feature selection criteria: the criteria that work with the FS criteria evaluation framework used in this study are underlined.*

Filters
<p><b>Unsupervised :</b></p> <p><b>Individual importance scores :</b>  <u>Pro :</u> Fast.  <u>Against :</u> Ranking selection, without considering redundancies between selected features.                      Solutions not parsimonious. Can be used to guide optimization.</p> <p><b>Feature subset scores and linear band combinations :</b>  <u>Pro :</u> Select the most representative and the less redundant features.  <u>Against :</u> How to balance between these two criteria ?  <u>Against :</u> Not necessarily the best solution (parsimonious/performance) for a classification problem.</p>
<p><b>Supervised :</b></p> <p><b>Individual importance scores :</b>  <u>Pro :</u> Fast.  <u>Against :</u> Ranking selection, without considering redundancies between selected features.                      Solutions not parsimonious. Can be used to guide optimization.</p> <p><b>Feature subset scores :</b>  <b>Separability measures :</b>  <u>Pro :</u> Fast.  <u>Pro :</u> Generally good solutions.  <u>Pro :</u> Consider redundancies between selected features.  <u>Against :</u> Classes modelled by an underlying parametric model (e.g. : Gaussian model).</p> <p><b>Correlation to ground truth :</b>  <u>Pro :</u> Fast.  <u>Pro :</u> Select features that are the most correlated to the ground truth and the least redundant.  <u>Against :</u> How to balance between these two criteria ?</p>
<p><b>Wrapper</b></p> <p><u>Pro :</u> Good results, especially for the classifier used to calculate the FS score.  <u>Pro :</u> Consider (implicitly) both the redundancy between selected features and their classification performance.  <u>Against :</u> Possible more important computing times depending on the used classifier  <u>Against :</u> Risk to be dependent on the used classifier</p>
<p><b>Embedded</b></p> <p><b>Regularization models :</b>  <u>Pro :</u> Good results.  <u>Pro :</u> Fast.  <u>Pro :</u> Consider redundancies between selected features.  <u>Against :</u> More difficult to adapt them (e.g. to introduce additional other constraints)  <u>Against :</u> Specific (not generic) optimization algorithm</p> <p><b>Built-in feature selection :</b>  <u>Pro :</u> Fast.  <u>Against :</u> Not parsimonious.</p> <p><b>Individual feature importance scores and pruning methods :</b>  <u>Pro :</u> Fast.  <u>Against :</u> Solutions not necessarily parsimonious (especially in case of a ranking FS).  <u>Against :</u> Specific optimization method, not generic in case of pruning methods.</p> <p><b>Generalization performance :</b>  <u>Pro :</u> can be used like a wrapper while being faster.  <u>Pro :</u> Consider (implicitly) redundancies between selected features.  <u>Against :</u> Possible difference between the generalization performance and the classification performance of a model.</p>

**Table 2.**  
*Pros and cons for the different families of FS criteria.*

unsupervised. For instance, the well-known ReliefF score [11, 12] or scores measuring the correlation between features and ground truth [23] are supervised ones. However, such individual feature importance measures do not take into account the correlations between selected features. Thus, a feature subset composed of the  $n$  best features according to such measures is not necessarily an optimal solution, in the sense that it is not parsimonious.

Other ranking methods are unsupervised: they use importance measures calculated from a feature extraction technique. For instance, [16] ranks bands according to a score of importance calculated from PCA decomposition. Correlated bands are then removed according to a divergence measure. Du et al. and Hasanlou et al. [17, 19] have a similar approach using ICA instead of PCA. Other

unsupervised approaches also use results of a PCA selecting the most similar features to the first PCA [48, 49].

Other filter approaches associate a score to feature subsets. In unsupervised case, [18] also performs a constrained energy minimization to select a set of bands having minimum correlation between each other. In supervised cases, separability measures such as Bhattacharyya or Jeffries-Matusita (JM) distances can be used in order to identify the best feature subsets for separating classes [25–27, 50]. Other separability measures based on the minimum estimated abundance covariance (related to the ability of the band subset to correctly unmix several sources) have also been used as in [47].

High-order statistics from information theory such as divergence, entropy and mutual information can also be used to select the best feature sets achieving the minimum redundancy and the maximum relevance, either in unsupervised situations as in [6, 38] or in supervised ones as in [14, 30, 31, 51, 52]. Martínez-Usó et al. [38] first clusters ‘correlated’ features and then selects the most representative feature of each group. Le Moan et al. [6] selects the three bands belonging to three red, green and blue spectral domains so that their correlation is minimized. In supervised cases, [14, 30, 31, 51, 53] select the set of bands that are more correlated to the ground truth and less correlated to each other. The most difficult is then to balance both criteria.

The orthogonal projection divergence [21] is another way to measure correlation between bands by the extent to which it is possible to express one band as a linear combination of the already selected bands. Last, [22] uses support vector clustering applied to features in order to identify the most relevant ones.

To sum it up, there are many various filter criteria corresponding to different approaches. Ranking methods according to an individual feature importance score remain limited, especially the ones only based on a supervised score, since they are not aware of the dependencies between selected features. Filter approaches associating a score to feature subsets are more interesting. Supervised and unsupervised approaches can be distinguished. Unsupervised approaches are interesting, but in a classification context, there is still a risk to select features that will not be all useful for the classification problem.

### *2.1.2 Wrapper*

Wrapper relevance score associated with a feature set simply corresponds to its corresponding classification performance (measured by an accuracy score). Examples of such scores can be found in [14, 15, 24, 54] using SVM classifier, [36, 39] maximum likelihood classifier, [37] random forests, [49] spectral angle mapper or [55] a target detection algorithm.

### *2.1.3 Embedded*

Embedded FS methods are also related to a classifier, but feature selection is performed using a feature relevance score different from a classification accuracy. Most of the time, embedded approaches directly select features during the classifier training step. Several types of embedded FS approaches can be distinguished [45].

Some embedded approaches are regularization based models. A classifier is trained according to an objective function where a fit-to-data term that minimizes the classification error is associated with a regularization function, penalizing models when the number of features increases or forcing model coefficients associated with some features to be small. Features with the coefficients close to 0 are eliminated. Examples of some approaches can be found in [29, 56, 57]. They also

include the L1-SVM [58] and the least absolute shrinkage and selection operator (LASSO) FS [41, 57] approaches. Such approaches are fast and efficient. However, it can be more difficult to adapt them, for instance, to take into account additional constraints, since FS criterion and optimization method are linked.

Other embedded approaches use the built-in mechanism for feature selection in the training algorithm of some classifiers. For instance, random forests (RF) [43] and decision trees can be considered as performing an embedded feature selection, since, when splitting a tree node, only the most discriminative feature according to Gini impurity criterion is used among a feature subset randomly selected [43]. This FS eliminates the less useful features, but there is no guarantee to select a parsimonious feature subset: redundant features can be selected.

Some embedded approaches also provide feature importance measures, such as random forest classifier [43]. It is processed on samples left out of the bootstrapped samples and is based on the permutation decrease accuracy: the importance of a feature is estimated by randomly permuting all its values in these samples for each tree, as the difference averaged over all the trees between prediction accuracy before and after permuting this feature. Other embedded approaches providing feature importance use them in a pruning process that first uses all features to train a model, before progressively eliminating some of them while maintaining model performance. SVM-RFE [13] is a well-known embedded approach where the importance of the different features in a SVM model is considered. Such approach has been extended to multiple kernel SVM by [42], associating a different kernel to each feature, estimating the model and then using the weights associated with these kernels as feature importance measures.

Other approaches do not calculate a score of importance for each feature individually, but evaluate the relevance of sets of features. Such scores often measure the generalization performance of the obtained model. Thus, the FS is not directly performed during the training step, but uses an intermediate result of the training step. For instance, [46, 59] use the generalization performance, e.g. the margin of a SVM classifier, as a separability measure to rank sets of features. The out-of-bag error rate of a random forest [43] can also be considered as such score. These scores are calculated for feature subsets and measure the generalization performance of the model provided by the classifier. Thus, they can be considered as an alternative between filter separability measures and wrapper scores.

Embedded approaches can also be extended to unmixing methods, as, for instance, in [60] where band selection is integrated into an endmember and abundance determination algorithm by incorporating band weights and a band sparsity term into an objective function.

#### *2.1.4 Optimization methods*

Another issue for band selection is to determine the best set of features corresponding to a given criteria. An exhaustive search is often impossible, especially for wrapper techniques. Therefore, heuristics have been proposed to find a near-optimal solution without visiting the entire solution space. Optimization methods can be either specific to a FS method (as for most embedded ones) or generic. Generic optimization methods can be divided into two groups: sequential and stochastic.

Several incremental search strategies have been detailed in [28], including the sequential forward search (SFS) starting from one feature and incrementally adding another feature making it possible to obtain the best score or on the opposite the sequential backward search (SBS) starting for all possible features and incrementally removing the worst feature. Variants such as sequential forward floating

search (SFFS) or sequential backward floating search (SBFS) are proposed in [28]. Serpico and Bruzzone [32] also proposes a variant of these methods called steepest ascent (SA) algorithms.

Among stochastic optimization strategies used for feature selection, several algorithms have been used for feature selection, including genetic algorithms [14, 15, 46, 54, 55], particle swarm optimization (PSO) [24, 47], clonal selection [39], ant colony [35] or even simulated annealing [27, 61].

In the specific case of hyperspectral data, adjacent bands are often very correlated to each other. Thus, hyperspectral band selection faces the problem of the clustering of the spectral bands. Band clustering/grouping has sometimes been performed in association with individual band selection. For instance, [15] first groups adjacent bands according to conditional mutual information and then performs band selection with the constraint that only one band can be selected per cluster. Su et al. [40] performs band clustering applying k-means to band correlation matrix and then iteratively removes the too inhomogeneous clusters and the bands that are too different from the representative of their cluster. Martínez-Usó et al. [38] first clusters 'correlated' features and then selects the most representative feature of each group, according to the mutual information. Chang et al. [61] performs band clustering using a more global criterion taking specifically into account the existence of several classes. Simulated annealing is used to maximise a cost function defined as the sum, over all clusters and over all classes, of correlation coefficients between bands belonging to a same cluster.

### 3. Which band selection criterion?

This study is a comparison of **FS criteria that can be optimized using generic optimization heuristics**, thus excluding several specific embedded or ranking approaches. The following FS criteria (listed in **Table 3**) were evaluated.

#### 3.1 Compared FS criteria

##### 3.1.1 Filter FS criteria

Filter criteria are independent from any classifier. Only scores assessing the relevance of feature subsets were considered, excluding filter FS methods ranking features independently according to an individual feature score (e.g. ReliefF).

##### 3.1.1.1 Separability

Separability measures are used to identify the feature subsets achieving the best class distinction. Fisher, Bhattacharyya and Jeffries-Matusita measures [25–27, 50] are such scores. They were used assuming Gaussian class models. Let  $\vec{\mu}_i$  and  $\Sigma_i$  be the mean and covariance matrices of the spectral distribution of class  $i$ . Fisher separability between classes  $i$  and  $j$  is defined in equation (1)

$$F_{ij} = \frac{(\vec{w} \cdot (\vec{\mu}_i - \vec{\mu}_j))^2}{\vec{w}^t (\Sigma_i + \Sigma_j) \vec{w}} \text{ where } \vec{w} = (\Sigma_i + \Sigma_j)^{-1} (\vec{\mu}_i - \vec{\mu}_j) \quad (1)$$

Bhattacharyya separability between classes  $i$  and  $j$  is defined by equation 2.



Wrapper criteria	
<i>svm.lin.K</i>	Kappa of linear 1vs1 SVM
<i>dt.K</i>	Kappa of Decision Tree (DT)
<i>rf.K</i>	Kappa of Random Forests (RF)
<i>ml.K</i>	Kappa of Maximum Likelihood classifier (ML)
<i>sam.K</i>	Kappa of Spectral Angle Mapper (SAM)
<i>sid.K</i>	Kappa of Spectral Information Divergence (SID)
<i>svm.lin.conf</i>	confidence FS score of linear 1vs1 SVM
<i>rf.conf</i>	confidence FS score of Random Forests (RF)
Embedded criteria	
<i>svm.lin.marg</i>	Margin of a linear 1vs1 SVM
<i>rf.oob</i>	Out-of-Bag accuracy of Random Forest
Filter criteria	
<i>fisher</i>	Mean Fisher separability measure $= \frac{1}{nb\_pairs\_of\_classes} \sum_{i=1}^{c-1} \sum_{j=i+1}^c F_{ij}$
<i>Bdist</i>	Mean Bhattacharyya separability measure [8] $= \frac{1}{nb\_pairs\_of\_classes} \sum_{i=1}^{c-1} \sum_{j=i+1}^c B_{ij}$
<i>jm</i>	Jeffries-Matusita separability measure [8] $JM = \sum_{i=1}^{c-1} \sum_{j=i+1}^c (1 - e^{-B_{ij}})$
<i>mi</i>	Mutual Information based criterion [23] $J(S) = \sum_{f \in S} I(C, f) - \frac{1}{\#S} \sum_{f \in S} \sum_{s \in S; s \neq f} \frac{I(f, s)}{H(f) \cdot H(s)}$

**Table 3.**  
Selected FS criteria to be compared.

$$B_{ij} = \frac{1}{8} (\vec{\mu}_i - \vec{\mu}_j)^t \Sigma^{-1} (\vec{\mu}_i - \vec{\mu}_j) + 0.5 \ln \left( \frac{\det \Sigma}{\sqrt{\det \Sigma_i \det \Sigma_j}} \right) \text{ where } \Sigma = \frac{\Sigma_i + \Sigma_j}{2} \quad (2)$$

As Bhattacharyya and Fisher separability measures are defined for binary problems, their mean overall possible pairs of classes were here used as FS criteria. To sum it up, the next separability measures were used as FS criteria:

- Mean Fisher (*fisher*) separability measures calculated over all pairs of classes (equation 3):

$$\frac{1}{nb\_pairs\_of\_classes} \sum_{i=1}^{c-1} \sum_{j=i+1}^c F_{ij} \quad (3)$$

- Mean Bhattacharyya (*Bdist*) separability measures calculated over all pairs of classes (equation 4):

$$\frac{1}{nb\_pairs\_of\_classes} \sum_{i=1}^{c-1} \sum_{j=i+1}^c B_{ij} \quad (4)$$

- Jeffries-Matusita measure (*jm*) defined in equation 5:

$$JM = \sum_{i=1}^{c-1} \sum_{j=i+1}^c (1 - e^{-B_{ij}}) \quad (5)$$

### 3.1.1.2 Mutual information

Another FS criterion based on high-order statistics from information theory, e.g. mutual information (MI), was adapted from [14] and tested: it took into account both feature-class dependencies and between feature correlations. It is defined in equation 6.

$$J(S) = \sum_{f \in S} I(C, f) - \frac{1}{\#S} \sum_{f \in S} \sum_{s \in S; s \neq f} \frac{I(f, s)}{H(f) \cdot H(s)} \quad (6)$$

for a feature subset  $S$ , where  $I(C, f)$  is the MI between feature  $f$  and classes,  $I(f, s)$  is the MI between features  $f$  and  $s$  and  $H(f)$  is the entropy of feature  $f$ . It is referred to as *mi* in **Table 3**.

### 3.1.2 Wrapper and embedded FS criteria

Several classifiers were used to define wrapper scores related to their classification performances achieved using feature subsets. Only fast classifiers which did not require an optimization of hyper-parameters were used:

- **Maximum likelihood classification (ML)**: assuming a Gaussian model for the spectral distribution of classes, mean vectors and covariance matrices are estimated for each class during the training step. Each new sample is then labelled by its most probable class according to the model.
- **SAM and SID**: these classifiers are specific to hyperspectral data. The *spectral angle mapper* (SAM) consists in classifying a sample according to the angle between its spectrum and reference spectra. The *spectral information divergence* [62] comes from dissimilarity measures between statistical distributions and more precisely the Kullback-Leibler measure.
- **Support Vector Machine (SVM)** [63]: SVM has been intensively used to classify remote sensing data and especially hyperspectral data [2, 15, 64]. Training a SVM classifier aims at estimating the best frontiers between classes. Only a one-against-one linear SVM was used here. Indeed, it is fast and enables to avoid an optimization of hyper-parameters, contrary to other kernels. Besides, using a linear SVM introduces a constraint to select bands achieving a linear separation between classes.
- **Decision trees (DT)** [44].
- **Random forests (RF)** [43] is a modification of bagging applied with decision trees. It can achieve a classification accuracy comparable to boosting [43] or SVM [65]. It does not require assumptions on the distribution of the data, which is interesting when different types or scales of input features are used. It was successfully applied to remote sensing data such as multispectral data, hyperspectral data or multisource data. This ensemble classifier is a combination of tree predictors built from multiple bootstrapped training samples. For each node of a tree, a subset of features is randomly selected. Then, the best feature with regard to Gini impurity measure is used for node splitting. For classification, each tree gives a unit vote for the most popular class at each input instance, and the final label is determined by a majority vote of all trees.

These different classifiers were chosen because their underlying principles were different from each other. SAM, SID and ML rely on class models, while the others use inter-class separation models. RF can model even complex class frontiers remaining quite fast, while linear SVM selects features achieving the most possible linear separation between classes.

Wrapper FS scores measuring classification performance were considered:

- **Kappa coefficient:** for all of these classifiers, the *Kappa* coefficient has been used as a FS score.
- **Classification confidence score:** in addition, another FS score taking into account the classification confidence was also used [66]. Indeed, most classifiers provide classification confidence indices and a class membership measuring the degree to which the sample belongs to the different classes according to the classifier. Let  $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n}$  be a set of labelled ground truth samples  $x_i$  and their associated true label  $y_i$ . Let  $m(x, c)$  be the class membership measuring the probability for  $\mathbf{x}$  to belong to class  $c$ . A possible feature selection score  $\mathcal{R}$  taking into account class membership measures and thus classification confidence can be defined by equation 7:

$$\mathcal{R}(\mathcal{X}) = \sum_{i=1}^n \delta(y_i, c(\mathbf{x}_i)) \cdot m(\mathbf{x}_i, c(\mathbf{x}_i)) \quad (7)$$

with  $\delta(i, j) = \{-1 \text{ if } i \neq j \text{ and } 1 \text{ otherwise}\}$  and  $c(\mathbf{x})$  the label given to  $\mathbf{x}$  by the classifier. Such score measures both the ability to well classify the test samples for a given feature set and the separability between classes. Indeed, the more the samples are well classified, the more the score increases. The more the classifier is confident for well-classified samples, the more the score increases. The more the classifier is confident for bad-labelled samples, the more the score decreases. This confidence score was used in our experiments only for RF and linear SVM classifiers.

Embedded FS criteria. The two following criteria measuring the generalization performance of two classifiers were also tested. They are not pure embedded but can be considered as intermediate between wrapper and embedded. However, differentiating them from previous common wrapper scores, they are here referred to as ‘embedded’ in the sense that they assess the classification performance directly using a measure calculated directly while training the classifier and not after an evaluation of the model on a test data set. These scores are:

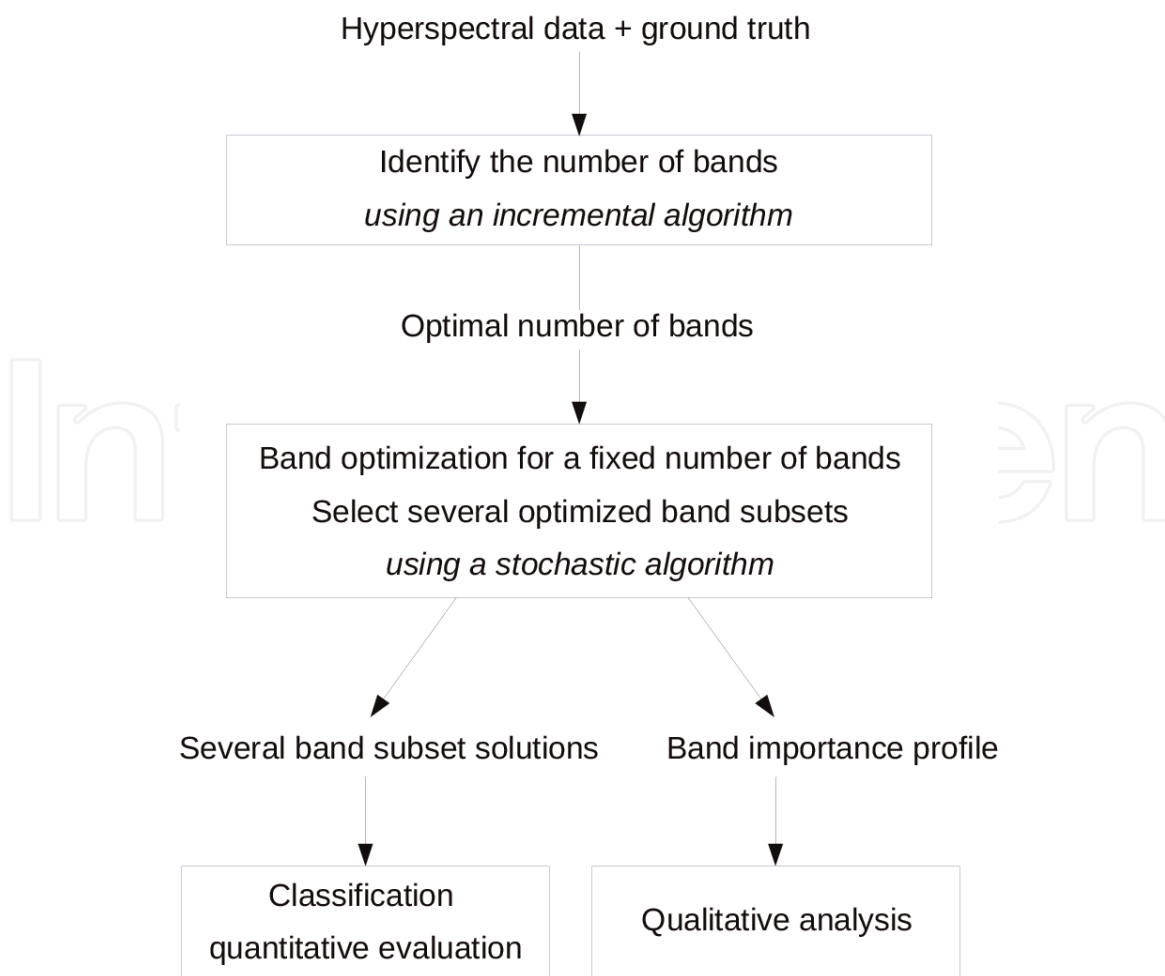
- The margin of a linear 1-vs-1 SVM classifier (without parameter optimization) (*svm.lin.marg*), that is to say the distance between the class frontier and its support vectors.
- The out-of-bag error [43] of a RF classifier (*rf.oob*). The out-of-bag samples are left out of the bootstrapped samples when training the RF.

### 3.2 Assessment approach

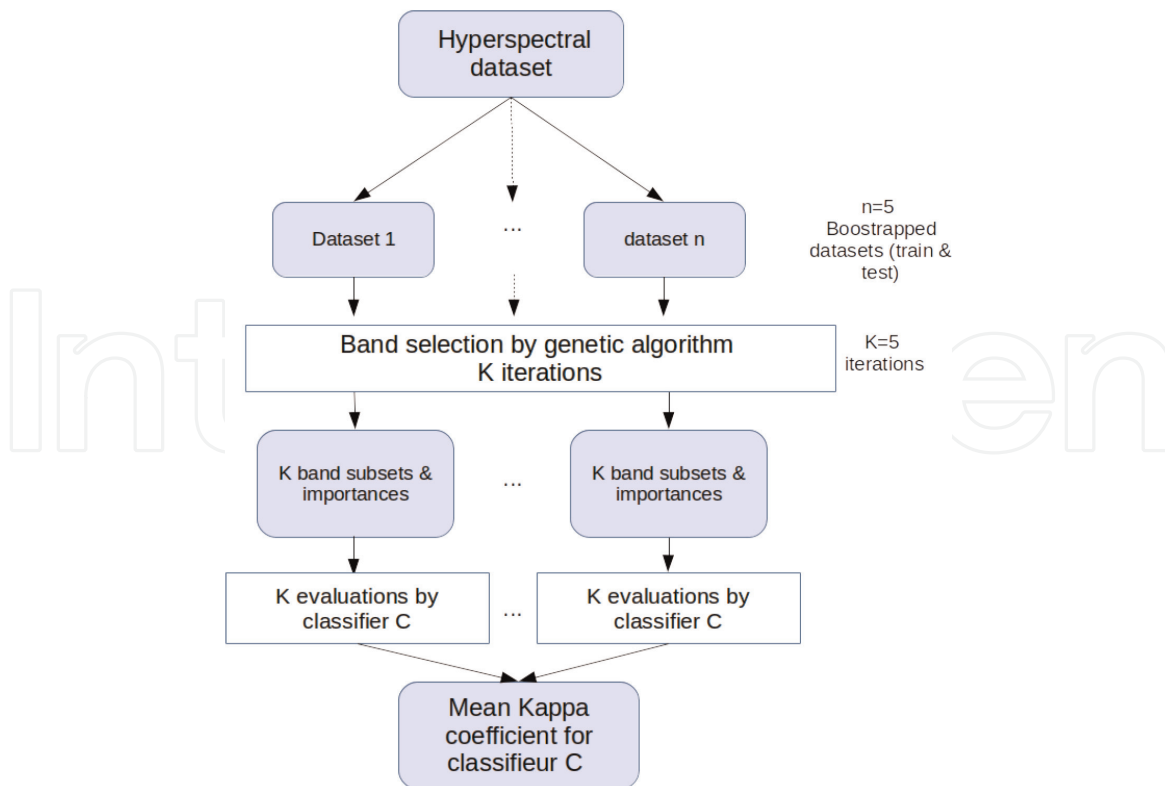
It must be kept in mind that study is a comparison of FS criteria and not of optimization methods. Thus all were optimized using the same optimization heuristics on the same classic hyperspectral data sets (3.3). The proposed workflow

(**Figure 1**) includes two steps. The suitable number of bands to select is first estimated for each data set, thanks to an incremental FS optimization algorithm called sequential forward floating search (SFFS) [28]. Then, the core comparison of FS criteria was performed. They were optimized to select this fixed number of bands using a stochastic FS optimization algorithm. A genetic algorithm (GA) (3.2.2) was used. Indeed, it proved to be efficient and generic enough to be used for all tested criteria. Besides it can provide valuable intermediate results (3.2.4) to assess FS stability. GA was launched several times to select this fixed number of bands for all tested FS criteria. It thus provided several possible band subset solutions. Indeed, performing FS several times was also a way to benefit from the stochastic nature of GA and thus to explore more band subset configurations. These different solutions were then quantitatively evaluated, according to different classifiers, to be able to draw conclusions about their relevance quite independently from a given classifier (3.2.3). Besides, to perform a qualitative analysis of the obtained solutions (and especially their stability), band importance measures were derived from intermediate results provided by this stochastic FS (3.2.2). It enabled to visually identify the parts of the spectrum considered as important by the FS criterion and to have a qualitative analysis concerning the stability of the proposed band subset solutions according to the FS criterion.

In practice, for each FS criterion, the GA feature selection process was launched five times on five limited data sets (100 training and 500 (300 for Indian Pines) testing samples) randomly selected with replacement among the whole data set. To sum it up, at the end, 25 ‘optimal’ feature subset solutions were thus obtained for each criterion and had to be evaluated (**Figure 2**).



**Figure 1.**  
*Assessment process.*



**Figure 2.** Evaluation of FS criteria using band subsets obtained using a GA optimization.

### 3.2.1 Optimal band subset size using a sequential FS algorithm

Intermediate results of a sequential FS algorithm were used to identify how many bands must be selected. In our experiments, the sequential forward floating search (SFFS) algorithm was used [28].

This optimization method provides useful intermediate results. Indeed, it selects the ‘best’ sets of bands for different band subset sizes, starting from 1. Thus, it provides for each of them both the selected band subset (that could then be evaluated according to the performance of several classifiers) and the value reached by the FS score. Therefore, it enables to observe the evolution of FS score and classification quality, with the number of selected bands and then to decide how many bands are necessary to obtain suitable results. Other sequential methods as SVM-RFE [13] or SFS could also provide such information, but contrary to them, SFFS has the advantage to question at each step the selected set of bands obtained at the previous step, which enables possible modifications in the already selected band subset.

### 3.2.2 Band subset solutions using a genetic algorithm

Genetic algorithm (GA) is a family of stochastic optimization heuristics simulating the evolution mechanisms on a population of individuals. A score measuring its adaptation and its aptitude to stay alive is associated with each individual. In FS context, each individual is a feature subset and the score is the FS score.

---

#### **Algorithm 1** Genetic algorithm.

---

It is intended to select less than  $p$  bands among a band set  $\mathcal{B}$ .  $J$  is the FS score to optimize.

```
Initialization: ( $t \leftarrow 0$ ) Randomly generate a population  $G(0)$  of  $N$  individuals,
i.e.  $N$  sets of  $p$  bands.
while  $t < t_{max}$  do
  //generation loop
   $t \leftarrow t+1$ 
  Calculate the score of each band subset of the current population.
  Keep only the  $n$  ( $n < N$ ) best band subsets of the current population. Let  $R(t)$ 
  be this remaining population.
  Generate a new population  $G(t)$  of  $N$  individuals from  $R(t)$ :
  for all new individual do
    Randomly select 2 parents among  $R(t)$ .
    Obtain a new individual by randomly crossing these 2 parents.
    Random mutations occur (randomly replacing a selected band by another
    one) in order to avoid to stay in a local optimum.
  end for
end while
```

---

### 3.2.2.1 GA-derived importance measures

The GA approach has some advantages for our problem. First, only the best solution is usually kept, while GA has visited many other candidates. Many of them have scores quite similar to the score of the best solution: they are almost as good as the final solution. Therefore, these intermediate results can be used to determine which bands are often selected in the solutions (see **Figure 3**) of these intermediate good band subset populations [67]. Thus, an individual band importance score  $I(b)$  (defined in equation 8) is calculated for each band  $b$ , measuring the occurrence at which it has been selected by GA among the different  $n$  best sets of bands obtained for all generations

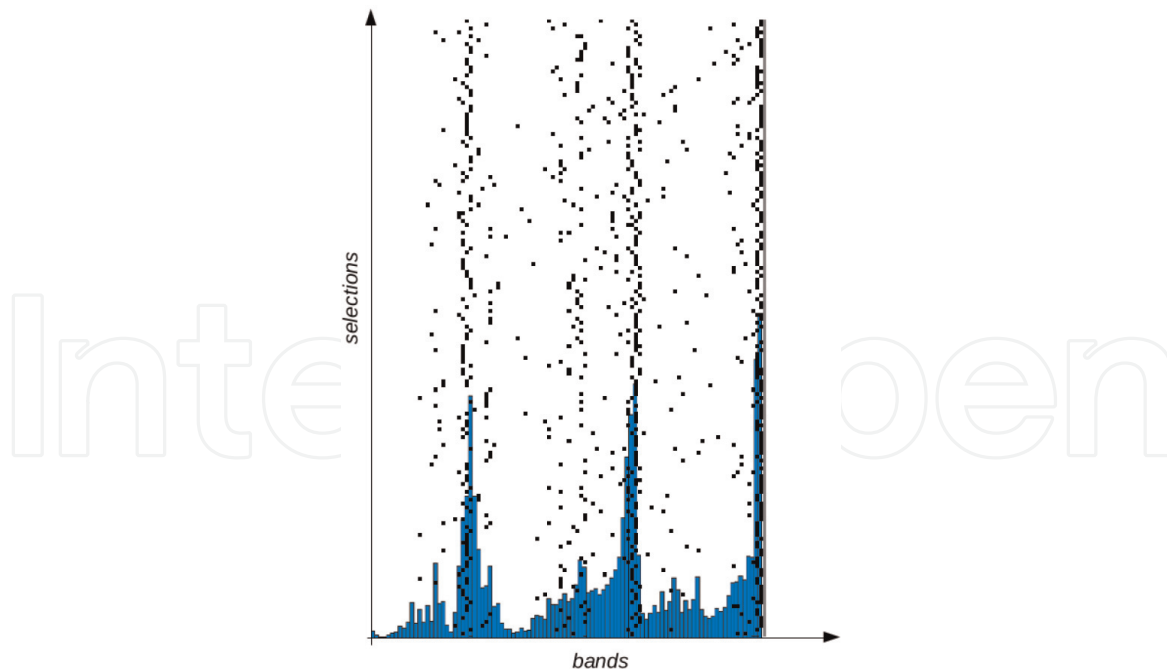
$$I(b) = \sum_t \sum_{R \in L(t)} \delta(b, R(t)) \text{ where } \delta(b, R) = 1 \text{ if } b \in R, 0 \text{ otherwise.} \quad (8)$$

To increase robustness, GA can be launched several times (i.e. so that different initializations and mutations occur) and over several training/testing sets randomly extracted from the whole data set. The proposed importance score is calculated for each of these results. Finally, the mean of these scores is considered for each band, giving the importance associated with this band.

### 3.2.3 Quantitative evaluation

In state of the art, FS is often considered as a first step in a specific classification workflow. In this context, wrappers are considered as achieving the best classification performance for a problem while sometimes lacking generality and being too classifier dependent. However, in our superspectral sensor design context, selected band subsets must be as efficient as possible for most classifiers and not only for the used FS criteria. Therefore, selected band subsets were here evaluated considering their classification quality reached with several classifiers.

Kappa coefficient was used as classification quality measure for the next classifiers: ML, RF and 1-vs-1 SVM with a radial basis function (RBF) kernel (with optimized parameters). It can here be noted that the latter was the only one not involved previously in a tested FS criterion. Thus, RBF SVM is the only classifier that is completely independent from all tested FS criteria. To come into details,



**Figure 3.** Each line is a band subset selected in the intermediate results of GA, and each black dot represents a selected band. Blue histogram represents the importance associated with each band.

evaluation was performed and averaged on five training/testing sample sets: for each of them, classifiers were trained using 50 samples per class (in order to be in a difficult case with few training samples), and results were evaluated on all remaining ground truth samples. For each FS criterion, all selected band subsets (obtained for the several launches of the algorithm) were evaluated, and the mean Kappa coefficient was then computed over all of them (see **Figure 2**).

#### 3.2.4 Selected band stability

Another evaluation criterion of the FS criteria quality was the stability of the selected features. As explained in Section 3.2.2, band importance profiles (**Figure 3**) can be derived from intermediate results of a GA feature selection. As the contiguous bands in hyperspectral data are correlated, such band importance profile should be quite regular and smooth (i.e. not too noisy). The smoothness/regularity of these profiles is thus related to the stability of the solutions obtained using a FS criterion. Furthermore, the final optimal solutions provided by the different launch of GA can also be examined. This analysis remains only qualitative.

### 3.3 Data sets

Three state-of-the-art available hyperspectral data sets were used for the experiments:

- **Pavia City Centre scene**<sup>1</sup>: This first data set is a hyperspectral scene acquired by the ROSIS sensor over the city centre of Pavia with a 1.3 m spatial resolution. It is a reflectance VNIR hyperspectral image with a

<sup>1</sup> Pavia data set is provided by Pavia University available at [http://www.ehu.us/ccwintco/index.php?title=Hyperspectral\\_Remote\\_Sensing\\_Scenes](http://www.ehu.us/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes).

spectral resolution ranging from 460 nm to 860 nm. Noisy bands have been discarded, and only 102 spectral bands from the original 115 bands have been kept. It covers an urban area (city centre). Its associated land cover ground truth consists of nine urban classes (materials and vegetation).

- **Indian Pines scene**<sup>2</sup>: This hyperspectral scene was collected by the AVIRIS sensor over the Indian Pines test site in North-western Indiana. It is a radiance VNIR-SWIR hyperspectral image consisting of 220 spectral bands ranging from 400 to 2500 nm. Its associated ground truth consists of agricultural classes and other classes concerning perennial vegetation (forest, grass). In our experiments, only nine classes out of the original were kept. The discarded classes concerned less than 400 samples, which were considered as too few for our experiments.
- **Salinas scene**<sup>3</sup>: This hyperspectral scene was collected by the AVIRIS sensor over the Salinas Valley in California at a 3.7 m spatial resolution. It is an at-sensor radiance VNIR-SWIR hyperspectral image consisting of 224 spectral bands ranging from 400 to 2500 nm. Its associated ground truth consists of agricultural classes, that is to say different kinds of culture at different growing steps.

### 3.4 Results and discussion

#### 3.4.1 Optimal number of bands using SFFS

An optimal number of bands to select was identified using SFFS incremental FS method, starting from one selected band and incrementing the band subset until a maximal number of bands. Indeed, this maximum number of bands was fixed to 20 considering the superspectral sensor design application, for which the number of possible spectral bands is limited. In practice, the influence of the number of selected bands on the FS score and on the classification performance (measured by Kappa and the F-score of the worst classified class) for a RBF SVM classifier using the best selected band subset was considered. The optimal number of bands was chosen as the one from which these scores virtually no longer increase. Results obtained using several FS scores were also considered to make this decision, and at the end, the number of bands to select is a trade-off between several FS criteria.

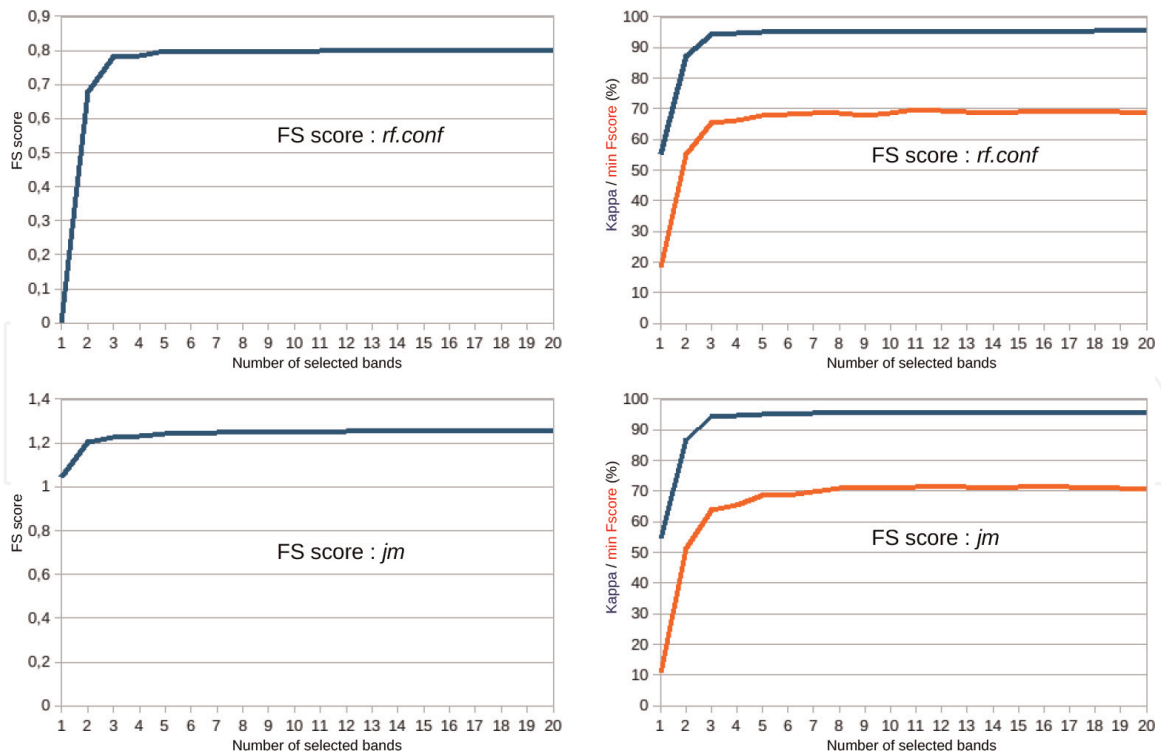
For Pavia data set, the influence of the number of selected bands on the FS score and on the classification performance (measured by Kappa and the F-score of the worst classified class) for a RBF SVM classifier using the best selected band subset can be seen in **Figure 4**. The different quality indices no longer evolve a lot from five bands, except the minimal F-score increasing slightly up to seven bands. Similar results were obtained using several FS criteria, even though some differences exist. For instance, the quality indices increased slower for *jm* than for *rf.conf* in **Figure 4**. Thus seven bands were selected for Pavia data set for further experiments.

---

<sup>2</sup> Indian Pines data set is provided by Purdue University and available at <https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html>.

<sup>3</sup> Salinas data set was downloaded from [http://www.ehu.es/ccwintco/index.php?title=Hyperspectral\\_Remote\\_Sensing\\_Scenes](http://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes).





**Figure 4.** Pavia test site: influence of the number of selected bands on the feature selection score (left) and on classification performance (using the best band subset with a RBF SVM classifier) (right with kappa coefficient for the blue line and F-score of the worst classified class for the red line). Two FS criteria tested: *rf.conf* (top) and *jm* (bottom).

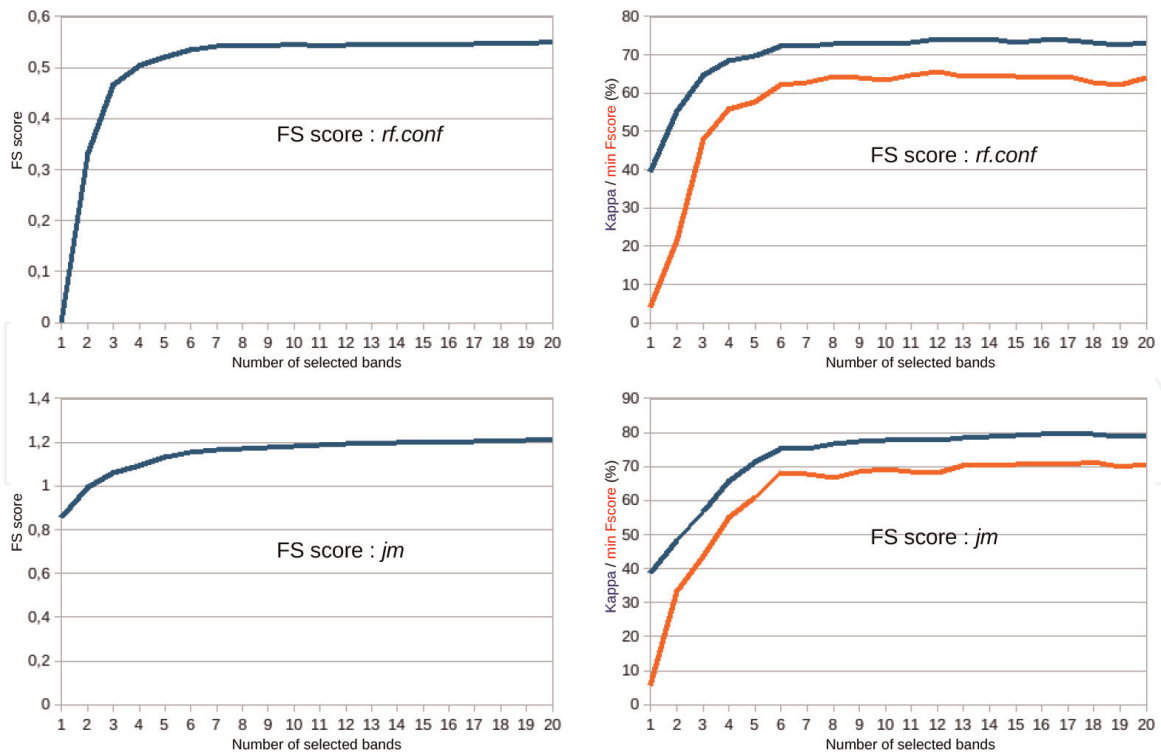
The same kind of results was obtained for Salinas, and seven bands were also selected for this data set in further experiments.

For Indian Pines, obtained results are slightly different as shown in **Figure 5**. The FS score increases fastly until seven bands are selected. Then, it remains quite constant for *rf.conf* but continues to very slightly increase for *jm*. The same phenomenon can be observed for classification accuracies reached by a RBF SVM classifier using the selected band subsets. For *rf.conf* FS criterion, a maximum is reached around 10–11 selected bands, while for *jm*, a stage is reached for these values followed by a new slight increase.

However, it must be kept in mind that this data set is more difficult than the other ones. Indeed, on the one hand, it offers less training/testing samples (and thus an increased risk of over-fitting). On the other hand, classes are more difficult to distinguish to each other, and raw classification results (that is to say without any regularization post-processing step) remain noisy. Thus 10 bands were selected in further experiments for Indian Pines data set.

### 3.4.2 Comparison of FS criteria

GA optimization heuristic was then launched to select 7 bands for Pavia, 10 bands for Indian Pines and 7 bands for Salinas. For each FS score, several feature subset solutions were proposed using GA. Their classification quality rate (Kappa) (averaged over all of them) using several classifiers is presented in **Figure 6**. At the first glance, most of the time, Kappa coefficients reached using features selected according to different FS scores are correlated over the different classifiers (RBF SVM, RF and ML) used for evaluation. Indeed, if a FS score leads to the best classification for a classifier, it will also generally be the best for the other classifiers. Thus the relevance of score appeared to be quite independent from the classifier used at validation step.



**Figure 5.** Indian Pines test site: Influence of the number of selected bands on the feature selection score (left) and on classification performance (using the best band subset with a RBF SVM classifier) (right with kappa coefficient for the blue line and F-score of the worst classified class for the red line). Two FS criteria tested: *rf.conf* (top) and *jm* (bottom).

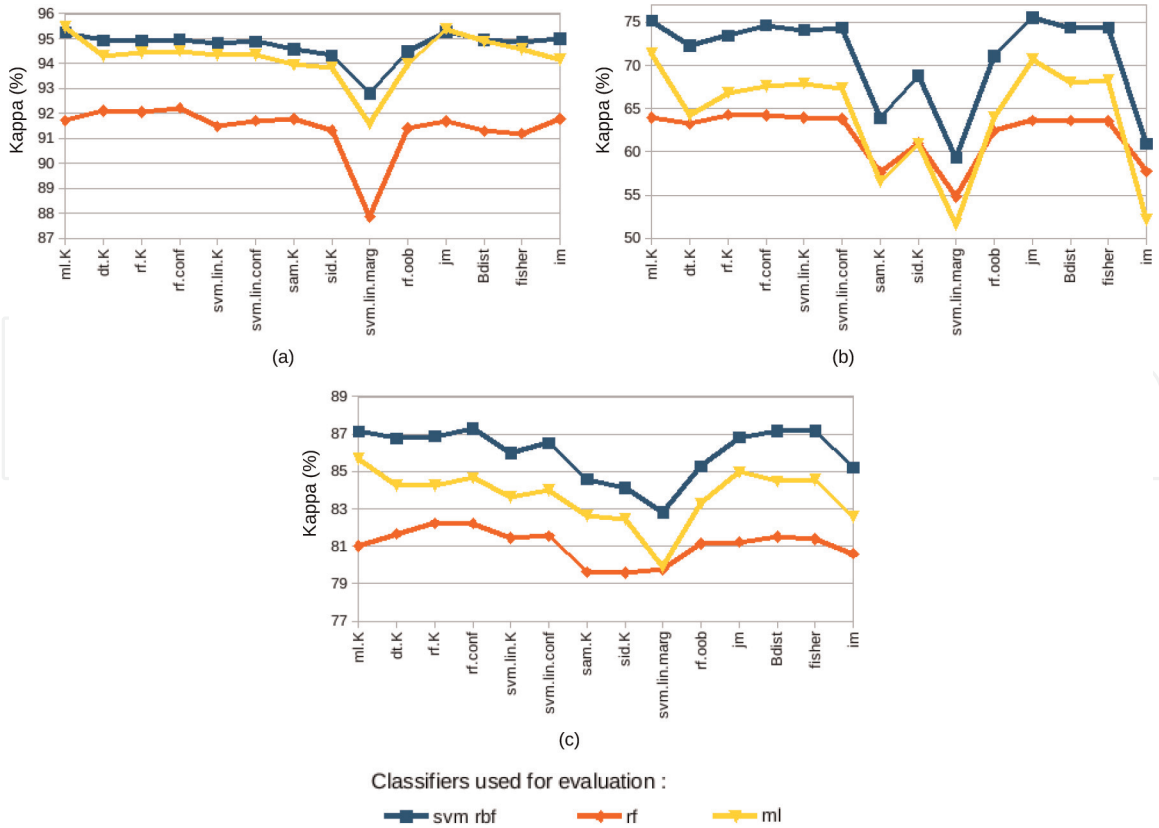
It can also be noticed from **Figure 6** that the best FS scores lead to quite equivalent classification quality. This is clearly visible for Pavia and to a less extent for Salinas. On the opposite, results are more contrasted on Indian Pines. This might be due to the fact that Indian Pines is a more difficult data set, with a stronger intra-class variability and inter-class similarity, whereas Pavia is a quite simple data set with few well-distinguished classes. These results will now be discussed for each category of FS criteria. Band importance provided by GA will also be considered.

### 3.4.2.1 Comparison of wrapper criteria

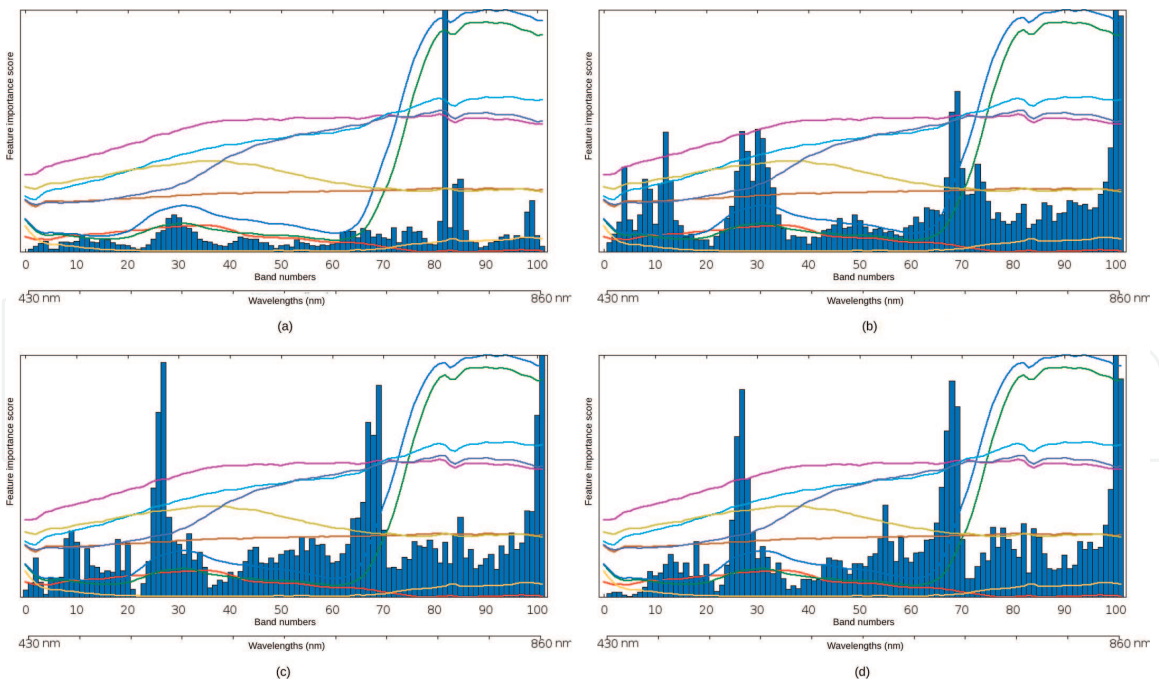
It can be seen from **Figure 6** that the FS scores *sam.K* and *sid.K* are less good than the other wrapper scores. This phenomenon appears strongly for Indian Pines and Salinas and is also a light trend for Pavia. The fact that it is more striking on Indian Pines scene can be related to the important intra-class variability of this data set.

The other wrapper scores relying on Kappa coefficient as a measure of classification performance lead to quite equivalent quantitative results. However, band importance profiles (**Figures 7 and 8**) provide other additional information. For instance, for Pavia data set (**Figure 7**), the FS score *svm.lin.K* tends to select the first bands (around band 5) of the spectrum, even though these bands are quite noisy. *ml.K* score performs very well considering classification performance but tends to be very sensitive to a probable atmospheric artefact, paying a lot of importance to bands from band 80 to band 85 and especially to band 82. This part of the spectrum corresponds to an atmospheric correction artefact, and not to a true discriminant phenomenon. This trend to select bands corresponding to this artefact is also observed for other FS scores.

Using classification confidence-based FS scores instead of classic classification accuracy scores tends to improve results. This trend can be observed in **Figure 6** both for RF and SVM: using *rf.conf* instead of *rf.K* or using *svm.lin.conf* instead of



**Figure 6.** Mean kappa coefficients obtained by classifiers RBF kernel SVM (red), RF (blue) and ML (yellow) using band subsets selected using the different FS criteria for the three data sets. From (a-c): Pavia, Indian Pines and Salinas.



**Figure 7.** Pavia test site: band importance profiles obtained using several FS criteria: (a) ml.K, (b) svm.lin.K, (c) rf.K and (d) rf.conf.

*svm.conf* tends to slightly improve classification quality. Considering band importance profiles obtained for Pavia (**Figure 7**), using *rf.conf* instead of *rf.K* avoids to select the noisy bands around band five. Band importance profiles obtained using *rf.conf* also seem to be slightly more regular than using *rf.K* both for Pavia (**Figure 7**)

and Indian Pines (**Figure 8**). Thus, using a confidence-based FS score tends to regularize feature importances and thus to stabilize feature selection.

### 3.4.2.2 Comparison of wrapper and embedded criteria

Classification qualities reached using both tested embedded criteria (*svm.lin.marg* and *rf.oob*) appeared to be generally less good than using the wrapper scores associated with these two classifiers. This is especially clear for *svm.lin.marg*, which is the worst FS score, for all classifiers used at evaluation step.

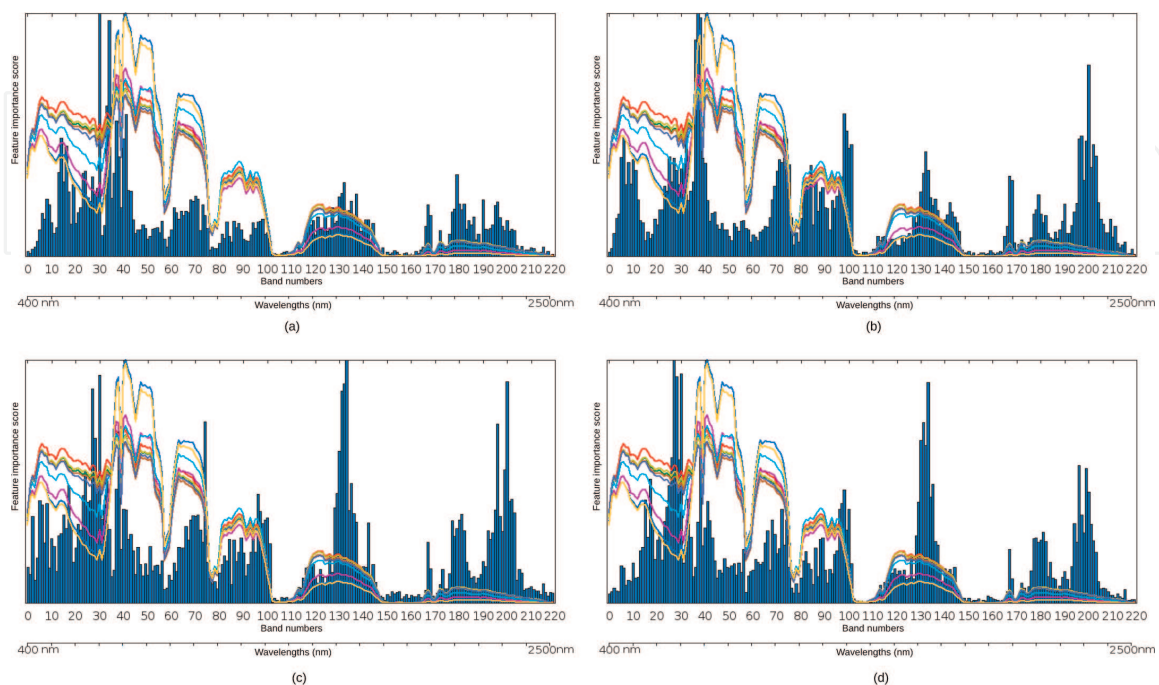
Even though it performs quite well, feature subsets selected using *rf.oob* lead generally to worse classification performance than using the best wrapper scores, and especially *rf.K* and *rf.conf*, also associated to random forests.

### 3.4.2.3 Comparison of wrapper and filter criteria

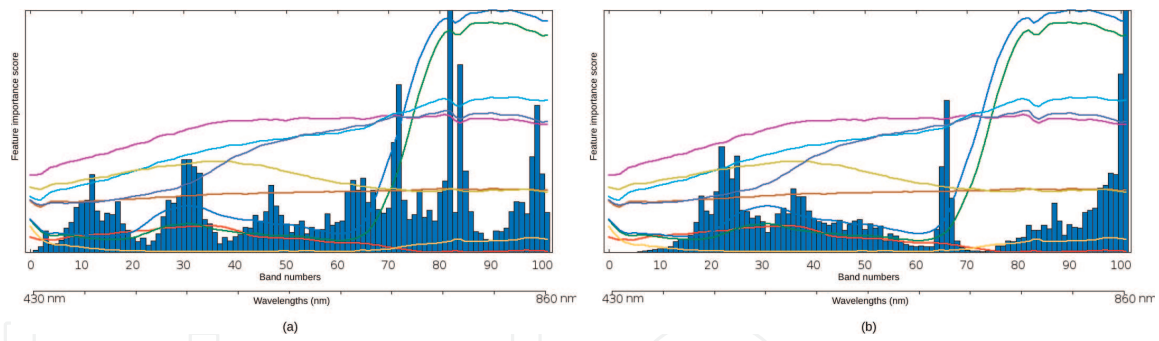
Considering classification quality (**Figure 6**), mutual information (*mi*) leads to different results for the various data sets: on Pavia data set, feature subsets selected according to this FS score enable to reach classification performance as good as the best wrapper scores, while on Indian Pines data set, obtained results are among the worst. Band importance profiles (**Figures 9 and 10**) obtained using *mi* are also very different from those obtained for the other FS scores: they tend to neglect wide parts of the spectrum. This is especially striking for Indian Pines data set, where bands from 30 to 100 are not considered as important, contrary to other FS scores.

The other tested filter FS scores are separability measures. They perform very well considering classification quality (**Figure 6**): they lead to classification results as good or better than those obtained using the best wrapper FS scores. In particular, the Jeffries-Matusita separability distance (*jm*) appears to be one of the best FS scores.

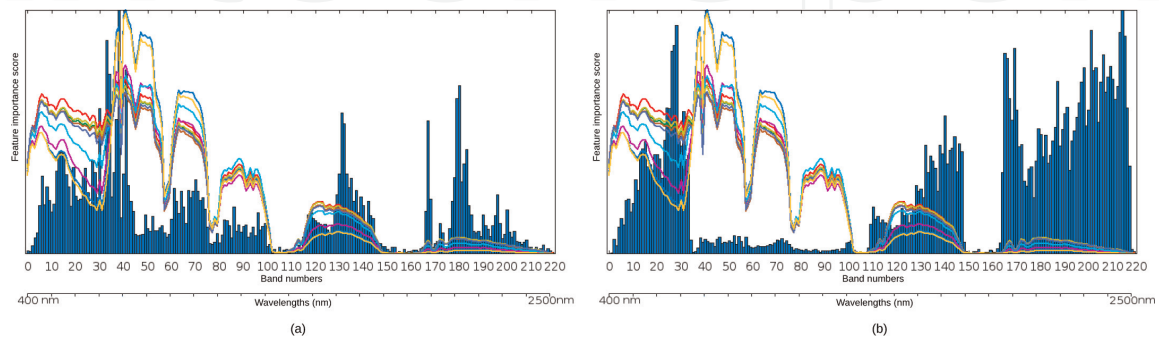
However, considering band importance profiles obtained for Pavia (**Figure 9**) using *jm*, it tends to strongly focus on a part of the spectrum (bands 80 to 85) concerned by artefacts caused by atmospheric corrections. This phenomenon also



**Figure 8.** Indian Pines test site: band importance profiles obtained using several FS criteria: (a) *ml.K*, (b) *svm.lin.K*, (c) *rf.K* and (d) *rf.conf*.



**Figure 9.** Pavia test site. Band importance profiles obtained using several FS criteria: (a) JM distance and (b) mutual information.



**Figure 10.** Indian Pines test site. Band importance profiles obtained using several FS criteria: (a) JM distance and (b) mutual information.

occurred for *bdist* and *fisher* and, as explained above, was also observed for some wrapper FS scores.

Furthermore, band importance profiles obtained using *jm* FS score seem slightly more noisy or more difficult to interpret than using the best wrapper FS scores (*rf.K*, *rf.conf*).

### 3.4.3 Conclusion

FS score comparison. Some wrapper, embedded and filter FS scores were tested and evaluated on several data sets:

- *svm.lin.marg* appears clearly as the worst of them, performing poorly on all data sets.
- Other ones (*sam.K*, *sid.K* and *im*) perform quite good on simple data sets but poorly on the most difficult one (Indian Pines).
- Most perform well, leading to good classification performance. The best FS scores are filter separability measures or wrapper FS scores. However some slight trends can be observed:
  - Filter separability scores tend to lead to slightly better classification results than wrapper scores. Especially *jm* often appears as the best FS score according to quantitative analysis. However, considering band importance profiles, it tends to lead to less regular profiles and thus to less stable solutions than some wrapper scores. Besides they appear to be sensitive to an atmospheric correction artifact for Pavia data set.

- Confidence-based wrapper scores taking into account classification confidence (*rf.conf* or *svm.lin.conf*) perform better than classic wrapper scores expressed as a simple classification “hard label” error rate. This trend could be observed both in quantitative (classification performance) and qualitative (band importance profiles) analyses. Indeed, taking into account classification confidence tends to regularize feature importances and provide more stable feature subsets.

At the end, the most interesting FS scores are *rf.conf* for wrappers and *jm* for filters, since they lead to the best quantitative results. *rf.conf* seems to provide more stable results than *jm*, considering its more regularized band importance profile. Besides it is more robust to some artefacts (e.g. atmospheric correction artefact for Pavia). However, even though computing times were not discussed in this study, it must be added that FS selection using filter separability measures (such as *jm*) is faster than using wrapper scores such as *rf.conf*.

Thematic comments. Conclusions about interesting spectrum parts can be drawn using the importance profiles provided by the different FS criteria:

- Optimized spectral configurations are different from one FS criterion to another. Indeed, some parts of the spectrum are identified as important by most FS criteria, but other ones correspond to a clear disagreement.
- Spectrum parts considered as important can often be understood considering the spectra of classes. Indeed, they can correspond to almost constant spectrum parts located before or after a strong variation of spectra of some classes. They can also correspond to intersections between the spectra of several classes.
- For Indian Pines and Salinas scenes, no precaution was taken to handle noisy bands corresponding to the main atmospheric absorption windows. However, importance measures associated with these bands were very weak for most FS criteria (except the worse of them). Such observation can be considered as an additional quality criterion for the tested FS scores.
- Band importance profiles obtained for Indian Pines are often more difficult to analyse than for Pavia. Nevertheless, some common trends could be observed, especially in the SWIR domain, where some blobs along the spectrum are visible for most FS criteria and might correspond approximately to the locations of some spectral bands of the WorldView-3 satellite.

#### **4. Exploring bandwidth and extracting optimal spectral bands using hierarchical band merging**

Works in the previous section were dedicated to the identification of a FS score. It was used for band selection, that is to say to select a subset of original bands out of a hyperspectral data set (without optimizing their weights). This section will focus on band extraction and will consider band subsets composed of spectral bands with different spectral widths. Indeed, optimizing spectral width is important to design a spectral sensor, as having wider bands is a way to limit signal noise while having too wide bands can also lead to a loss a useful information.

## **4.1 Band grouping and band extraction: state of the art and proposed strategy**

### *4.1.1 State of the art*

Band grouping and clustering. In the specific case of hyperspectral data, adjacent bands are often very correlated to each other. Thus, band selection encounters the question of the clustering of the spectral bands of a hyperspectral data set. This can be a way to limit the band selection solution space. Band clustering/grouping has sometimes been performed in association with individual band selection. For instance, [15] first groups adjacent bands according to conditional mutual information and then performs band selection with the constraint that only one band can be selected per cluster. Su et al. [40] performs band clustering applying k-means to band correlation matrix and then iteratively removes the too inhomogeneous clusters and the bands too different from the representative of the cluster to which they belong. Martínez-Usó et al. [38] first clusters ‘correlated’ features and then selects the most representative feature of each group, according to mutual information. Chang et al. [61] performs band clustering using a more global criterion taking specifically into account the existence of several classes: simulated annealing is used to maximise a cost function defined as the sum, over all clusters and over all classes, of the sum of correlation coefficients between bands belonging to a same cluster. Bigdeli et al. and Prasad et al. [68, 69] perform band clustering, but not for band extraction: a multiple SVM classifier is defined, training one SVM classifier per cluster. Bigdeli et al. [68] has compared several band clustering/grouping methods, including k-means applied to the correlation matrix or an approach considering the local minima of mutual information between adjacent bands as cluster borders. Prasad and Bruce [69] proposes another band grouping strategy, starting from the first band of the spectrum and progressively growing it with adjacent bands until a stopping condition based on mutual information is reached.

Band extraction. Specific band grouping approaches have been proposed for spectral optimization. De Backer et al. [27] defines spectral bands by Gaussian windows along the spectrum and proposes a band extraction optimizing score based on a separability criterion (Bhattacharyya error bound) thanks to a simulated annealing. [70] merges bands according to a criteria based on mutual information. Jensen and Solberg [71] merges adjacent bands decomposing some reference spectra of several classes into piece-wise constant functions. Wiersma and Landgrebe [34] defines optimal band subsets using an analytical model considering spectra reconstruction errors. Serpico and Moser [50] proposes an adaptation of his steepest ascent algorithm to band extraction, also optimizing a JM separability measure. Minet et al. [55] applies genetic algorithms to define the most appropriate spectral bands for target detection. Last, some studies have also studied the impact of spectral resolution [72], without selecting an optimal band subset.

### *4.1.2 Proposed approach*

The approach proposed in this study consists in first building a hierarchy of groups of adjacent bands. Then, band selection is performed at the different levels of this hierarchy.

Thus, it is here intended to use the hierarchy of groups of adjacent bands as a constraint for band extraction and a way to limit the number of possible combinations, contrary to some existing band extraction approaches such as [50] that extract optimal bands according to JM information using an adapted optimization method or [55] that directly use a genetic algorithm to optimize a wrapper score.

## 4.2 Hierarchical band merging

The first step of the proposed approach consists in building a hierarchy of groups of adjacent bands that are then merged. Even though it is here intended to be used to select an optimal band subset, this hierarchy of merged bands can also be a way to explore several band configurations with varying spectral resolution, that is to say with contiguous bands with different bandwidth.

### 4.2.1 Proposed algorithm

Notations. Let  $B = \{\lambda_i\}_{0 \leq i \leq n_{bands}}$  be the (ordered) set of original bands. Let  $H = \{H^{(i)}\}_{0 \leq i < n_{levels}}$  be the hierarchy of merged bands.  $H^{(i)} = \{H_j^{(i)}\}_{1 \leq j \leq n_i}$  is the  $i$ th level of this hierarchy of merged bands. It is composed of  $n_i$  merged bands, that is to say  $n_i$  ordered groups of adjacent bands from  $B$ .

Thus, each  $H_j^{(i)}$  is defined as a spectral domain:

$$H_j^{(i)} = [H_j^{(i)} \cdot \lambda_{min}; H_j^{(i)} \cdot \lambda_{max}]$$

Thus, the merged band  $B_1 \oplus B_2$  obtained when merging two such adjacent merged bands  $B_1$  and  $B_2$  is  $B_1 \oplus B_2 = [B_1 \cdot \lambda_{min}; B_2 \cdot \lambda_{max}]$ .

Let  $J(\cdot)$  be the score that has to be optimized during the band merging process.

The proposed hierarchical band merging approach is a bottom-up one. The algorithm is defined below:

**Initialization:**  $H^{(0)} = B$  (that is to say that each merged band of the first level of the hierarchy only contains one individual original band).

**Band merging: create level  $l + 1$  from level  $l$ :**

Find the pair of adjacent bands at level  $l$  that will optimize the score if they are merged: find  $\hat{k} = \underset{k}{\operatorname{argmin}} J(T(H^{(l)}, k))$  with

$$T(H^{(l)}, k) = [H_0^{(l)}; \dots; H_{k-1}^{(l)}; H_k^{(l)} \oplus H_{k+1}^{(l)}; H_{k+2}^{(l)}; \dots; H_{n_l}^{(l)}].$$

$$\text{Then } H^{(l+1)} = T(H^{(l)}, \hat{k}).$$

A table  $L_l^{l+1}$  is defined to link the different merged bands at consecutive hierarchy levels:

$$\text{for } 1 \leq j \leq \hat{k}, L_l^{l+1}(H_j^{(l)}) = H_j^{(l+1)}$$

$$L_l^{l+1}(H_{\hat{k}}^{(l)}) = H_{\hat{k}}^{(l+1)}$$

$$L_l^{l+1}(H_{\hat{k}+1}^{(l)}) = H_{\hat{k}+1}^{(l+1)}$$

$$\text{for } \hat{k} + 2 \leq j \leq n_l, L_l^{l+1}(H_j^{(l)}) = H_{j-1}^{(l+1)}.$$

At the end, the value of a pixel in a merged band is defined as the mean of its values over the different bands it contains.

### 4.2.2 Band merging criteria

Several optimization scores  $J$  were examined. (In the algorithm described in Section 4.2.1, this score is aimed to be minimized.) They can be either



supervised or unsupervised, depending whether classes are considered or not at this step.

#### 4.2.2.1 Correlation between bands

Between band correlation (either the classic normalized correlation coefficient or mutual information) (see **Figure 11**) measures the dependence between bands. So a first band merging criterion intends to merge adjacent bands considering how they are correlated to each other. Thus, it tries to obtain consistent groups of adjacent correlated bands.

Such measure inspired from [61] can be defined by the next function in equation 9 (intended to be minimized):

$$J(H^{(l)}) = \sum_{i=1}^{n_l} \sum_{b_1=H_i^{(l)}.\lambda_{min}}^{H_i^{(l)}.\lambda_{max}} \sum_{b_2=H_i^{(l)}.\lambda_{min}}^{H_i^{(l)}.\lambda_{max}} (1 - c(b_1, b_2)) \quad (9)$$

where  $c(b_1, b_2)$  is the correlation score between bands  $b_1$  and  $b_2$ .

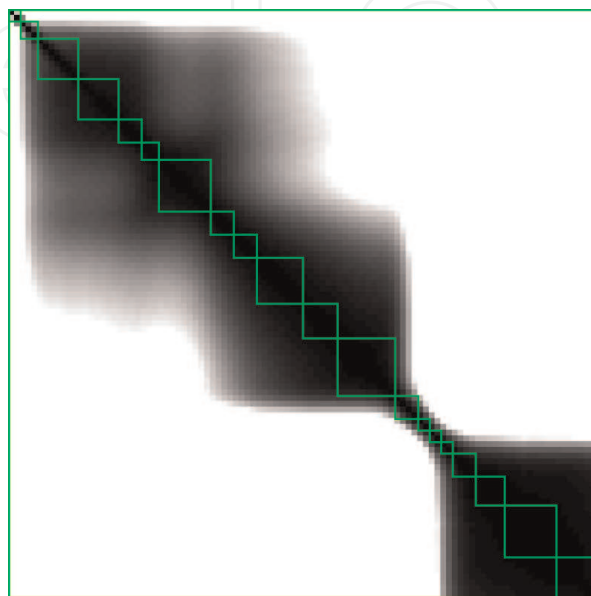
#### 4.2.2.2 Spectra approximation error

Band merging can also use the method as described in [71] to decompose some reference spectra of several classes into piece-wise constant functions (**Figure 12**). Adjacent bands are then merged trying to minimize the reconstruction error between the original and the piece-wise constant reconstructed spectra.

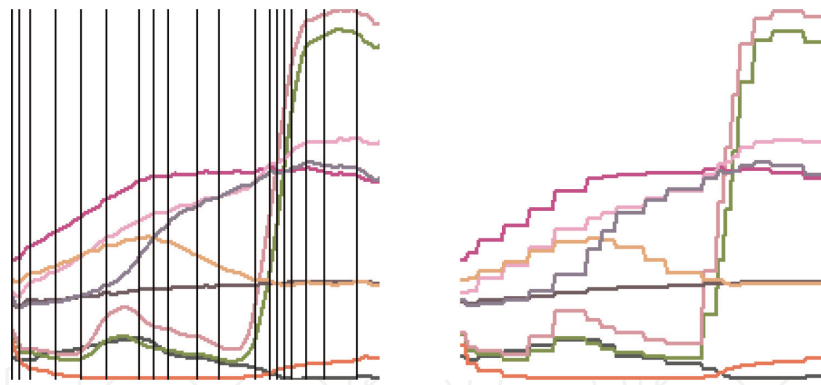
Such measure is defined by the next function (see equation 10) for a set  $s_{j1 \leq j \leq n_s}$  of  $n_s$  spectra:

$$J(H^{(l)}) = \sum_{j=1}^{n_s} \sum_{i=1}^{n_l} \sum_{b=H_i^{(l)}.\lambda_{min}}^{H_i^{(l)}.\lambda_{max}} |s_j(b) - \text{mean}(s_j, H_i^{(l)})| \quad (10)$$

where  $\text{mean}(s_j, H_i^{(l)})$  denotes the mean of spectra  $s_j$  over spectral domain  $H_i^{(l)}$ .



**Figure 11.** Examples of groups of bands superimposed on the band correlation matrix (for Pavia data set).



**Figure 12.** On the left, examples of merged bands superimposed on the original reference spectra. On the right, piece-wise constant reconstructed spectra for these merged bands (Pavia data set).

#### 4.2.2.3 Separability

Another criterion to merge adjacent band is their contribution to separability between classes. Possible separability measures are the Bhattacharyya distance (*B-distance*) or the Jeffries-Matusita distance [25, 50] already used as FS score in 3.

At a level of the band merging hierarchy, the best set of merged bands is the one that maximizes class separability. So a possible criterion  $J$  (to minimize) for band merging can be defined by equation 11 as

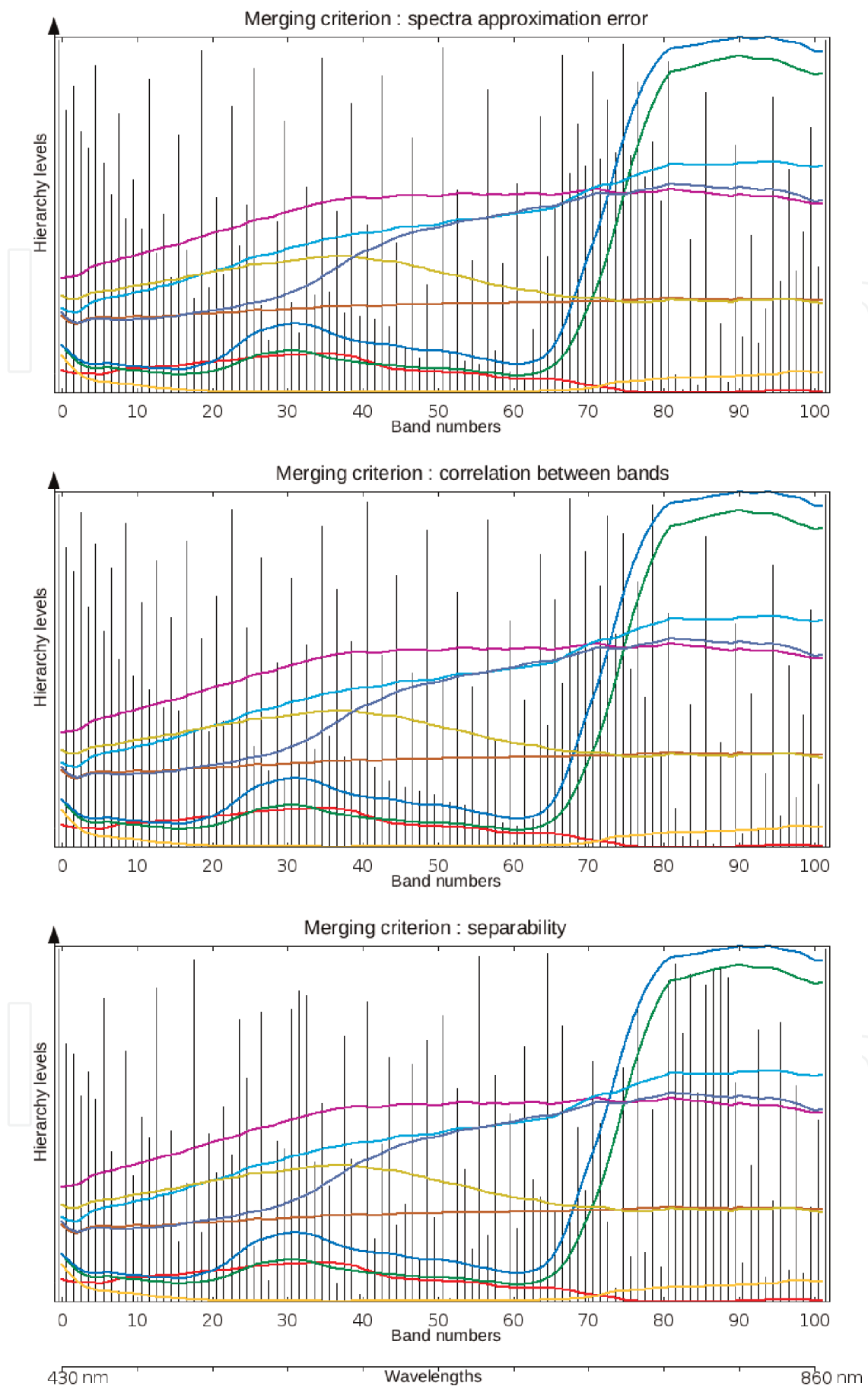
$$J(H^{(l)}) = -JM(H^{(l)}) \quad (11)$$

#### 4.2.3 Results

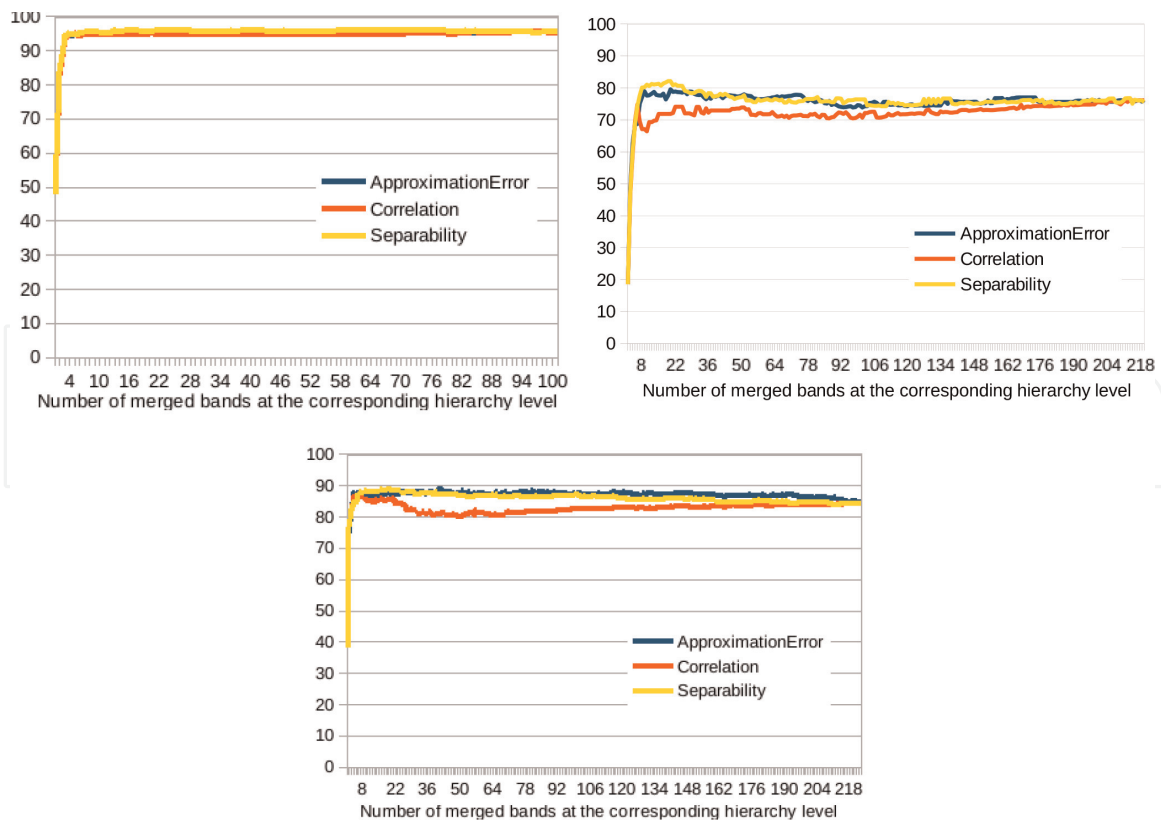
**Figure 13** shows results on Pavia data set for the three criteria described in the previous section. The separability-based criterion tends to lead to more different results than the other ones. The different criteria do not consider the same parts of the spectrum as having to be kept at fine resolution. For instance, correlation or spectra reconstruction criteria tend to fast merge bands between number 30 and 32, while separability tends to preserve them at fine resolution. On the opposite, separability tends to fast merge some bands in the red-edge domain, while the other criteria keep this domain at fine resolution. This can be understood considering the underlying criteria; indeed adjacent bands are not very correlated to each other in this domain, and the slope of spectra is strong for vegetation classes; thus they cannot be merged easily according to correlation or spectra approximation error band merging criteria. On the opposite, the only interesting information for classification (e.g. for class separability) is the fact there is a slope there and thus the values of the bands before and after this domain. Thus, merging these red-edge bands will have a little impact on class separability.

As the hierarchy of merged bands can also be a way to explore several band configurations with varying contiguous bands with different spectral resolution, the different band configurations corresponding to the different levels were evaluated using a classification quality measure. Thus, for each level, a classification was performed using a support vector machine (SVM) classifier with a radial basis function (rbf) kernel and evaluated. Its Kappa coefficient was considered.

Such results are presented on **Figure 14**. It can be seen that some spectral configurations made it possible to obtain better results than at original spectral resolution. Configurations obtained using the correlation coefficient are generally less good than for the two other criteria. Except for Pavia, the spectra piece-wise approximation



**Figure 13.** Hierarchies of merged bands obtained for different criteria for Pavia data set: Spectra piece-wise approximation error (top), between band correlation (middle) and class separability (bottom). X-axis corresponds to the band numbers/wavelengths. y-axis corresponds to the level in the band merging hierarchy (bottom, finest level with original bands; top, only a single merged band). Vertical black lines are the limits between merged bands: the lower the hierarchy, the more the merged bands are. Reference spectra of the classes are displayed in colour.



**Figure 14.** *Kappa (in %) reached by a rbf SVM for the different band configurations of the hierarchy (x-axis = number of merged bands in the spectral configuration corresponding to the hierarchy level): for Pavia (top), Indian Pines (middle) and Salinas (bottom) data sets.*

error merging criterion tends to lead to the best results. But for Pavia, the classification Kappa reached using the different criteria remained very similar.

### 4.3 Band selection within the hierarchy

#### 4.3.1 Greedy algorithm

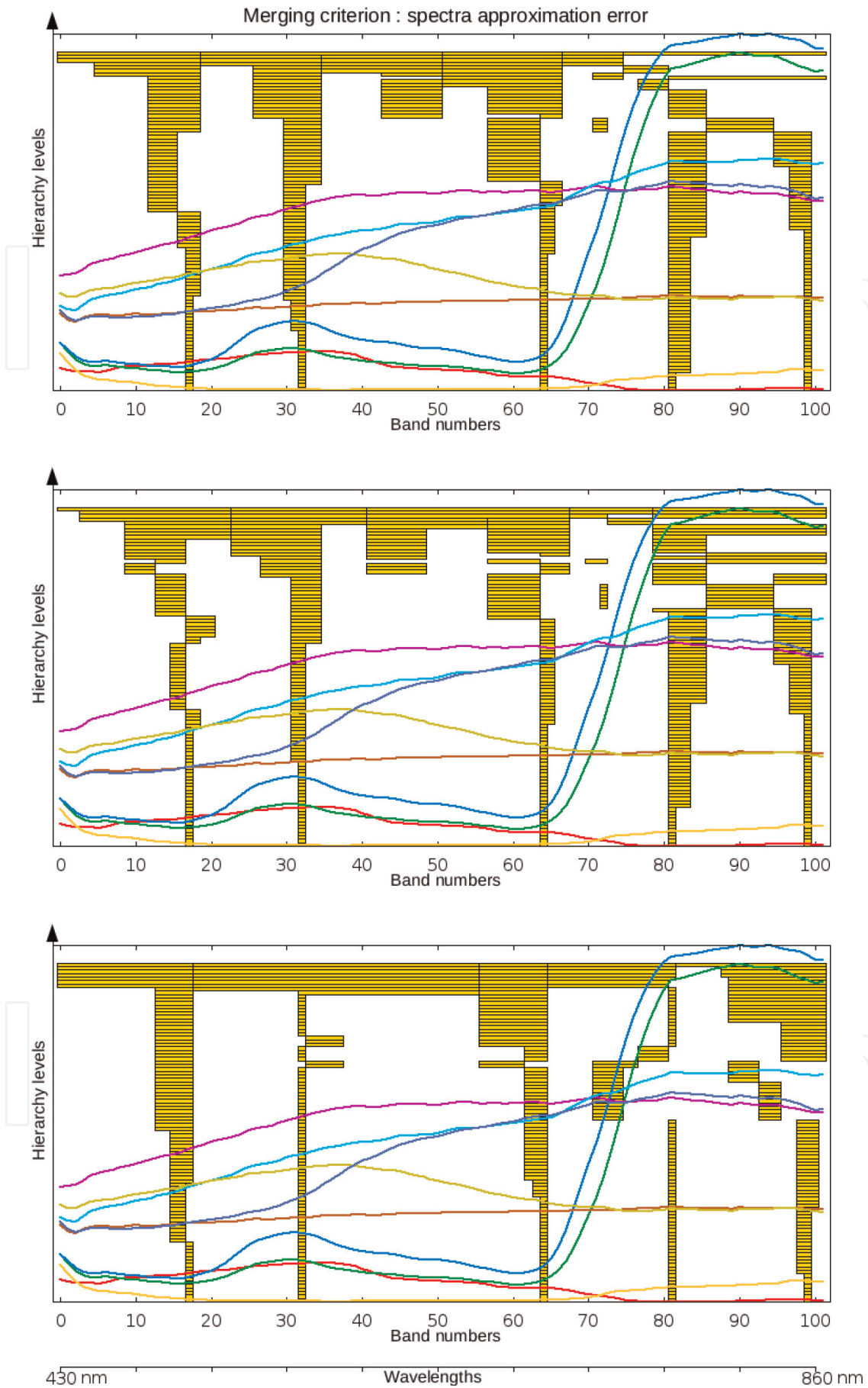
To optimize spectral configuration for a limited number of merged bands, a greedy approach was first used: it performed band selection at the different levels of the hierarchy of merged bands, paying no attention at results obtained at the previous level. Thus a set of merged bands was selected at each level of the hierarchy.

The feature selection (FS) score to optimize was the JM separability measure. It was optimized at each level of the hierarchy using an SFFS incremental optimization heuristic [28].

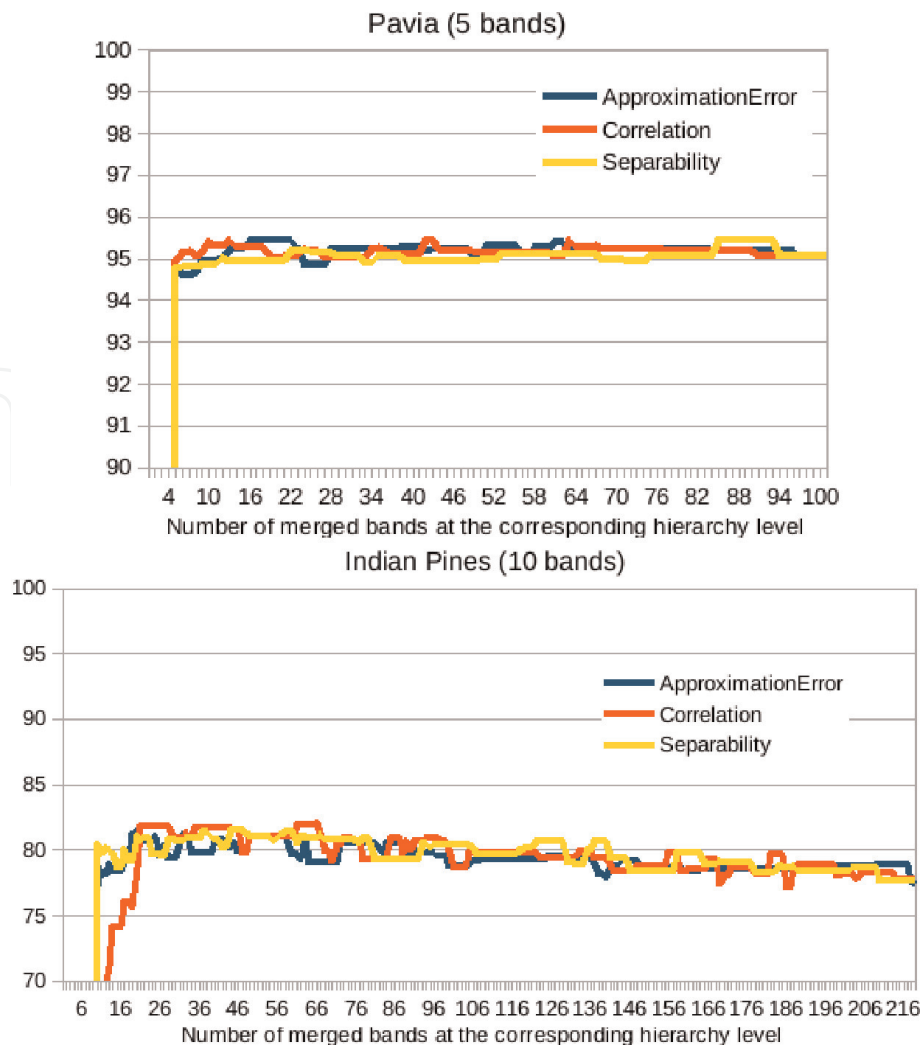
##### 4.3.1.1 Results

Obtained results on Pavia data set are presented on **Figure 15**: five merged bands (as in [67]) were selected at each level of the hierarchy of merged bands. The positions of the selected merged bands do not change a lot when climbing the hierarchy, except when reaching the lowest spectral resolution configurations. At some levels of the hierarchy, the position of some selected merged bands can also move and then come back to its initial position when climbing the hierarchy.

Thus, it can be possible to use the selected bands at a level  $l$  to initialize the algorithm at the next level  $l + 1$ . This modified method will be presented in Section 4.3.2.



**Figure 15.** Pavia data set: selected bands at the different levels of the hierarchy using the greedy approach for hierarchies of merged bands obtained using different band merging criteria: spectra piece-wise approximation error (top), between band correlation (middle) and class separability (bottom); x-axis corresponds to the band numbers/wavelengths; y-axis corresponds to the level in the band merging hierarchy (bottom—finest level with original bands; and top—only a single merged band).



**Figure 16.** Kappa (in %) reached for rbf SVM classification for merged band subsets selected at the different levels of the hierarchy for Pavia and Indian Pines data sets using the greedy FS algorithm (x-axis = number of merged bands in the spectral configuration corresponding to the hierarchy level).

The merged band subsets selected at the different levels of the hierarchy were evaluated according to a classification quality measure. As in the previous section, the Kappa coefficient reached by a rbf SVM was considered. Results for Pavia and Indian Pines data sets can be seen in **Figure 16**. At each level of the hierarchy, 5 bands were selected for Pavia, and 10 bands for Indian Pines. It can be seen that these accuracies remain very close to each other whatever the band merging criterion used, and no band merging criterion tends to really be better than the other ones. Results obtained using merged bands are generally better than using the original bands.

#### 4.3.2 Taking into account the band merging hierarchy during selection

##### 4.3.2.1 Proposed algorithm

The previous merged band selection approach is greedy and computing time expensive. So an adaptation of the SFFS heuristic was proposed to directly take into account the band merging hierarchy in the band selection process. As for the hierarchical band merging algorithm, a bottom-up approach was chosen. Contrary to the greedy approach, this one uses the band subset selected at the previous lower level when performing band selection at a new level of the hierarchy of merged bands. This algorithm is described below:

Let  $S^{(l)} = \{S_i^{(l)}\}_{1 \leq i \leq p}$  be the set of selected merged bands at level  $l$  of the hierarchy. (NB: the same number  $p$  of bands is selected at each level of the hierarchy.)

**Initialization:** standard SFFS band selection algorithm is applied to the base level  $H^{(0)}$  of the hierarchy.

**Iterations over the levels of the hierarchy:**

Generate  $S^{(l+1)}$  from  $S^{(l)}$ :

$$S^{(l+1)} \leftarrow \{L_i^{l+1}(S_i^{(l)})\}_{1 \leq i \leq p}$$

Remove possible duplications from  $S^{(l+1)}$ .

**if**  $\#S^{(l+1)} < p$ ,

**find**

$$s = \operatorname{argmax}_{b \in H^{(l+1)} \setminus S^{(l+1)}} J(S^{(l+1)} \cup b)$$

$$S^{(l+1)} \leftarrow \{S^{(l+1)}; s\}$$

**endif**

Question  $S^{(l+1)}$ : find band  $s \in S^{(l+1)}$  such that  $S^{(l+1)} \setminus \{s\}$  maximizes FS score, i.e.

$$s = \operatorname{argmax}_{z \in S^{(l+1)}} J(S^{(l+1)} \setminus \{z\}).$$

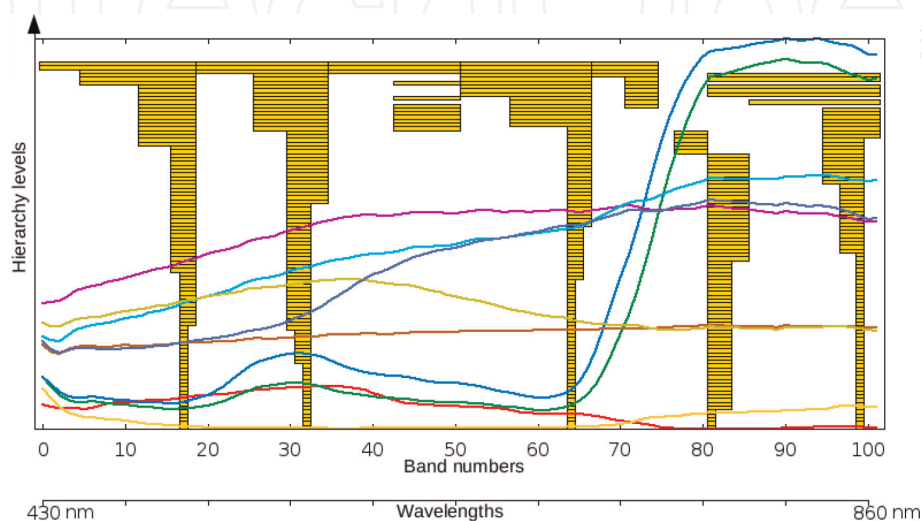
$$S^{(l+1)} \leftarrow S^{(l+1)} \setminus \{s\}$$

Then apply classic SFFS algorithm until  $\#S^{(l+1)} = p$ .

#### 4.3.2.2 Results

Obtained results on Pavia scene for the band merging criterion ‘spectra piece-wise approximation error’ are presented in **Figure 17**: five merged bands were selected at each level of the hierarchy, starting from an initial solution obtained at the bottom level of the hierarchy.

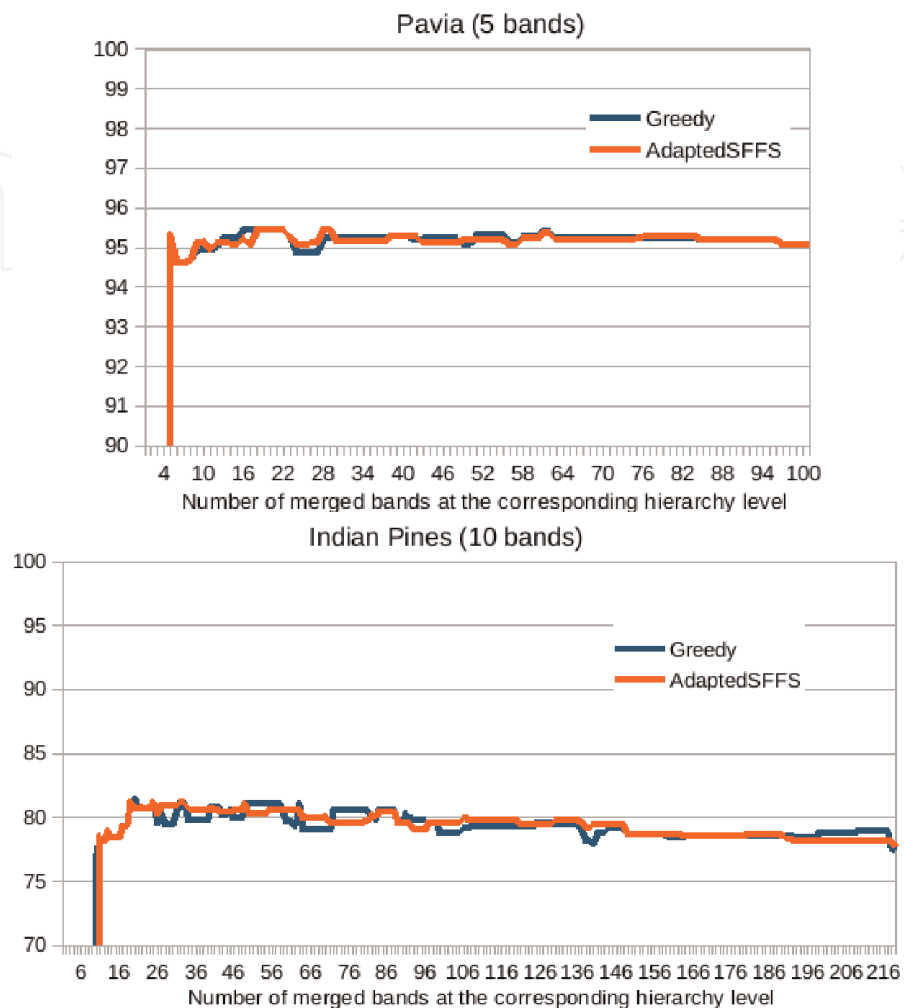
As for previous experiments, obtained results were evaluated both for Pavia (5 selected bands) and Indian Pines (10 selected bands) data sets. Kappa reached for rbf SVM classification for merged band subsets selected at the different levels of the hierarchy (built for band merging criterion ‘spectra piece-wise approximation



**Figure 17.** Pavia data set: Selected bands at the different levels of the hierarchy using the proposed hierarchy aware algorithm for a hierarchy of merged bands obtained using spectra piece-wise approximation error band merging criteria.

error’) can be seen both for the greedy FS algorithm and for the hierarchy aware one in **Figure 18**: obtained results remain very close, whatever the optimization algorithm.

Both algorithms lead to equivalent results considering classification performance (see **Table 4**), while the proposed hierarchy aware algorithm is really faster.



**Figure 18.** *Kappa (in %) reached for rbf SVM classification for merged band subsets selected at the different levels of the hierarchy (built for band merging criterion ‘spectra piece-wise approximation error’) for Pavia and Indian Pines data sets, using the hierarchy aware band selection algorithm.*

	Original bands	Greedy SFFS	Adapted SFFS
Pavia (5 bands)			
Kappa (%)	95.05	95.45	95.44
Computing times	2min	1h10min	9min
Indian Pines (10 bands)			
Kappa (%)	77.69	81.41	81.21
Computing times	4min	7h	40min

**Table 4.** *Computing times and best kappa coefficients reached on Pavia (for a 5-band subset) and Indian Pines (for a 10-band subset) data sets for band merging criterion ‘spectra piece-wise approximation error’.*

## 5. Conclusion

Hyperspectral imagery consists of hundreds of contiguous spectral bands, but only a subset of well-chosen bands is generally sufficient for a specific classification



problem. So it is possible to design superspectral sensors dedicated to specific land cover classification tasks. This chapter presented a spectral optimization strategy to identify the most relevant spectral band subset for such sensor, optimizing both band position and width. Spectral optimization involves a band subset relevance score as well as a method to optimize it.

This study first focused on the definition of this relevance score. Several filter, wrapper and embedded scores compatible with generic optimization heuristics were compared, and both their classification performance and selection stability were considered for band selection problem. At the end, most of them brought good results. Jeffries-Matusita distance score tended to lead to slightly better quantitative classification results than the best wrapper scores but also being less stable. Wrapper scores taking into account classification confidence performed better than classic wrapper scores expressed as a simple classification “hard label” error rate. For instance, a random forest confidence-based score was identified as one of the best criteria, considering both quantitative and qualitative analyses. As an intermediate result of this FS criteria comparison, a method to create band importance profiles according to the different criteria was proposed providing visual hints about the relevance of the different parts of the spectrum. Then the study focused on the optimization of bandwidth, which is important in a spectral sensor design context, as having wider bands is a way to limit signal noise while having too wide bands can also lead to a loss a useful information. A strategy consisting in building a hierarchy of groups of adjacent bands before applying band selection at the different levels of this hierarchy using an adaptation of an incremental algorithm for this problem. This band grouping strategy enabled to limit the problem’s combinatory while considering relevant band subsets composed of spectral bands with different spectral widths. It was also a way to consider several possible solutions and evaluate their impact.

To conclude, algorithms proposed in this study were applied to design a sensor dedicated to classify urban materials [20, 33].

## Author details

Arnaud Le Bris<sup>\*†</sup>, Nesrine Chehata<sup>1,2†</sup>, Xavier Briottet<sup>3†</sup> and Nicolas Paparoditis<sup>1</sup>

1 Université Paris-Est, LASTIG STRUDEL, IGN, ENSG, Saint-Mande, France


2 EA G&E Bordeaux INP, Bordeaux Montaigne University, Pessac, France

3 ONERA—The French Aerospace Lab, Toulouse, France

\*Address all correspondence to: [arnaud.le-bris@ign.fr](mailto:arnaud.le-bris@ign.fr)

† These authors contributed equally.

## IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*. 1968;**14**(1):55-63
- [2] Camps-Valls G, Bruzzone L. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2005; **43**(6):1351-1362
- [3] Melgani F, Bruzzone L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*. 2004;**42**(8):1778-1790
- [4] Pal M, Foody G. Feature selection for classification of hyperspectral data by svm. *IEEE Transactions on Geoscience and Remote Sensing*. 2010;**48**(5): 2297-2307
- [5] Demir B, Celebi A, Ertürk S. A low-complexity approach for the color display of hyperspectral remote-sensing images using one-bit-transform-based band selection. *IEEE Transactions on Geoscience and Remote Sensing*. 2009; **47**(1):97-105
- [6] Le Moan S, Mansouri A, Voisin Y, Hardeberg J. A constrained band selection method based on information measures for spectral image color visualization. *IEEE Transactions on Geoscience and Remote Sensing*. 2011; **49**(12):5104-5115
- [7] Clark RN, Swayze GA, Livo KE, Kokaly RF, Sutley SJ, Dalton JB, et al. Imaging spectroscopy: Earth and planetary remote sensing with the usgs tetracorder and expert systems. *Journal of Geophysical Research*. 2003;**108** (E12):5-1-5-44
- [8] Bassani C, Cavalli R, Cavalcante F, Cuomo V, Palombo A, Pascucci S, et al. Deterioration status of asbestos-cement roofing sheets assessed by analyzing hyperspectral data. *Remote Sensing of Environment*. 2007;**109**:361-378
- [9] Mohammadi M. Road classification and condition determination using hyperspectral imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 2012;**XXXIX-B7**
- [10] Heiden U, Segl K, Roessner S, Kaufmann H. Determination of robust spectral features for identification of urban surface materials in hyperspectral remote sensing data. *Remote Sensing of Environment*. 2007;**111**(4):537-552
- [11] Kira K, Rendell L. A practical approach to feature selection. In: *Proceedings of the 9th International Workshop on Machine Learning*. 1992. pp. 249-256
- [12] Kononenko I, Simec E, Robnik-Sikonja M. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*. 1997;**7**(1): 39-55
- [13] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2002;**46**: 289-422
- [14] Estévez PA, Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*. 2009; **20**(2):189-201
- [15] Li S, Wu H, Wan D, Zhu J. An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Systems*. 2011;**24**:40-48
- [16] Chang C-I, Du Q, Sun T-L, Althouse M. A joint band prioritization

- and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*. 1999; **37**(6):2631-2641
- [17] Du H, Qi H, Wang X, Ramanath R, Snyder W. Band selection using independent component analysis for hyperspectral image processing. In: *Proceedings of the 32nd Applied Imagery Pattern Recognition Workshop*; 2003. pp. 93-98
- [18] Chang C-I, Wang S. Constrained band selection for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*. 2006; **44**(6):1575-1585
- [19] Hasanlou M, Samadzadegan F. ICA/PCA base genetically band selection for classification of hyperspectral images. In: *Proc. of the Asian Conference on Remote Sensing (ACRS)*; 2010
- [20] Le Bris A, Chehata N, Briottet X, Paparoditis N. Spectral band selection for urban material classification using hyperspectral libraries. *ISPRS Annals*. 2016; **3**(7):33-40
- [21] Du Q, Yang H. Similarity-based unsupervised band selection for hyperspectral image analysis. *IEEE Geoscience and Remote Sensing Letters*. 2008; **5**(4):564-568
- [22] Campedel M, Maître H, Moulines E. *Indexation Des Images Satellitaires—Comparaison et évaluation des caractéristiques pour la Classification*. Télécom Paris: Tech. rep; 2004
- [23] Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*. 2003; **15**(6):1437-1447
- [24] Yang H, Zhang S, Deng K, Du P. Research into a feature selection method for hyperspectral imagery using PSO and SVM. *Journal of China University of Mining and Technology*. 2007; **17**(4): 473-478
- [25] Bruzzone L, Serpico SB. A technique for feature selection in multiclass problem. *International Journal of Remote Sensing*. 2000; **21**(3):549-563
- [26] Herold M, Gardner ME, Roberts DA. Spectral resolution requirements for mapping urban areas. *IEEE Transactions on Geoscience and Remote Sensing*. 2003; **41**(9): 1907-1919
- [27] De Backer S, Kempeneers P, Debruyne W, Scheunders P. A band selection technique for spectral classification. *IEEE Geoscience and Remote Sensing Letters*. 2005; **2**(3): 319-323
- [28] Pudil P, Novovicova J, Kittler J. Floating search methods in feature selection. *Pattern Recognition Letters*. 1994; **15**:1119-1125
- [29] Tuia D, Volpi M, Dalla Mura M, Rakotomamonjy A, Flamary R. Automatic feature learning for spatio-spectral image classification with sparse SVM. *IEEE Transactions on Geoscience and Remote Sensing*. 2014; **52**(10): 6062-6074
- [30] Guo B, Damper R, Gunn S, Nelson J. A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. *Pattern Recognition*. 2008; **41**:1653-1662
- [31] Cang S, Hongnian Y. Mutual information based input feature selection for classification problems. *Decision Support Systems*. 2012; **54**: 691-698
- [32] Serpico SB, Bruzzone L. A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*. 2001; **39**:1360-1367

- [33] Le Bris A, Chehata N, Briottet X, Paparoditis N. Hierarchically exploring the width of spectral bands for urban material classification. In: Proc. of JURSE; 2017
- [34] Wiersma D, Landgrebe D. Analytical design of multispectral sensors. *IEEE Transactions on Geoscience and Remote Sensing*. 1980; **GE-18**(2):180-189
- [35] Zhou S, Zhang J, Su B. Feature selection and classification based on ant colony algorithm for hyperspectral remote sensing images. In: Proc. of the 2nd International Congress on Image and Signal Processing (CISP'09); 2009. pp. 1-4
- [36] Fauvel M, Zullo A, Ferraty F. Nonlinear parsimonious feature selection for the classification of hyperspectral images. In: Proc. of the 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS'14); 2014
- [37] Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006;**7**(3):1-13
- [38] Martínez-Usó A, Pla F, Martínez Sotoca J, García-Sevilla P. Clustering-based hyperspectral band selection using information measures. *IEEE Transactions on Geoscience and Remote Sensing*. 2007;**45**(12):4158-4171
- [39] Zhang L, Zhong Y, Huang B, Gong J, Li P. Dimensionality reduction based on clonal selection for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*. 2007;**45**(12):4172-4186
- [40] Su H, Yang H, Du Q, Sheng Y. Semisupervised band clustering for dimensionality reduction of hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*. 2011;**8**(6): 1135-1139
- [41] Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; **58**:267-288
- [42] Tuia D, Camps-Valls G, Matasci G, Kanevski M. Learning relevant image features with multiple-kernel classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2010; **48**(10):3780-3791
- [43] Breiman L. Random forests. *Machine Learning*. 2001;**45**(1):5-32
- [44] Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees*. Boca Raton: CRC Press; 1984
- [45] Tang J, Alelyani S, Liu H. Feature selection for classification: A review. In: *Data classification: Algorithms and applications*. Chapman and Hall/CRC data mining and knowledge discovery series. Boca Raton: CRC Press; 2014. pp. 37-64
- [46] Fröhlich H, Chapelle O, Schölkopf B. Feature selection for support vector machines by means of genetic algorithms. In: Proc. of the 15th IEEE International Conference on Tools with Artificial Intelligence; 2003. pp. 142-148
- [47] Yang H, Du Q, Chen G. Particle swarm optimization-based hyperspectral dimensionality reduction for urban land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2012;**5**(2):544-554
- [48] Kandasamy S, Minghelli-Roman A, Tavin F, Mathieu S, Baret F, Gouton P. Optimal band selection for future satellite sensor dedicated to soil science. In: *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009. WHISPERS '09. First Workshop; 2009. pp. 1-4

- [49] Kandasamy S, Tavin F, Minghelli-Roman A, Mathieu S, Weidong L, Baret F. et al. Optimization of image parameters using a hyperspectral library application to soil identification and moisture estimation. In: *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009; 2009*. Vol. 3; pp. III-141-III-144
- [50] Serpico SB, Moser G. Extraction of spectral channels from hyperspectral images for classification purposes. *IEEE Transactions on Geoscience and Remote Sensing*. 2007;45(2):484-495
- [51] Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*. 1994; 5(4):537-550
- [52] Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005;27(8): 1226-1238
- [53] Sotoca J, Filiberto P. Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition*. 2010; 43(6):2068-2081
- [54] Zhuo L, Zheng J, Wang F, Li X, Bin A, Qian J. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 2008;37 (B7):397-402
- [55] Minet J, Taboury J, Pealat M, Roux N, Lonnoy J, Ferrec, Y. Adaptive band selection snapshot multispectral imaging in the vis/nir domain. In: *Proceedings of SPIE the International Society for Optical Engineering; 2010*. 7835; p. 10
- [56] Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*. 2008;8(5):392-403
- [57] Tuia D, Courty N, Flamary R. A group-lasso active set strategy for multiclass hyperspectral image classification. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 2014;II-3: 1-8
- [58] Zhu J, Rosset S, Hastie T, Tibshirani R. 1-norm support vector machines. *Advances in Neural Information Processing Systems*. 2004; 16(1):49-56
- [59] Pal M. Margin-based feature selection for hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*. 2009; 11:121-220
- [60] Zare A, Gader P. Hyperspectral band selection and endmember detection using sparsity promoting priors. *IEEE Geoscience and Remote Sensing Letters*. 2007;5(2): 256-260
- [61] Chang Y-L, Chen K-S, Huang B, Chang W-Y, Benediktsson J, Chang L. A parallel simulated annealing approach to band selection for high-dimensional remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2011; 4(3):579-590
- [62] Chang C. An information-theoretic approach to spectral variability, similarity and discrimination for hyperspectral image analysis. *IEEE Transactions on Information Theory*. 2000;46(15):1927-1932
- [63] Boser E, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: *Fifth annual workshop on computational learning theory*. 1992;5: 144-152

- [64] Fauvel M. Spectral and spatial methods for the classification of urban remote sensing data [Ph.D. thesis]. Institut National Polytechnique de Grenoble; 2007
- [65] Pal M. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*. 2005;26(1):217-222
- [66] Le Bris A, Chehata N, Briottet X, Paparoditis N. A random forest class memberships based wrapper band selection criterion: application to hyperspectral. In: Proc. of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'15); 2015
- [67] Le Bris, A, Chehata N, Briottet X, Paparoditis N. Identify important spectrum bands for classification using importances of wrapper selection applied to hyperspectral data. In: Proc. of the 2014 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM'14); 2014
- [68] Bigdeli B, Samadzadegan F, Reinartz P. Band grouping versus band clustering in svm ensemble classification of hyperspectral imagery. *Photogrammetric Engineering and Remote Sensing*. 2013;79(6):523-533
- [69] Prasad S, Bruce LM. Decision fusion with confidence-based weight assignment for hyperspectral target recognition. *IEEE Transactions on Geoscience and Remote Sensing*. 2008; 46(5):1448-1456
- [70] Cariou C, Chehdi K, Le Moan S. Bandclust: An unsupervised band reduction method for hyperspectral remote sensing. *IEEE Geoscience and Remote Sensing Letters*. 2011;8(3): 565-569
- [71] Jensen A-C, Solberg A-S. Fast hyperspectral feature reduction using piecewise constant function approximations. *IEEE Geoscience and Remote Sensing Letters*. 2007;4(4): 547-551
- [72] Adeline K, Gomez C, Gorretta N, Roger J. Sensitivity of soil property prediction obtained from VNIR/SWIR data to spectral configurations. In: Proc. of the 4th International Symposium on Recent Advances in Quantitative Remote Sensing: RAQRS'IV; 2014