Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Spectral region identification versus individual channel selection in supervised dimensionality reduction of hyperspectral image data

S. Enayat Hosseini Aria
Massimo Menenti
Ben G. H. Gorte

SPIE.

S. Enayat Hosseini Aria, Massimo Menenti, Ben G. H. Gorte, "Spectral region identification versus individual channel selection in supervised dimensionality reduction of hyperspectral image data," *J. Appl. Remote Sens.* **11**(4), 046010 (2017), doi: 10.1117/1.JRS.11.046010.

Spectral region identification versus individual channel selection in supervised dimensionality reduction of hyperspectral image data

S. Enayat Hosseini Aria,* Massimo Menenti, and Ben G. H. Gorte
Delft University of Technology, Civil Engineering and Geosciences Faculty,
Geoscience and Remote Sensing Department, Delft, The Netherlands

Abstract. Hyperspectral images may be applied to classify objects in a scene. The redundancy in hyperspectral data implies that fewer spectral features might be sufficient for discriminating the objects captured in a scene. The availability of labeled classes of several areas in a scene paves the way for a supervised dimensionality reduction, i.e., using a discrimination measure between the classes in a scene to select spectral features. We show that averaging adjacent spectral channels and using wider spectral regions yield a better class separability than the selection of individual channels from the original hyperspectral dataset. We used a method named spectral region splitting (SRS), which creates a new feature space by averaging neighboring channels. In addition to the common benefits of channel selection methods, the algorithm constructs wider spectral regions when it is useful. Using different class separability measures over various datasets resulted in a better discrimination between the classes than the best-selected channels using the same measure. The reason is that the wider spectral regions led to a reduction in intraclass distances and an improvement in class discrimination. The overall classification accuracy of two hyperspectral scenes gave an increase of about two-percent when using the spectral regions determined by applying SRS. [DOI: 10.1117/1.JRS.11.046010]

Keywords: dimensionality reduction; hyperspectral images; spectral region splitting; class separability; search algorithms; filter approach.

Paper 170504 received Jun. 8, 2017; accepted for publication Oct. 2, 2017; published online Oct. 30, 2017.

1 Introduction

Hyperspectral imagery can provide a complete spectral coverage in the 400- to 2500-nm spectral range with narrow spectral channels, causing an increase in the data dimensionality.[1–3] On the one hand, the high dimensionality of hyperspectral data has the potential to identify observed targets and patterns reliably. On the other hand, it escalates the complexity of the data that must be analyzed.[4–6] The main challenge of using hyperspectral images is reducing the cost of data inspection without degrading the potential of image classification and pattern recognition.[7–9] Dimensionality reduction (DR) techniques are applied to hyperspectral images before classification procedures to mitigate the so-called curse of the dimensionality of hyperspectral datasets.[10–12] In classification procedures, for example, the problem of hyperdimensionality can usually be observed, i.e., the classification accuracy increases gradually with the number of features or dimensions but decreases when the number of features (spectral bands) becomes large.[13,14] Sometimes, using algorithms less dependent on the number of dimensions, such as the support vector machine (SVM),[10,15,16] is a solution. Another way is to obtain a limited number of features and apply frequently used classification algorithms to achieve a high classification accuracy.[8,17–20]

*Address all correspondance to: S. Enayat Hosseini Aria, E-mail: S.E.Hosseiniaria@tudelft.nl

The main objective of the DR method used in our approach was to evaluate the potential of wider spectral bands versus the narrow spectral channels in the original hyperspectral image data to achieve a better class separability and a good classification accuracy with a limited number of features. Having a sufficient number of labeled classes based on ground truth in an image captured by a hyperspectral sensor, it is feasible to measure the discrimination between the classes in the feature space. In general, the objective is to achieve maximum separability of the labeled classes since, by maximizing class separability, the minimum-error-rate classification is expected,[21–23] which leads to more accurate and reliable classification results. It can be achieved by selecting the spectral bands that increase the class discrimination. This type of feature selection using a set of predefined classes is a supervised method.[19,20,24]

Ideally, the best bands for an image classification and class separability are those that provide larger interclass distances and smaller intraclass diameters in the selected feature space.[18,21,24] For example, in one-dimensional feature space, the interclass distance can be obtained simply by measuring the Euclidean distance between the class means, and the class variances can provide the intraclass diameter. Separability measures such as Mahalanobis (Mh) distance and Jeffreys–Matusita (JM)[23] account for both the inter- and the intraclass distance implicitly, i.e., a larger Mh distance gives a higher ratio of inter- to intraclass distances. These separability measures can be plugged in to a DR method to obtain a channel set that maximizes class discrimination.[19]

There are several supervised channel selection methods,[5,18–20,25] mostly aiming at maximizing class separability. In this paper, we apply a method called the spectral region splitting (SRS) to identify spectral regions that maximize class separability. This method allows the use of different metrics to obtain the required spectral configuration, given a hyperspectral dataset. SRS, similar to other DR techniques, mitigates the curse of dimensionality;[13] however, it may not be categorized into the standard DR methods.

Usually, DR is obtained by feature extraction (FE) and feature selection (FS) techniques with the aim of exploiting the most useful information carried by hyperspectral data.[4,8,17] FE methods transform the data into a new coordinate system to provide the required features, but the transformation may compromise some critical information associated with the physical meanings of original spectral bands.[26] On the other hand, feature (channel) selection techniques identify a subset of the original channels for a given purpose, having the advantage of preserving the original information in the dataset.[27,28]

SRS is an alternative method to channel selection techniques, which averages adjacent channels into spectral regions (bands), instead of selecting a subset of the original spectral channels. Each band covers a number of adjacent channels of the original image, and each channel is used in exactly one band. The final band configuration is also contiguous, so it covers the whole of the spectrum captured by an imaging spectrometer. At every pixel, the signal value in a band is the average of the values of the adjacent channels included in this band. The primary advantage over channel selection is that no information gets thrown away.

There are also other methods that cluster adjacent spectral channels for different purposes and substitute them with a value per pixel, e.g., the mean reflectance spectrum.[5,29–31] For example, Jensen and Solberg[29] averaged adjacent spectral channels to minimize the square representation error of spectral curve in a scene. Cariou et al.[31] used an unsupervised method to find the least-dependent spectral regions by averaging adjacent channels. The final bandset obtained by these methods including SRS is a set of contiguous spectral regions with different widths covering the whole spectrum. In addition to the higher signal-to-noise ratio of the wider bands than of the narrow channels of the original hyperspectral data, in this approach, we demonstrate the advantage of averaging adjacent channels in increasing class discrimination.

We applied different separability metrics with SRS, and we evaluated all the metrics with two hyperspectral datasets. The separability metrics used were Euclidean, Mh, divergence, Bhattacharyya, transformed divergence (TD), and JM distance.[23] We also compared SRS with three well-known search algorithms used in channel selection: branch and bound (BB),[32] sequential forward selection (SFS),[33] and sequential forward floating selection (SFFS).[34]

The selected feature sets were also evaluated regarding classification accuracy, although we obtained the features completely independent of the classification method. This type of FS process is called filter approach;[33–35] there is no classifier involved in the FS procedure, and the features selected are entirely independent of the classifier applied in the subsequent classification procedure. As mentioned, in our study, we identify features by focusing on class separability, which is only one of the factors that determines the final results of the classification. The other approaches are so-called wrapper models[33–35] that select features on the basis of classification accuracy. In general, the wrapper approaches search for a subset of features using the induction algorithm and utilize its performance, usually accuracy, as the criterion to select features. We also compared our classification results with the ones obtained with an algorithm based on the wrapper approach.

This study is organized as follows. Section 2 reviews frequently used search algorithms in channel selection and the separability metrics utilized in this study. Section 3 explains the concept of the SRS method and its advantage, and Sec. 4 describes the datasets used. Section 5 presents five different experiments and the results, including an evaluation of SRS by comparing it with the BB, SFS, and SFFS techniques and the classification accuracy achieved. A discussion is presented at the end of this section. Finally, the findings of this study are summarized in Sec. 6.

2 Review

In this section, we review search algorithms widely applied in supervised channel selection methods. These algorithms select the channels where spectral information is more relevant for assigning pixels to predefined classes in the given scene and "relevance" is measured by a specific separability metric. The selected channels can later be used in a classification procedure. The review also covers class separability metrics and provides a framework to evaluate the application of SRS in combination with a separability metric for a supervised selection of spectral features.

2.1 *Search Algorithms*

The aim of the feature (channel) selection methods is to select the best combination of n out of N variables on the basis of a predefined metric of performance. In the supervised FS for DR of hyperspectral images, finding a truly optimal combination is very challenging due to the dimensionality of data. The selection of a limited number of channels, e.g., 10, from an original hyperspectral image, with, e.g., 224 channels, requires the evaluation of a huge number of combinations, i.e., 7.15×10^{16}, which makes the evaluation of all of them impossible. This problem is known to be NP-hard.[36,37] Therefore, instead of an exhaustive search, a greedy algorithm, which solves the problem heuristically to find a near optimal solution, is usually apply.[38]

There are several search strategies to select n out of N spectral channels of an original hyperspectral dataset. Here, we briefly explain three frequently used search algorithms, which later will be compared with the SRS method.

2.1.1 *Branch and bound*

This technique was developed to select the best subset of n features from an N-feature set.[32] The algorithm avoids the exhaustive search by rejecting suboptimal subsets without direct evaluation and guarantees that the selected subset yields the globally best value of any monotonic metric. However, the BB algorithm is applicable for small datasets, and, if the number of features in the original dataset is high, the utilization of BB is expensive.[34] This is due to the rapid growth with the number of features of the enumeration scheme (solution tree) in the BB algorithm, leading to a dramatic increase in computational cost.[39,40]

2.1.2 *Sequential forward selection*

This method has been used in many FS approaches[4,18,20,41] and can be applied to large datasets. This method is much faster than BB.[34] SFS is an iterative process that selects a single element (such as a channel) that appears to be the best when combined with a previously selected subset of elements.[33] The principle of SFS and SRS is similar (see Sec. 3.1). The method reduces the complexity of selection significantly by progressively ranking the evaluated selections. It permits an analyst to make trade-offs between system complexity and performance. However, SFS suffers a problem, the so-called "nesting effect," i.e., a feature once selected cannot be discarded later.

2.1.3 *Sequential forward floating selection*

This method is also a sequential search method that is characterized by dynamically changing the number of features included or eliminated at each step.[34] Features selected can be later discarded in a search strategy, which avoids the nesting effect problem. There is no guarantee, however, that SFFS always provides a better subset of features than SFS.[42] SFFS is also faster than BB.[34]

2.2 *Separability Metrics*

The separability distance shows how well the given classes are separated in the feature space, which provides guidance for the actual classification of images. Swain and Davis defined the separability analysis as a performance assessment based on the training data to estimate the expected error in the classification for various feature combinations.[22]

There are several separability metrics, with the simplest ones only taking into account the interclass distances, while most consider both interclass distance and intraclass diameter. Figure 1 schematically shows two normally distributed classes $[N(\mu_a, \sigma_a^2), N(\mu_b, \sigma_b^2)]$ in one-dimensional feature space. The distance obtained from the mean values $(\mu_a - \mu_b)$ gives the interclass distance, and the variances of the classes (σ^2) indicate the intraclass diameters.

We applied six separability metrics in this study for FS. The equations of the metrics are given in this review, and more details on them can be found in Refs. 18, 22, 24, and 43. Let a and b be two given classes, $\mu = \{\mu_1, \mu_2, \ldots, \mu_n\}$ and Σ be the mean vector and covariance matrix of a class in an n-dimensional feature space, respectively, and σ_i be the variance of a class in the i'th dimension. The six separability metrics are as follows:

1. Euclidean distance

$$\mathrm{ED} = \|\mu_a - \mu_b\| = [(\mu_a - \mu_b)^{\mathrm{T}}(\mu_a - \mu_b)]^{1/2}. \tag{1}$$

2. Mh distance

$$\mathrm{Mh} = \left[(\mu_a - \mu_b)^{\mathrm{T}}\left(\frac{\Sigma_a + \Sigma_b}{2}\right)^{-1}(\mu_a - \mu_b)\right]^{1/2}. \tag{2}$$

3. Divergence distance

$$D = \frac{1}{2}tr[(\Sigma_a - \Sigma_b)(\Sigma_a^{-1} - \Sigma_b^{-1})] + \frac{1}{2}tr[(\Sigma_a^{-1} + \Sigma_b^{-1})(\mu_a - \mu_b)(\mu_a - \mu_b)^{\mathrm{T}}]. \tag{3}$$

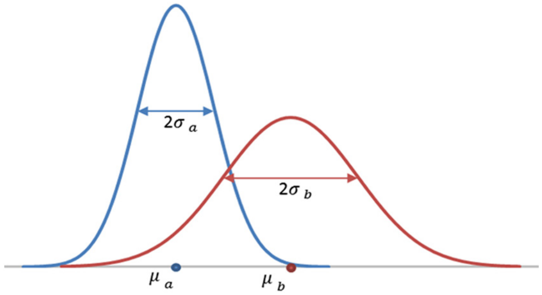


Fig. 1 Two hypothetical classes with normal distribution in one-dimensional feature space.

4. Bhattacharyya distance

$$B = \frac{1}{8}\mathrm{Mh} + \frac{1}{2}\ln\left[\frac{|(\Sigma_a + \Sigma_b)/2|}{(|\Sigma_a||\Sigma_b|)^{1/2}}\right]. \tag{4}$$

5. TD distance

$$D^t = 2[1 - e^{-D/8}]. \tag{5}$$

6. JM distance

$$\mathrm{JM} = [2(1 - e^{-B})]^{1/2}. \tag{6}$$

All of the mentioned separability metrics compute the distance between two classes. In multi-class cases, a common solution is to calculate the average distance over all class pairs[43]

$$L_{\mathrm{avg}} = \frac{\sum_{a=1}^{m-1} \sum_{b=a+1}^{m} L_{ab}}{C}, \tag{7}$$

where L_{ab} is a separability measure between the classes a and b, m is the number of classes, and C is the number of class pairs and $\sum$ is the mathematical summation indicator in this equation.

3 Proposed Method

3.1 *Spectral Region Splitting*

In the SRS method, the spectrum is split into spectral bands with different spectral widths. This method is an iterative top-down algorithm, and, therefore, a termination point must be chosen to stop the iterations. We use the term "bands" to refer to the spectral regions, either narrow or wide, obtained by the SRS algorithm and the term "channels" to refer to the narrow spectral channels in the original hyperspectral image data.

Let R be an original hyperspectral image with N_c channels and N_p pixels: $R = \{R_{ij} | 1 \leq i \leq N_c, 1 \leq j \leq N_p\}$, and R_{ij} is the i'th signal value in the j'th pixel. Assume that a reduced spectral band configuration (A) with k bands, $k \in \mathbb{Z}^+$ and $k \leq N_c$, is $A = \{A_{tj} | 1 \leq t \leq k, 1 \leq j \leq N_p\}$, and A_{tj} is the value of the t'th band in the j'th pixel. The number of splits is $k-1$, with $N_c - 1$ possible spectral locations. The set $S = \{s_1, s_2, \dots, s_{k-1}\}$ gives the split locations sorted with respect to their wavelength, so $s_1 < s_2 < \dots < s_{k-1}$. Determining the splits does not necessarily occur in the same order, e.g., the first split found can be any $s \in S$. By defining s_0 and s_k as the beginning and the end of the spectrum in S, then $S = \{s_0, s_1, \dots, s_{k-1}, s_k\}$, and A is defined by

$$A_{tj} = \begin{cases} \langle R_{ij} \rangle_{\, i \in [s_{t-1}, s_t)} & \text{if } t < k \\ \langle R_{ij} \rangle_{i \in [s_{t-1}, s_t]} & \text{if } t = k \end{cases}, \tag{8}$$

where $\langle . \rangle$ is the average of all elements in the subset.

If one band is required, there will be no split and the signal per pixel in this band will be the mean value of all spectral channels. It has no spectral detail whatsoever. If the measurements in the individual channels are noisy due to the narrow spectral bandwidth of the spectrometer, the wide band will have a much better signal to noise ratio than the individual channels. Therefore, an advantage of the broadband over the narrow channels is that at least the radiance is computed with less random noise. However, the loss of all spectral details is the disadvantage.

If $k > 1$, then the algorithm will search for the location of the first split. The location is determined based on the predefined criterion. The critical point is to translate the criterion into a score that gives a value depending on the split location for the entire spectrum. Thus, by scanning all possible locations of the split, the maximum or the minimum score determines

Table 1 SRS algorithm (pseudocode).

```
Algorithm SRS($R, T$)          // $R$ is the original dataset, and $T$ is the termination point
begin
  $S \leftarrow \{s_0, s_k\}$                // The set of split locations
  $P \leftarrow \{p_1, \ldots, p_{N_c-1}\}$          // The set of potential split locations
  $Sc = Sc[A(S)]$              // Computing the score for one broad band covering the entire spectrum
  while $T$ is not valid
    for all $p$ in $P$
      $A(S \cup p)$ is generated [Eq. (8)] and the score $Sc = Sc[A(S \cup p)]$ is computed
    end_for
    Split at the position $p_{\max}$ with the highest score ($Sc_{\max}$)
    $S \leftarrow S \cup p_{\max}$             // Adding the location of the new split to $S$
    $P \leftarrow P/p_{\max}$              // Subtracting the new split from $P$
  end_while
  return $S$
end
```

the best location of the split in that iteration. The new band configuration is generated using Eq. (8) for every location of the split and then calculates the score. Finally, the best location of the split at that iteration will be found; then the subsequent locations of splits are determined sequentially, and at each iteration, a new split is placed in the spectrum. After each iteration, the termination point must be checked. In conclusion, the algorithm yields a continuous bandset comprising several narrow and wide bands, identified by the spectral locations of the splits. The SRS procedure is described schematically in Table 1.

The search is based on the given objective and corresponding metric, which in this paper is the class discrimination. The first two spectral regions give the maximum separability on average of the classes defined in the scene. Next, the algorithm searches for the second spectral location giving the highest separability, taking into account the location of the first split. This process continues iteratively until either the average separability [Eq. (7)] achieved with the selected bands or the number of spectral regions reaches the predefined termination point. Finally, we have a set of continuous bands with different widths covering the entire spectrum.

3.2 *Spectral Region Splitting with the Class Separability Metric*

The channel selection methods operate keeping the original channel width.[18,19] Such methods search in the original hyperspectral space consisting of hundreds of narrow channels and select the channels that provide the optimal value of the given metric. If the target application is classification, this metric is usually a separability metric, which is calculated based on the known classes in the scene. In each channel, the class attributes, such as mean or variance, are constant, and the search algorithm finds the best channel combination maximizing separability.

In the SRS algorithm, in addition to the possibility of selecting narrow channels, there is also the freedom to average the narrow channels and construct wider spectral regions with new class means and variances. The algorithm can average the channels to yield a better class discrimination than the selection of individual channels. By averaging the channels, two situations may cause a better discrimination among the classes: (1) a larger distance between the class means (i.e., an increase in the interclass distance) and (2) a smaller class variance (i.e., a reduction in the intraclass distance).

Here, we prove that the intraclass distance becomes smaller in the new feature space. Let $X = \{X_1, X_2, \ldots, X_n\}$ be a class in a hyperspectral dataset having n independent spectral channels with m sample pixels per channel. The class variance is a vector given as $\sigma^2 = \{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\}$. The variance of the linear combination $Y = \sum_{i=1}^{n} a_i X_i$, where $a_1, a_2, \ldots .a_3$ are real constants ($a_i \in R$), is

$$\sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2. \tag{9}$$

In SRS, when the neighboring channels are averaged to yield a wider band, the variance of each class will be the summation of all the class variances in the channels divided by the factor n^2, where n is the number of constitutive channels of the band. In fact, if the class variances in all channels are equal to σ^2, the class variance for a band obtained by averaging those channels is σ^2/n^2, leading to a small intraclass distance. This argument can be the main point to achieve a better class separability when the SRS method is applied.

3.3 *Spectral Region Splitting Iterations*

SRS, likewise most search algorithms, detects a suboptimal feature subset based on the selected criterion, to avoid an exhaustive search for all possible combinations. The total number of subsets of size n out of N possible elements, $n < N$, is $C_n^N = \frac{N!}{(N-n)!n!}$. In the case of hyperspectral images, the number of combinations is very large, e.g., the number of possible selections of 10 out of 200 channels is 2.25×10^{16}. Whitney[33] proposed the SFS method to reduce the number of subsets evaluated and obtain a suboptimal subset. This iterative process selects a single element (like a channel) that appears to be the best when combined with a previously selected subset of elements. SRS follows the same rule as well. Since SRS searches for the location of the split to divide the spectrum, the number of possible locations is $N - 1$ at each iteration, where N is the number of channels. Therefore, the number of subsets searched to find a subset of n bands from an original dataset ($2 \le n \le N$) is given by

$$\sum_{i=1}^{n-1} (N - i) = (n - 1)\left(N - \frac{n}{2}\right). \tag{10}$$

Thus, the previous example of selecting 10 bands out of 200 channels requires evaluating 1755 band sets, i.e., 13 orders of magnitude (10^{13}) less than the number of bandsets searched by an exhaustive search strategy.

4 Dataset

The algorithm was evaluated by applying it to two hyperspectral datasets acquired by AVIRIS over a mostly vegetated area in northwestern Indiana and the Salinas Valley, California. The AVIRIS sensor covers the spectrum from 400 to 2500 nm in 224 spectral channels.[1,2] The details of the images are as follows:

Indian Pines: The scene consists of 145×145 pixels with a spatial resolution of about 20 m. Two-thirds of the Indian Pines scene is covered by agriculture, and one-third by forest and other natural perennial vegetation. The available ground truth on land cover is based on sixteen classes, but it is not mutually exclusive. Since three classes in the scene contain less than 50 samples, we did not use them for the experiments. After the atmospheric correction and the removal of noisy channels, the number of channels was reduced to 178. We removed water absorption channels (104 to 108, 150 to 163, and 220), noisy channels (1 to 4, 103, 109 to 111, 148 to 149, 164 to 166, and 217 to 219), and seven channels that are spectrally overlapping (32, 33, 95, 96, 158, 191, and 192). The Indian Pines dataset is available free of charge via Purdue University's website.[44] Figure 2 shows the scene and its reference ground truth data with 13 classes. Figure 3 shows the legend of the reference data, giving information about all the available classes

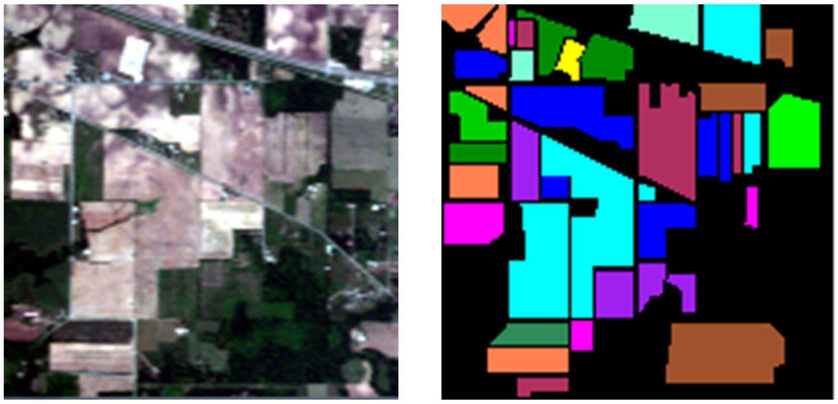

(a) (b)

Fig. 2 True color image of (a) Indian Pines scene and (b) the reference data.

#	*Class*	*Samples*	*Color*
1	Alfalfa	46	⬛
2	Corn-notill	1428	🟦
3	Corn-mintill	830	🟧
4	Corn	237	🟩
5	Grass-pasture	483	🟪
6	Grass-trees	730	🟪
7	Grass-pasture-mowed	28	⬛
8	Hay-windrowed	478	🟩
9	Oats	20	⬛
10	Soybean-notill	972	🟥
11	Soybean-mintill	2455	🟦
12	Soybean-clean	593	🟩
13	Wheat	205	🟩
14	Woods	1265	🟫
15	Buildings-Grass-Trees-Drives	386	🟩
16	Stone-Steel-Towers	93	🟨
17	Uknown samples	10776	⬛

Fig. 3 The legend of the Indian Pines reference data and their respective sample numbers.

in the scene in addition to the unclassified pixels (unknown class). The classes not included in our evaluations are shown as black, the same as the unclassified pixels.

Salinas: This scene is characterized by high spatial resolution (3.7 m). The area covered comprises 512 lines by 217 samples. The dataset is available in Ref. 45 only as at-sensor radiance with 20 water absorption channels (108 to 112, 154 to 167, and 224) discarded. We corrected the data atmospherically and removed the noisy (1 to 4, 107, 113, and 220 to 223) and duplicated channels (33, 34, 97, and 98). The final dataset has 190 channels. The ground truth on the land cover is also available; it contains 16 classes including vegetables, bare soils, and vineyard fields, which are all the classes used in our experiments. Figure 4 shows the scene and the ground truth reference data, and Fig. 5 shows the legend.

5 Experiments and Results

SRS was applied to the two hyperspectral datasets with the criterion of optimal separability between the classes in the scenes. To assess SRS, we compared it with the search algorithms

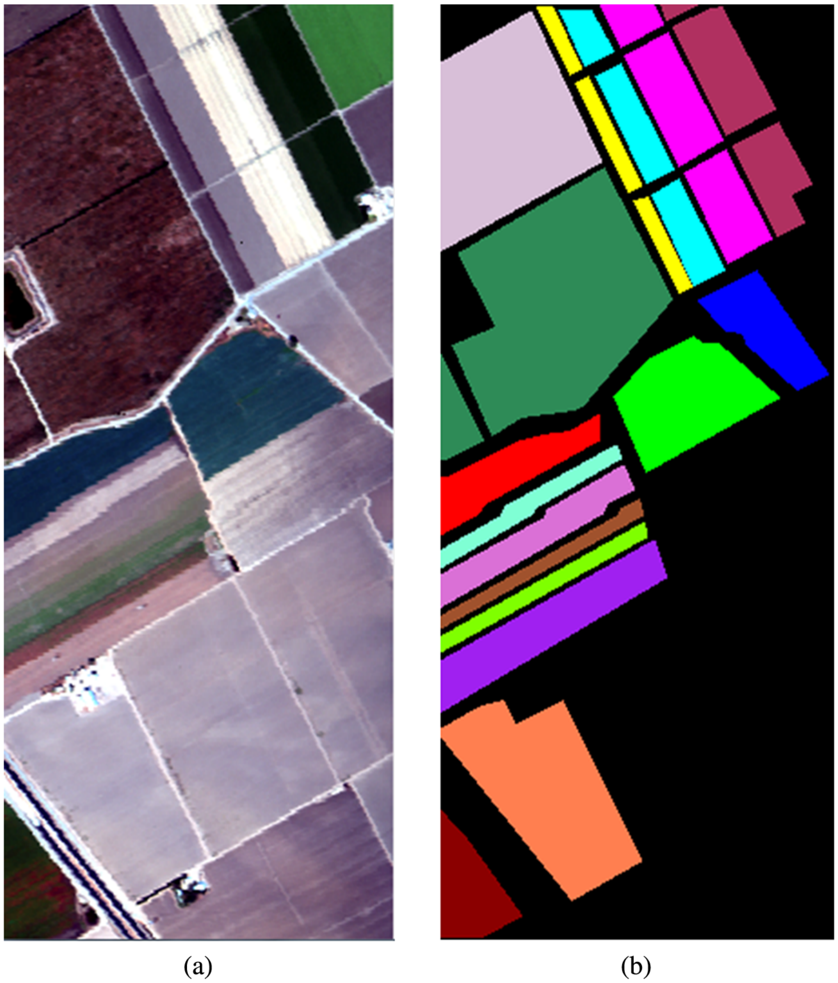


Fig. 4 True color image of the (a) Salinas scene and (b) the reference data.

used for channel selection and described in Sec. 2.1. This evaluation was performed by comparing the mean separability [Eq. (7)] and classification accuracy versus the number of features in each step.

For the experiments, we separated the available samples (pixels) for each land-cover class into training and testing data using a random subsampling method, so the training set represents the distribution of class attributes well. The training datasets were generated by taking 35%, 50%, and 70% of the total number of samples (pixels) per class and the remaining samples were considered validation data. Finally, the results obtained with the three datasets were averaged. We repeated random subsampling validation using different partitions of the samples per class and averaged the results[46] to derive a more accurate estimate of the model performance and generalize the statistical analyses.

5.1 *Spectral Region Splitting Versus Best-Selected Channels*

In the first experiment, we compared the spectral configuration constructed by SRS with the best channels giving the maximum separability selected by the BB search algorithm. As mentioned, BB selects the best n features with the highest value of the given metric out of an N-feature dataset. All other channel selection methods would choose n channels giving a lower or at best equal value of the metric than the channel set selected by BB. However, BB is costly when applied to large datasets.[34] Therefore, we used smaller datasets with 20 channels in this experiment.

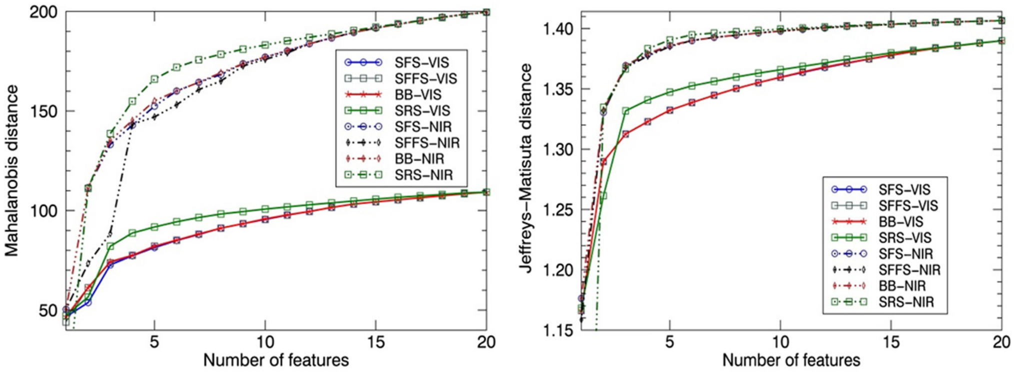
The 20-channel datasets were chosen from two spectral regions of the Indian Pines scene: (1) visible (VIS) spectral region: 400 to 637 nm and (2) near-infrared (NIR) spectral region: 647

#	*Class*	*Samples*	*Color*
1	Brocoli_green_weeds_1	2009	🟥
2	Brocoli_green_weeds_2	3726	🟩
3	Fallow	1976	🟦
4	Fallow_rough_plow	1394	🟨
5	Fallow_smooth	2678	🟦
6	Stubble	3959	🟪
7	Celery	3579	🟥
8	Grapes_untrained	11271	🟩
9	Soil_vinyard_develop	6203	🟧
10	Corn_senesced_green_weeds	3278	🟪
11	Lettuce_romaine_4wk	1068	🟩
12	Lettuce_romaine_5wk	1927	🟪
13	Lettuce_romaine_6wk	916	🟫
14	Lettuce_romaine_7wk	1070	🟩
15	Vinyard_untrained	7268	⬜
16	Vinyard_vertical_trellis	1807	🟥
17	Uknown samples	56975	⬛

Fig. 5 The legend of the Salinas reference data and their respective sample numbers.

to 822 nm. The two common separability measures, Mh and JM, were applied as separability metrics with the SRS and BB algorithms. We also applied the two other search algorithms to select channels: SFS and SFFS. These two algorithms do not select the channels having a better result than BB since they select a suboptimal channel set (see Sec. 2.1). We used them in this experiment to benchmark their performance against BB, prior to applying them to the entire scene in the second experiment.

The first experiment shows that the SRS method can provide a better class discrimination than the best selection of spectral channels (Fig. 6). When the number of bands is small (in the worst case, it was three), SRS may not give a better class separability than channel selection. After a few iterations, however, the SRS method has a rapid improvement and creates spectral regions providing better separability. This means that the spectral regions identified by SRS would discriminate the classes better than the same number of selected spectral channels by



(a) (b)

Fig. 6 The mean class separability versus the number of features obtained by applying the (a) Mh and (b) JM distances and the four algorithms described in the text; the datasets used are two spectral subsets with 20 channels of the Indian Pines scene from VIS and NIR spectral regions.

Table 2 Mean JM distance between the class pairs using the bands and channels obtained by different algorithms over the NIR dataset.

No. of features	SFS	SFFS	BB	SRS
1	1.1762	1.1585	1.1762	0.7635
2	1.3304	1.3314	1.3325	1.3347
3	1.3692	1.3676	1.3694	1.3666
4	1.3784	1.3772	1.3793	1.3833
5	1.3858	1.3850	1.3859	1.3905
6	1.3899	1.3901	1.3901	1.3948
7	1.3926	1.3926	1.3926	1.3962
8	1.3945	1.3944	1.3945	1.3974
9	1.3963	1.3963	1.3963	1.3984
10	1.3978	1.3979	1.3979	1.3995
11	1.3992	1.3994	1.3994	1.4004
12	1.4004	1.4006	1.4006	1.4013
13	1.4017	1.4017	1.4017	1.4022
14	1.4026	1.4025	1.4026	1.4030
15	1.4034	1.4034	1.4034	1.4037
16	1.4042	1.4042	1.4042	1.4044
17	1.4049	1.4049	1.4049	1.4050
18	1.4055	1.4055	1.4055	1.4056
19	1.4061	1.4061	1.4061	1.4061
20	1.4065	1.4065	1.4065	1.4065

BB. This result was obtained in all cases, regardless of whether Mh or JM was used. Table 2 also gives the value of JM on the NIR subset for more clarification.

The performance with both SFS and SFFS is almost as good as the BB selection in most cases. SFFS sometimes gave a lower separability than SFS, e.g., when Mh is used over NIR spectral regions. The result of the final selection, i.e., 20 features, was the same for all the algorithms since all available features were applied to determine separability with all the algorithms.

Overall, a spectral configuration determined by SRS identifies features that give better class separability than the other search algorithms used for channel selection in this study. The widely used search methods, SFS and SFFS, result in channel sets providing lower class separability than the channels selected by BB to maximize separability, although the differences are rather small.

5.2 *Spectral Region Splitting Versus Conventional Channel Selection Methods*

The second experiment compares the SRS with SFS and SFFS by applying them to the complete datasets. The BB algorithm is not applicable for a large dataset (see Sec. 2.1). We used the six separability measures mentioned in Sec. 2.2, and both the Indian Pines and Salinas datasets were used. Figures 7 and 8 show the general trends in mean separability versus the number of features for the two different hyperspectral scenes, with the number of features being from two to a maximum of 30.

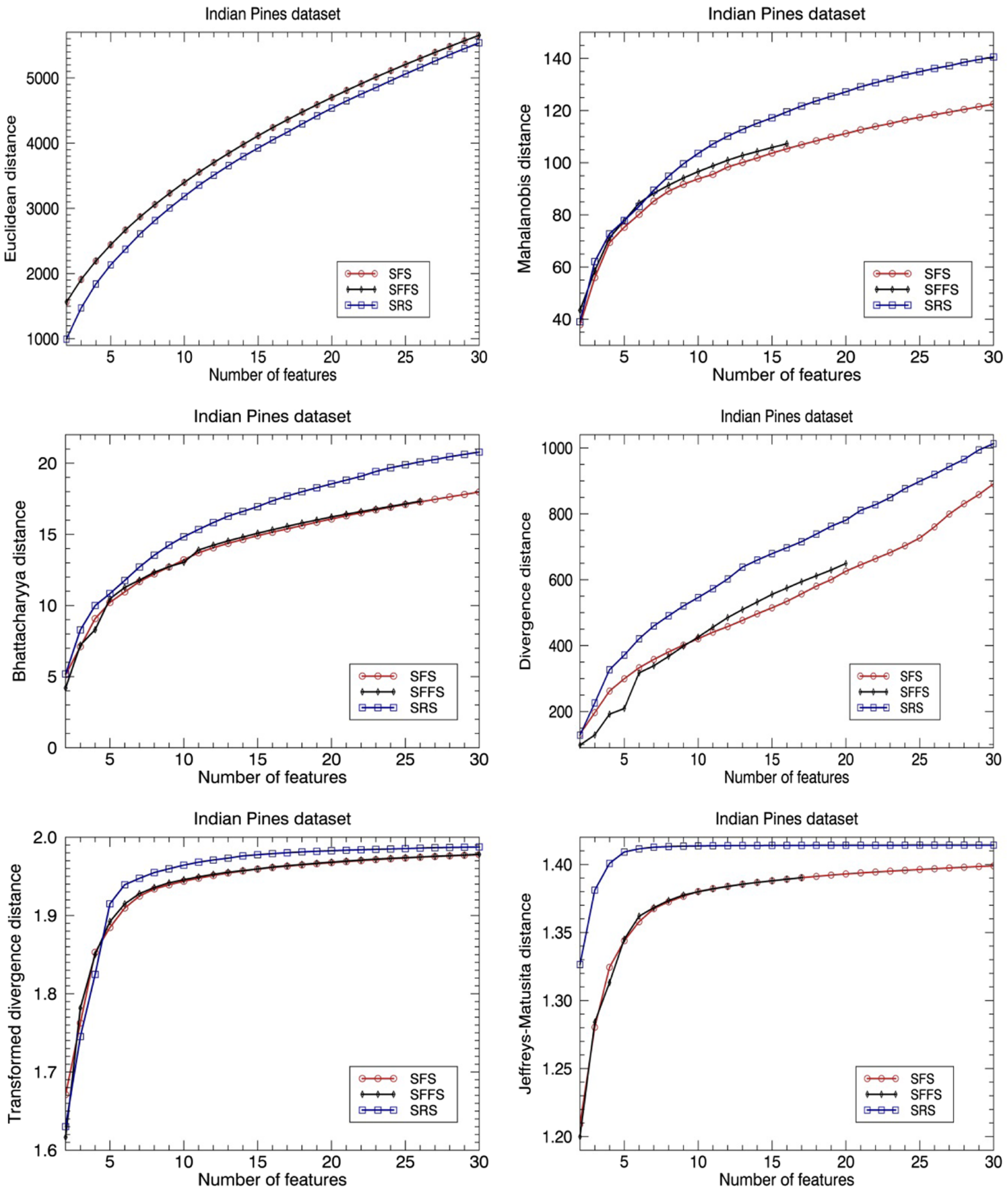


Fig. 7 The mean separability distance of the features obtained with different algorithms for the Indian Pines dataset.

Overall, SRS yields a better separability with a higher number of bands in all cases except when Euclidean distance was applied. The main difference between the Euclidean distance and the other metrics is the class variance that is not taken into account by the Euclidean distance, which only accounts for interclass distance. This means that a new feature space is generated by SRS, but, if based on the class mean only, it gives a worse discrimination. On the other hand, when other separability metrics are used, SRS achieves a better class discrimination than SFS and SFFS. These metrics consider both the class mean and the class variance, which contribute to improving the separability. Initial selections usually give better performance with individual channels than with wider spectral bands. The number of wider bands needed for higher separability than with spectral channels is different from case to case, but at some points SRS always provides a better separability. For example, even with two bands, SRS gave a better result than channel selection methods when Bhattacharyya or divergence metrics were applied with the Indian Pines dataset. In the worst case, at least seven bands were needed to achieve a better SRS performance for the Salinas scene, when the TD metric is applied.

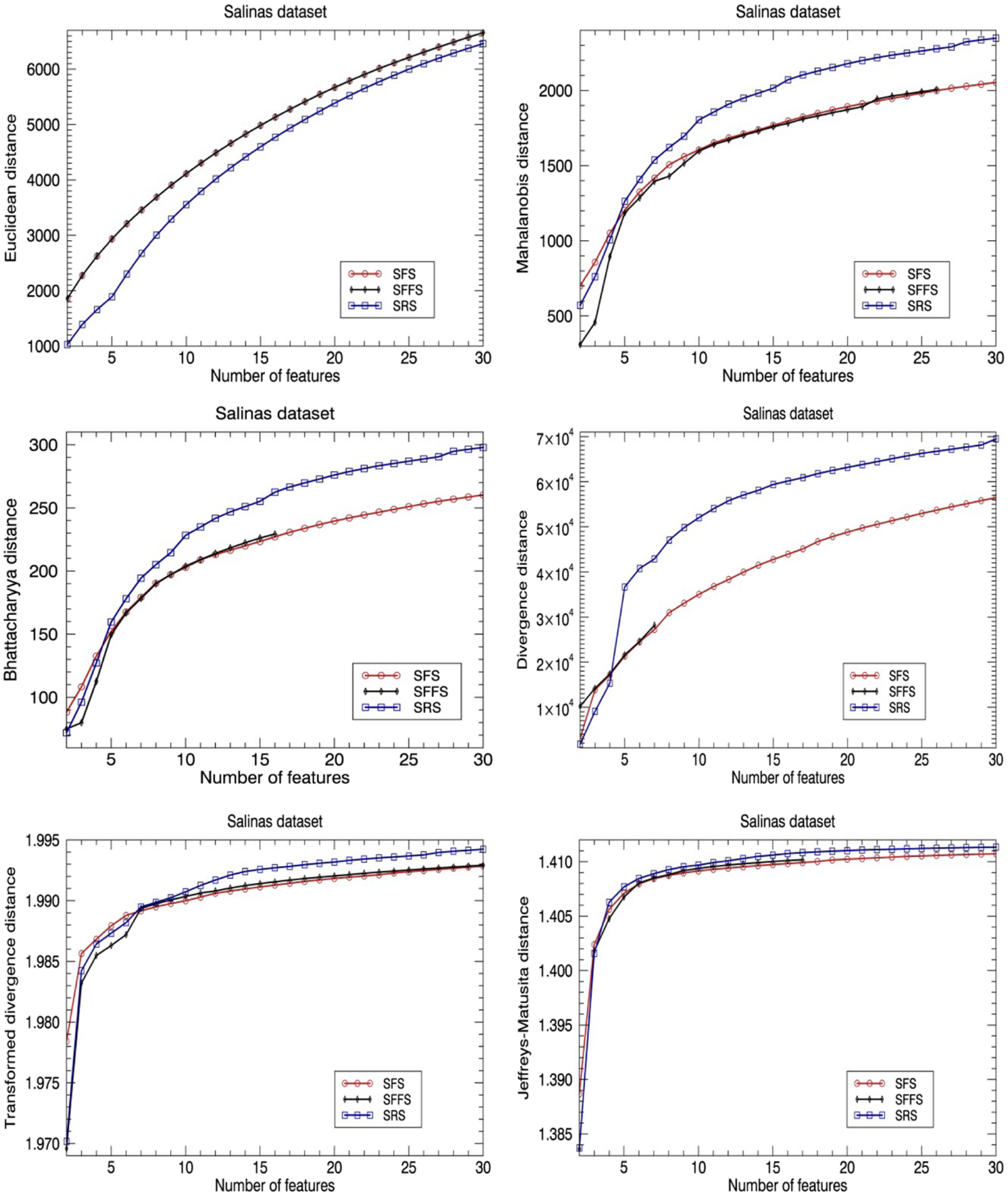


Fig. 8 The mean separability distance of the features obtained with different algorithms for the Salinas dataset.

Sometimes SFFS does not select the number of channels predefined by the termination point since SFFS remains in a local loop during the search and cannot determine the required number of features. This occurs during the "conditional exclusion" of the features already selected in the backward process of the algorithm (see Ref. 34 for the details).

We developed a hypothetical example of the steps in a run of the SFFS algorithm since, in the practical cases of this experiment, the number of steps is too large to reveal this issue. Let k be the number of features in the set S_k and $J(S_k)$ be the criterion function for the given set. Then the example is as follows (Table 3):

At step 2, the features are {1,2}, which are again obtained 10 steps later, at step 12. When the iteration continues, the same loop will be repeated, and the other features of the original dataset will not be selected. This situation is more likely to occur when the number of features is large. In these cases, we considered the maximum number of channels provided by SFFS.

Using the Euclidean distance, the results of SFS and SFFS were the same. Furthermore, SFFS does not always provide a higher separability than SFS,[39] and the results with the two methods were almost the same, whereas the SRS results were better. Table 4 also gives

Table 3 A hypothetical example of the steps in a run of the SFFS algorithm remaining in a local loop.

Step	Action	k	S_k	$J(S_k)$	$J(S_{k-1})$
1	Add feature 1	1	{1}	1	—
2	Add feature 2	2	{1,2}	3	1
3	Add feature 3	3	{1,2,3}	5	3
4	Add feature 4	4	{1,2,3,4}	7	5
5	Add feature 5	5	{1,2,3,4,5}	9	7
6	Remove feature 2	4	{1,3,4,5}	8	4
7	Add feature 6	5	{1,3,4,5,6}	10	8
8	Add feature 2	6	{1,3,4,5,6,2}	12	10
9	Remove feature 5	5	{1,3,4,6,2}	11	6
10	Remove feature 6	4	{1,3,4,2}	7	4
11	Remove feature 4	3	{1,3,2}	5	2
12	Remove feature 3	2	{1,2}	3	1

Table 4 The spectral locations (nm) of the channels and splits selected by SFS-JM and SRS-JM, respectively, for the Indian Pines dataset in the order of selection.

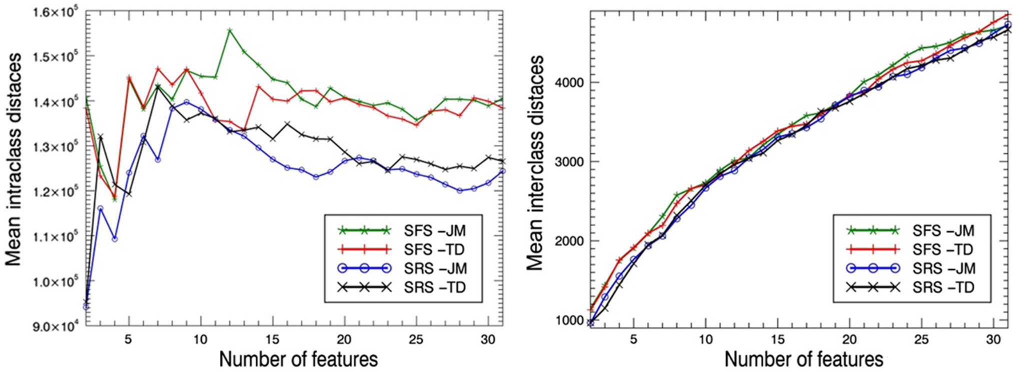
SFS-JM	677, 976, 1641, 2113, 754, 528, 697, 1983, 716, 1091, 627, 783, 1293, 1512, 1770, 587, 2262, 1139, 706, 1760, 1993, 548, 667, 2093, 2202, 725, 745, 617, 947, 1283.
SRS-JM	697, 1148, 2003, 735, 687, 1571, 841, 657, 577, 1789, 1323, 1120, 1621, 1283, 1730, 1983, 1442, 985, 899, 617, 1462, 488, 1091, 2153, 2083, 2242, 2192, 956, 677.

more details about the spectral location of the channels and the splits selected by the two algorithms using the JM metric.

5.3 *Separability with Wider Spectral Bands*

The second experiment suggests that the class variance (intraclass distance) is the major determinant of the better separability achieved by SRS since we obtained the worst result by SRS when the Euclidean distance was used. We evaluated in detail the role of the class variance (intraclass distance) in determining class separability by analyzing the variations of both intra- and interclass distances. This evaluation was performed by averaging all the intra- and interclass distances for the Indian Pines dataset using the features identified by SRS and SFS at each iteration. There are 13 intraclasses (class variances) per feature and 78 interclass distances per feature set. The number of features in a set increases in each iteration. Therefore, for a given feature set with k bands or channels, we calculated the mean value of $k \times 13$ intraclass and 78 interclass distances. We consider the results of the two most comprehensive separability measures: JM and TD.

As expected (see Sec. 3.2), the mean intraclass distance (class variance) obtained with the spectral bands is smaller than with individual channels in most cases [Fig. 9(a)], i.e., the mean intraclass distance becomes smaller when the channel signals are averaged. The difference increases with the number of bands. Analyzing the interclass distances [Fig. 10(b)] reveals that none of the methods performs better than the other ones since each method sometimes gives a higher mean interclass distance (distance between the class centroids).



(a) (b)

Fig. 9 The mean (a) intraclass and (b) interclass distance of 13 classes for the Indian Pines dataset with a different number of features when JM and TD measures are applied with the SRS and SFS search algorithms.

From the three experiments, it can be concluded that a bandset created by SRS with the proper number of bands gives a better class separability than the other channel selection methods with the same criterion, and this is mainly due to the reduction of the intraclass distances because of averaging the channel signals. The SRS separability is lower in the initial iterations because the initial SRS bands are usually very broad. In this case, the class variances of the constitutive channels are not usually equal, and an individual channel having a small class variance may yield a better class separability. For example, a vegetation class has a larger NIR than VIS variance, so a narrow VIS channel is likely to give a smaller variance than the average of all the variances across the spectrum. After the first iterations, however, the newly formed spectral bands comprise channels with comparable variances, and averaging the channels in every band yields a better separability of the classes in a scene.

5.4 *Classification Accuracy*

Finally, we applied two classifiers to the selected bands and channels and computed the accuracy of classification to evaluate and compare performance. In this experiment, maximum likelihood classifier (MLC) and SVM were applied to the final channels and bands identified by SFS, SFFS, and SRS using the JM and TD metrics. The classification accuracies are shown in Fig. 10 for the two scenes. Table 5 also gives the accuracy of six instances of feature sets determined by SFS-JM and SRS-JM.

The classification accuracy obtained with the spectral bands identified by SRS was higher than the accuracy obtained with the channel selection methods. When the number of features was small, SFS and SFFS might give better results. On the other hand, with a maximum of four spectral features, SRS gave higher accuracy than both SFS and SFFS for the Indian Pines scene when MLC was used as the classifier. For the Salinas scene, there were larger fluctuations in the classification accuracy, i.e., sometimes the channel selection methods gave better results when the number of bands was low, while with higher number of bands SRS had better accuracy. MLC gave a larger improvement in classification accuracy with the bandset created by SRS than with SVM. On average, the improvements were about 3.24% and 0.96% when MLC and SVM were used, respectively. This is due to the reduction in class variances having a larger impact on a parametric classifier like MLC that considers the distributions of class attributes explicitly.

5.5 *Comparison with a Wrapper Approach*

The last results revealed that the band configurations identified by SRS gave a noticeable improvement in classification accuracy compared with the channel selected by SFS and SFFS, especially when the MLC was used. Although our focus was on FS based on the filter approach, we compared the results of classification with an algorithm based on the wrapper

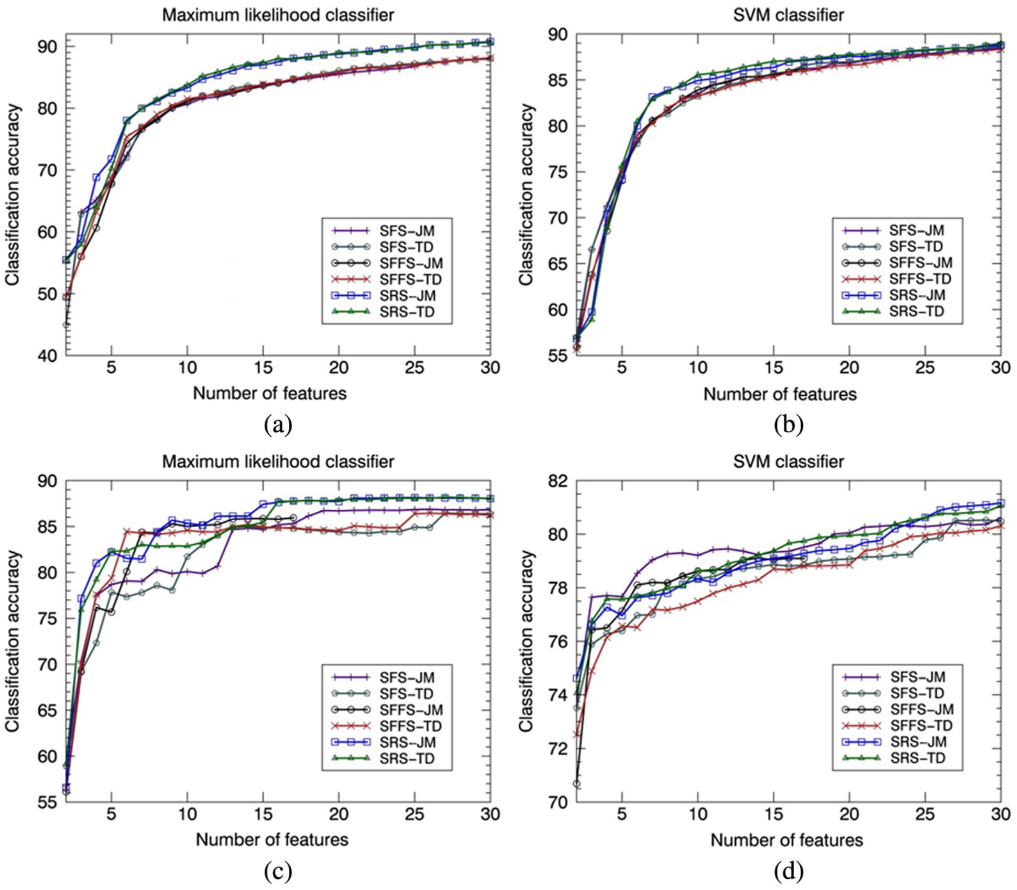


Fig. 10 The classification accuracy obtained with the classifiers MLC and SVM applied to the bands and channels identified by SRS, SFS, and SFFS using the JM and TD separability metrics; (a) and (b) Indian Pines and (c) and (d) Salinas scenes.

Table 5 The classification accuracy of two DR methods (SFS-JM and SRS-JM) based on the filter approach and one method (SVM-RFE) based on the wrapper approach.

Dataset		Indian Pines					Salinas				
Classifier		MLC		SVM			MLC		SVM		
Method		SFS-JM	SRS-JM	SFS-JM	SRS-JM	SVM-RFE	SFS-JM	SRS-JM	SFS-JM	SRS-JM	SVM-RFE
Number of features	5	68.73	71.78	75.53	74.18	43.01	78.69	82.22	77.67	76.96	73.38
	10	80.65	83.27	83.43	84.94	72.22	80.09	85.35	79.21	78.33	81.28
	15	83.53	86.99	83.43	84.94	76.70	84.71	87.42	79.33	79.10	83.16
	20	85.50	88.72	86.80	87.55	81.21	86.71	87.72	80.04	79.47	87.03
	25	86.76	89.86	87.68	88.19	84.21	86.85	88.15	80.28	80.62	88.05
	30	87.95	90.76	88.41	88.78	87.76	86.82	88.04	80.61	81.17	88.17

models[35,47,48] as well. In the wrapper models, features (channels) are usually selected given the classifier based on the accuracy of classification. It is usually assumed that the wrapper approaches achieve better classification accuracy than the filter approaches since they identify features that better suit the classification algorithm regarding the performance.[49–51] The wrapper approaches, on the other hand, are computationally more expensive than the filter

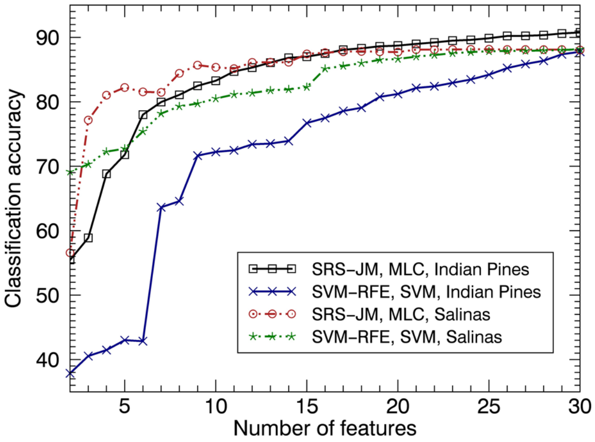


Fig. 11 The classification accuracy obtained by a filter approach, i.e., features identified by SRS-JM and the classification with MLC and one wrapper approach (SVM-RFE) for two datasets. The legend gives in each row the feature identification algorithm, the classification method, and the dataset used, respectively.

approaches.[52,53] We examined a well-known wrapper algorithm for channel selection on the basis of classification accuracy in comparison with a method giving a better classification result in our experiments: SRS-JM combined with MLC. Infact, SRS-JM identifies features based on maximizing class separability, to be applied with a classification algorithm, whereas a wrapper algorithm selects features to maximize classification accuracy.

We used SVM-RFE (recursive feature elimination) as a wrapper algorithm[50,51,54–56] in this study. In RFE, the decision function of the SVM, i.e., finding an optimal hyperplane that maximizes the marginal distance between two classes, is used as the criterion to select features in a backward elimination approach. It computes ranking weights based on the training samples for all the features and sorts the features according to weight vectors.[51,56]

For the SVM-RFE algorithm implemented in this experiment, we considered the commonly used "one-against-all' strategy.[43,54] The "one-against-all" strategy converts the problem of k classes ($k > 2$) into k dual-class problems. The radial basis function[46,52,53] is utilized as the kernel function to map the original feature space into the higher dimensional space. For the kernel width factor (γ) and regularization parameter (C), we applied different values suggested in the literature and chose the ones giving the best classification performance, i.e., $\gamma = 0.1$ and $C = 2000$ proposed by Ref. 54

In this experiment, we investigated which combination of features and classifiers would give a higher classification accuracy. Thus, the accuracy achieved with the spectral configuration constructed by SRS-JM, combined with MLC, was compared with the accuracy obtained by SVM-RFE for each dataset (Fig. 11). Table 5 also gives the accuracy obtained with six feature sets determined by SVM-RFE in comparison with SRS-JM.

It was observed in this experiment that the combination of the features identified by SRS-JM and MLC gave a better or comparable classification accuracy than the wrapper method. The SVM-RFE combination starts with a classification accuracy of about 70% for the Salinas scene, increases gradually, and reaches the accuracy of SRS-JM with MLC when the number of features is about 25. For the Indian Pines dataset, SVM-RFE had a lower accuracy than SRS-JM in all cases, whereas the difference became less by increasing the number of features. The SRS-JM-MLC gave about 3% higher classification accuracy with 30 features.

In comparison with the SVM-RFE algorithm, the spectral configuration determined by SRS gave a better or comparable classification accuracy when MLC was used as the classifier and the number of features was less than 30.

5.6 *Discussion*

We compared SRS with three search strategies used in channel selection: BB, SFS, and SFFS, by applying different separability metrics: Euclidean, Mh, Bhattacharyya, divergence, TD, and JM.

In the first experiment, the comparison between SRS and BB indicated that SRS gives better results than the best-selected channels over small datasets with the same criterion. Then, SRS was compared with search algorithms widely used in channel selection, i.e., SFS and SFFS, using different separability metrics over complete hyperspectral datasets. The second experiment had a very similar outcome, i.e., better class discrimination by SRS with a higher number of bands. On average, at least four spectral bands are needed to have better separability than by selecting narrow channels.

The third experiment analyzed the effect of the broader spectral regions formed by SRS on two main factors of class discrimination: interclass and intraclass distances. The results of this experiment and the second one revealed that the main reason that SRS provides a better class separability is the reduction of intraclass distances due to the broader spectral bands identified by SRS. Having a better separability between the classes in a scene leads to a higher classification accuracy in a filter-based approach and, finally, a better identification of observed targets, as shown by the fourth experiment. Comparison with a wrapper approach, in the fifth experiment, revealed that the combination of SRS and MLC gave a better or comparable accuracy of classification.

6 Conclusion

The approach we proposed demonstrates the importance of averaging narrow channels in improving the class separability by utilizing a DR method of hyperspectral data, which identifies spectral regions with the aim of optimal class discrimination. We have shown that the algorithm, i.e., the SRS, applied with a class separability metric can provide a spectral configuration with a better class separability than the channels selected from the original dataset with the same criterion. The reason is that SRS does not preserve the width of the original spectral channels, unlike the channel selection methods, and it defines a new feature space by merging the adjacent spectral channels if it is necessary. Averaging the narrow adjacent channels results in a smaller intraclass distance by reducing the class variances, leading to an increase in class separability. We conclude that whenever the separability measures include the intraclass parameters, SRS provides a better class discrimination when there is an adequate number of bands. The experiments were implemented on two different hyperspectral datasets including various types of classes. Eventually, the scenes were classified applying the selected bands showing that SRS increased the accuracy of classification by about 2%, when compared with the other filter-based approaches.

Disclosures

The authors declare no conflict of interest.

References

1. R. O. Green et al., "Imaging spectroscopy and the airborne visible infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.* **65**(3), 227–248 (1998).
2. G. Vane et al., "The airborne visible infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.* **44**(2–3), 127–143 (1993).
3. A. F. H. Goetz, "Three decades of hyperspectral remote sensing of the earth: a personal view," *Remote Sens. Environ.* **113**, S5–S16 (2009).
4. S. Le Moan et al., "A constrained band selection method based on information measures for spectral image color visualization," *IEEE Trans. Geosci. Remote Sens.* **49**(12), 5104–5115 (2011).
5. S. Kumar, J. Ghosh, and M. M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.* **39**(7), 1368–1379 (2001).
6. C. Chen et al., "Reconstruction of hyperspectral imagery from random projections using multihypothesis prediction," *IEEE Trans. Geosci. Remote Sens.* **52**(1), 365–374 (2014).

7. M. Fauvel et al., "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE* **101**(3), 652–675 (2013).
8. S. Jia et al., "Unsupervised band selection for hyperspectral imagery classification without manual band removal," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing* **5**(2), 531–543 (2012).
9. G. A. Shaw and H. K. Burke, "Spectral imaging for remote sensing," *Lincoln Lab. J.* **14**(1), 3–28 (2003).
10. T. Oommen et al., "An objective analysis of support vector machine based classification for remote sensing," *Math. Geosci.* **40**(4), 409–424 (2008).
11. A. Plaza et al., "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. Geosci. Remote Sens.* **43**(3), 466–479 (2005).
12. J. T. Peng, Y. C. Zhou, and C. L. P. Chen, "Region-kernel-based support vector machines for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.* **53**(9), 4810–4824 (2015).
13. G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory* **14**(1), 55–63 (1968).
14. M. Kamandar and H. Ghassemian, "Maximum relevance, minimum redundancy band selection for hyperspectral images," in *19th Iranian Conf. on Electrical Engineering (ICEE)*, pp. 1–5 (2011).
15. M. C. Alonso, J. A. Malpica, and A. M. de Agirre, "Consequences of the Hughes phenomenon on some classification techniques," in *ASPRS 2011 Annual Conf.*, Milwaukee, Wisconsin, pp. 1–5 (2011).
16. I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognit.* **43**(1), 5–13 (2010).
17. A. Martinez-Uso et al., "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.* **45**(12), 4158–4171 (2007).
18. J. S. Han, S. W. Lee, and Z. Bien, "Feature subset selection using separability index matrix," *Inf. Sci.* **223**, 102–118 (2013).
19. R. Huang and M. Y. He, "Band selection based on feature weighting for classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.* **2**(2), 156–159 (2005).
20. H. Yang et al., "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.* **8**(1), 138–142 (2011).
21. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York (1973).
22. P. H. Swain and S. M. Davis, *Remote Sensing: The Quantitative Approach*, McGraw-Hill, New York (1978).
23. R. A. Schowengerdt, *Remote Sensing, Models, and Methods for Image Processing*, Academic Press, San Diego (1997).
24. J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*, Springer, Berlin (2006).
25. C. I. Chang et al., "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.* **37**(6), 2631–2641 (1999).
26. J. Wang and C. I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.* **44**(6), 1586–1600 (2006).
27. Y. Yuan, G. K. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Trans. Geosci. Remote Sens.* **53**(2), 631–644 (2015).
28. B. F. Guo et al., "Band selection for hyperspectral image classification using mutual information," *IEEE Geosci. Remote Sens. Lett.* **3**(4), 522–526 (2006).
29. A. C. Jensen and A. S. Solberg, "Fast hyperspectral feature reduction using piecewise constant function approximations," *IEEE Geosci. Remote Sens. Lett.* **4**(4), 547–551 (2007).
30. S. Prasad and L. M. Bruce, "Decision fusion with confidence-based weight assignment for hyperspectral target recognition," *IEEE Trans. Geosci. Remote Sens.* **46**(5), 1448–1456 (2008).

31. C. Cariou, K. Chehdi, and S. Le Moan, "BandClust: an unsupervised band reduction method for hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Lett.* **8**(3), 565–569 (2011).
32. P. Narendra and K. Fukunaga, "Branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.* **C-26**(9), 917–922 (1977).
33. A. W. Whitney, "Direct method of nonparametric measurement selection," *IEEE Trans. Comput.* **C-20**(9), 1100–1103 (1971).
34. P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature-selection," *Pattern Recognit. Lett.* **15**(11), 1119–1125 (1994).
35. R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.* **97**(1–2), 273–324 (1997).
36. E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theor. Comput. Sci.* **209**(1–2), 237–260 (1998).
37. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
38. T. H. Cormen, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts (2009).
39. D. A. Bader, W. E. Hart, and C. A. Phillips, *Parallel Algorithm Design for Branch and Bound*, Springer, New York (2005).
40. P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(7), 900–912 (2004).
41. J. M. Sotoca, F. Pla, and J. S. Sanchez, "Band selection in multispectral images by minimization of dependent information," *IEEE Trans. Syst., Man, Cybern. C* **37**(2), 258–267 (2007).
42. C. Spence and P. Sajda, "The role of feature selection in building pattern recognizers for computer-aided diagnosis," *Proc. SPIE* **3338**, 1434–1441 (1998).
43. J. R. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*, Prentice Hall Press, Upper Saddle River, New Jersey (2015).
44. K. Biehl, "MultiSpec$^{©}$, a freeware multispectral image data analysis system," Copyright © 1994-2017 Purdue Research Foundation, https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html.
45. D. Manuel Graña Romay, "Hyperspectral remote sensing scenes," http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes (7 April 2014).
46. W. Bubitzky, M. Granzow, and D. P. Berrar, *Fundamentals of Data Mining in Genomics and Proteomics*, Springer, New York (2007).
47. M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognit.* **35**(4), 835–846 (2002).
48. S. P. Lodha and S. Kamlapur, "Dimensionality reduction techniques for hyperspectral images," *Int. J. Appl. Innovation Eng. Manage.* **3**(10), 92–99 (2014).
49. Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **44**(11), 3374–3385 (2006).
50. B.-C. Kuo et al., "A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(1), 317–326 (2014).
51. I. Guyon et al., "Gene selection for cancer classification using support vector machines," *Mach. Learn.* **46**(1), 389–422 (2002).
52. E. Sarhrouni, A. Hammouch, and D. Aboutajdine, "Band Selection and classification of hyperspectral Images using mutual Information: an algorithm based on minimizing the error probability using the inequality of Fano," in *Int. Conf. on Multimedia Computing and Systems (ICMCS)*, pp. 156–160 (2012).
53. A. Santos et al., "Feature selection for classification of remote sensed hyperspectral images: a filter approach using genetic algorithm and cluster validity." in *Proc. of the Int. Conf. on Image Processing, Computer Vision, and Pattern Recognition (IPCV '12)*, p. 1 (2012).
54. R. Zhang and J. Ma, "Feature selection for hyperspectral data based on recursive support vector machines," *Int. J. Remote Sens.* **30**(14), 3669–3677 (2009).

55. M. Pal, "Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data," *Int. J. Remote Sens.* **27**(14), 2877–2894 (2006).
56. M.-L. Huang et al., "SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier," *Sci. World J.* **2014**, 795624 (2014).

S. Enayat Hosseini Aria received his BSc degree in geomatics engineering and his MSc degree in remote sensing from the Faculty of Engineering, Tehran University, Iran. He is now a PhD student in the Geoscience and Remote Sensing Department, TUDelft, the Netherlands. He has been involved in several projects in the field of remote sensing, including design and implementation of hyperspectral nanosatellite and monitoring environmental changes using satellite images. His research interests include image processing and machine learning.

Massimo Menenti is an internationally renowned scientist in the fields of Earth observation and global terrestrial water cycle. His achievements have been attained in the aspects of surface parameter retrievals from remote sensing-based evapotranspiration (ET) estimation and time series analysis of satellite data. He is one of the earliest researchers in using laser radar technology to measure surface aerodynamic roughness. He also presented the surface energy balance index (SEBI) theory for ET estimation.

Ben G. H. Gorte received his MSc degree in applied mathematics/computer science and his PhD with a thesis on classification and segmentation of remote sensing images. He has been working on various subjects in computer vision, photogrammetry, and remote sensing for applications in traffic monitoring, precision agriculture, urban hydrology, and so on. His current research is on algorithms for 3-D-reconstruction, change detection, and deformation monitoring from images and point clouds.