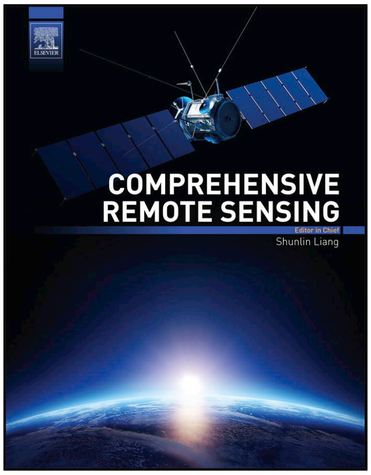


From Tochon, G., Dalla Mura, M., Veganzones, M. A., Valero, S., Salembier, P., & Chanussot, J. (2018) Advances in Utilization of Hierarchical Representations in Remote Sensing Data Analysis. In S. Liang (Ed.), Comprehensive Remote Sensing, vol. 2, pp. 77–107. Oxford: Elsevier.

ISBN: 9780128032206

Elsevier

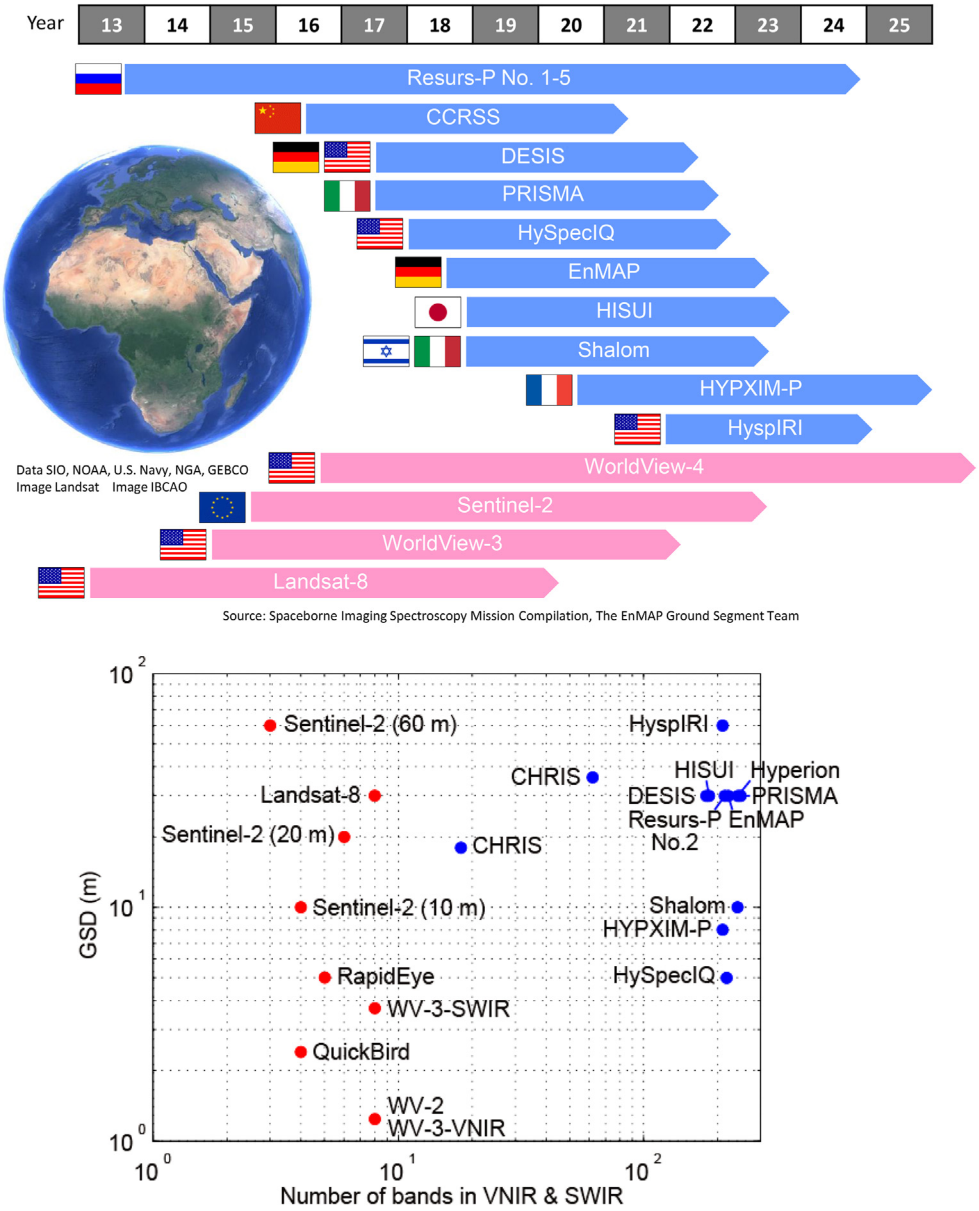
# 2.05 Advances in Utilization of Hierarchical Representations in Remote Sensing Data Analysis

**G Tochon,** LRDE, EPITA, Le Kremlin-Bicêtre, France
**M Dalla Mura and MA Veganzones,** GIPSA-Lab, Saint Martin d'Hères, France
**S Valero,** CESBIO, Toulouse, France
**P Salembier,** Technical University of Catalonia (UPC), Barcelona, Spain
**J Chanussot,** GIPSA-Lab, Saint Martin d'Hères, France; and University of Iceland, Reykjavík, Iceland

## 2.05.1 Introduction

Thanks to the recent technical advances in sensor design, the new generation of remote sensing imagers for Earth observation missions comes with a notably increased spatial resolution for the collected data. Ground sampling distances are now reaching the order of a few meters for the latest spaceborne hyperspectral and radar sensors, and even submetric resolution (down to a few dozen centimeters) for panchromatic and multispectral images. The amount of data to process is additionally constantly growing, resulting from the multitude of current acquisition campaigns as well as the forthcoming ones, as illustrated by **Fig. 1**.

Acquiring images with finer spatial resolutions has for direct corollary that pixels correspond to smaller on-ground sites. This allows to make a more accurate analysis of the geometrical features of the objects composing the scene under study. However, the intrinsic mixture of land covers in natural landscapes and the huge amount of details present in urban areas make the analysis both particularly complex and demanding. In addition, pixels no longer map to real semantic objects, but rather to small portions of them, preventing a high-level analysis of images at the pixel scale. Over the last decade, real efforts have been devoted to the integration of spatial information to typical pixel-based processing, such as image denoising (Xie et al., 2002), superresolution (Gu et al., 2008), or classification (Fauvel et al., 2013) and have demonstrated the benefits of taking into account the contextual structure of the scene to improve its analysis and interpretation. Object-based image analysis (OBIA) is a widely accepted paradigm in image processing where the image analysis is no longer based on its pixels, but it is rather conducted at a higher level of abstraction through the definition of some suited regions in the image. In OBIA, the regions (being groups of adjacent pixels) bear some semantic meaning, and open the door to tasks, which were impossible to achieve when working at the pixel level, such as object recognition and identification.



Source: Spaceborne Imaging Spectroscopy Mission Compilation, The EnMAP Ground Segment Team

**Fig. 1** Principal ongoing and forthcoming spaceborne Earth observation missions (*top*) with the corresponding sensor characteristics (ground sampling distance and number of bands) (*bottom*, image borrowed from (Yokoya et al., 2016)).

When applied to remote sensing images, OBIA approaches are commonly used to define features or descriptors for each region (such as size, shape, orientation, etc.), that are typically valuable for land cover applications (Blaschke, 2010; Lang, 2008). The critical stage of OBIA is obviously the segmentation step, namely, the definition of the regions comprised in the image, as poorly estimated regions would impact all subsequent processing relying on this information. A major issue is that image segmentation is an ill-posed problem, as an image can be segmented appropriately in many different ways. This multiplicity of acceptable solutions comes from the notion of scale of analysis: regions of interest can be defined in the image following various levels of details, and the best level to conduct the analysis depends upon the application. A potential solution to this intrinsic multiscale nature of images is to use hierarchical representations, which encompass in their structure all potential scales of interest in the image. The hierarchical representation can then be built once, regardless of the application, and then stored for further processing. This latter is, on the other hand, driven by the underlying application in order to produce the desired result. Thus, a single hierarchical representation for a given image can lead to various outputs.

The first instance of decomposing an image into a hierarchical structure to analyze it at multiple scales goes back to Finkel and Bentley (1974), which introduced the quad-tree to that purpose. If they have found many applications, such as image compression or segmentation (Samet, 1984), quad-trees cannot efficiently account for complicate and irregular region boundaries. Later on, hierarchical structures such as component trees (Salembier et al., 1998) (also called min-trees and max-trees) and inclusion trees (Monasse and Guichard, 2000) (also known as the trees of shapes) were introduced in the mathematical morphology community. While the former describe a gray level image as a set of regional minima or maxima included within each other, the latter are based on the notion of shapes within the image and can be viewed as the merging of a min-tree and a max-tree, with the desirable property of being self-dual (dark and bright components are processed the same way). The reader is referred to the recent work of Cavallaro et al. (2015) for an extensive review regarding the use of component trees and inclusion trees for very-high-resolution remote sensing data analysis. Another popular hierarchical representation is the binary partition tree (BPT), initially proposed by Salembier and Garrido (2000). Contrary to components and inclusion trees that rely on local minima and maxima in the image, defined according to the absolute pixel values (hence being easily computable for grayscale images (Carlinet and Géraud, 2014) but very challenging to extend to multichannel images), the BPT representation relies on the notion of similarity between adjacent pixels or regions. Within the BPT, the regions of interest are those that sufficiently differ from their surroundings, with respect to a similarity measure that can be defined according to the specificities of the handled image. Thanks to this flexibility in its definition, the BPT structure has been utilized for various applications in image and video processing, such as image segmentation (Valero et al., 2013), filtering (Alonso-González et al., 2012), compression (Salembier and Garrido, 2000) as well as object detection (Vilaplana et al., 2008), and object tracking (Palou and Salembier, 2013).

This article draws a comprehensive review of the most recent works involving the BPT representation for various remote sensing data analysis tasks, as there is up to now no such review in the literature. This article is organized as follows: section "**Hierarchical Representations**" features the introduction of all notations and definitions used throughout this article, as well as the description of the most common hierarchical representations under a mathematical morphology perspective. Section "**The Binary Partition Tree Representation**" focuses more in detail on the BPT structure, introducing from a conceptual point of view the key aspects related to its construction and analysis. Section "**Applications of the BPT Representation in Remote Sensing Data Analysis**" presents in detail several works related to the exploitation of BPT representations in remote sensing, for a broad range of applications. It includes the segmentation of hyperspectral and SAR polarimetry (PolSAR) images by means of an energetic minimization formulated on the BPT representation (sections "**Hyperspectral Image Segmentation**" and "**Segmentation of PolSAR images**," respectively), the classification of hyperspectral images (section "**Hyperspectral Image Classification**"), the recognition and extraction of buildings from a urban scene (section "**Object Recognition in Urban Hyperspectral Scene**"), the detection of anomalies within hyperspectral images (section "**Anomaly Detection in Hyperspectral Data**"), and the filtering of speckle noise in PolSAR data (section "**Speckle noise filtering**"). Each application is organized following the same scheme, namely, the general problem statement and the potential of hierarchical representations with respect to it, a detailed description of the implemented solution using the BPT representation, and illustrative examples selected from the mentioned references. Finally, section "**Conclusion**" concludes this article, that is expected to present the reader with the huge potential of BPT structures, and provide a real knowhow related to BPT processing as well as their possible tailoring to any remote sensing data analysis application.

## 2.05.2 Hierarchical Representations

This section introduces hierarchical representations through the prism of mathematical morphology, being the field which has historically been mostly concerned with the development of such structures (Najman and Cousty, 2014).

### 2.05.2.1 Base Definitions and Notations

In mathematical morphology, a generic image $\mathcal{I}$ is defined as a function

$$\begin{array}{rccc} \mathcal{I}: & E & \rightarrow & V \\ & x & \mapsto & \mathcal{I}(x) \end{array} \tag{1}$$

where $E$ is the spatial support of $\mathcal{I}$ (and is commonly defined as a subset of $\mathbb{Z} \times \mathbb{Z}$ to represent the pixel grid), and the space of pixel values $V$ depends on the recorded data type. For multichannel optical images for instance, $V \subseteq \mathbb{R}^N$ with $N$ being the number of channels (note that, in that case, the pixel vector $\mathcal{I}(x) \epsilon \mathbb{R}^N$ will be denoted by a bold symbol $\mathbf{x} = (x(1), \ldots, x(N))$ if there is no ambiguity).

A region (or class) $\mathcal{R}$ of the image $\mathcal{I}$ is defined as a subset $\mathcal{R} \subseteq E$ of its spatial support, that is, a set of pixels $\mathcal{R} = \{x_i \in E\}$. There is no requirement for $\mathcal{R}$ to be connected (in one single piece), although it will be implicitly assumed in the following.

A partition of $E$, denoted $\pi$, is a family $\{\mathcal{R}_i \subseteq E\}$ of regions of $E$ such that $\mathcal{R}_i \cap \mathcal{R}_{j \neq i} = \varnothing$ and $\cup_i \mathcal{R}_i = E$, that is, a division of $E$ into nonoverlapping regions, which entirely cover $E$. The set of all partitions of $E$ is denoted $\Pi_E$. It is possible to compare partitions through the binary ordering $\leq$ defined on $\Pi_E$, called the refinement ordering. For any two $\pi_i, \pi_j \in \Pi_E$, one says that $\pi_i$ refines (or is a refinement of) $\pi_j$, and one writes $\pi_i \leq \pi_j$, whenever for each $\mathcal{R}_i \in \pi_i$, there exists $\mathcal{R}_j \in \pi_j$ such that $\mathcal{R}_i \subseteq \mathcal{R}_j$. In other words,

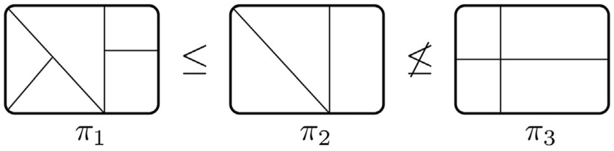


**Fig. 2** Illustration of the refinement ordering: $\pi_1 \leq \pi_2$, but $\not\leq \pi_3$..

$\pi_i$ is a refinement of $\pi_j$ if every individual region $\mathcal{R}_j \in \pi_j$ can be fragmented into one or several regions $\mathcal{R}_i \in \pi_i$, as it is the case of $\pi_1$ and $\pi_2$ in Fig. 2. However, the refinement ordering $\leq$ is only a partial order and not every two partitions are comparable. This is the case in particular for $\pi_1$ and $\pi_3$ displayed by Fig. 2.

## 2.05.2.2 Tree-Based Image Representations

Natural images accommodate well with hierarchical representations, since regions of interest within an image are very often either disjoint or nested within each other. To support this observation, tree-based representations have been developed as efficient image representation structures.

### *2.05.2.2.1 Generalities on tree-based representations*

A tree-based representation $\mathcal{T}$ of $E$ is a collection of regions $\mathcal{T} = \{\mathcal{R} \subseteq E\}$ such that $\emptyset \notin \mathcal{T}$, $E \in \mathcal{T}$ and $\forall \mathcal{R}_i$. In other words, a tree-based representation of $E$ is a decomposition of $E$ into regions that are either disjoint, or nested. A tree-based representation $\mathcal{T}$ can be represented by an undirected acyclic graph $\mathcal{G}_\mathcal{T} = (\mathcal{V}_\mathcal{T}, \mathscr{E}_\mathcal{T})$, or dendrogram, where each vertex (also called node) $v \in \mathcal{V}_\mathcal{T}$ is associated with a region $\mathcal{R} \in \mathcal{T}$ and each edge $e_{i,j} \in \mathscr{E}_\mathcal{T}$ means that either $\mathcal{R}_i \subseteq \mathcal{R}_j$ or $\mathcal{R}_j \subseteq \mathcal{R}_i$. In order to simplify the notations, we will denote by $\mathcal{R}$ both the regions of the tree-based representation $\mathcal{T}$ and the vertices of its dendrogram $\mathcal{G}_\mathcal{T}$. The terminology related to tree-based representations is largely borrowed from graph theory, thanks to this tree/graph equivalence:

- The children of $\mathcal{R}$ correspond to all regions $\mathcal{R}' \in \mathcal{T}$ that are directly connected to $\mathcal{R}$ in $\mathcal{G}_\mathcal{T}$ and such that $\mathcal{R}' \subseteq \mathcal{R}$. The set of children of $\mathcal{R}$ is denoted $\mathsf{C}(\mathcal{R})$.
- If $\mathcal{R}$ has no children, that is, $|\mathsf{C}(\mathcal{R})| = 0$, then $\mathcal{R}$ is called a leaf of $\mathcal{T}$. $\mathsf{leaves}(\mathcal{R})$ is the set of leaves of $\mathcal{T}$ that are included in $\mathcal{R}$.
- The father of $\mathcal{R}$ is the node $\mathsf{F}(\mathcal{R})$ to which $\mathcal{R}$ is connected and such that $\mathcal{R} \subseteq \mathcal{F}(\mathcal{R})$. In a tree-based representation, each region has exactly one father, except for the root of $\mathcal{T}$, $\mathsf{root}(\mathcal{T})$, which has none.
- The sibling of $\mathcal{R}$ is the set of regions $\mathsf{Sib}(\mathcal{R})$ that have the same father as $\mathcal{R}$, that is, $\mathcal{R}' \in \mathsf{Sib}(\mathcal{R}) \Leftrightarrow \mathsf{F}(\mathcal{R}') = \mathsf{F}(\mathcal{R})$.
- The branch of $\mathcal{R}$, denoted $\mathsf{br}(\mathcal{R})$ is the set of regions $\{\mathcal{R}, \mathsf{F}(\mathcal{R}), \mathsf{F}(\mathsf{F}(\mathcal{R})), \ldots, \mathsf{root}(\mathcal{T})\}$. Elements of $\mathsf{br}(\mathcal{R}) \backslash \{\mathcal{R}\}$ are called ancestors of $\mathcal{R}$.
- The height of $\mathcal{R}$, $\mathsf{h}(\mathcal{R})$, is number of elements in $\mathsf{br}(\mathcal{R})$ minus 1, that is, $\mathsf{h}(\mathcal{R}) = |\mathsf{br}(\mathcal{R})| - 1$. It corresponds to the length of the path linking $\mathcal{R}$ to the root node. The height of the root node is set by convention to 0.
- The subtree rooted at $\mathcal{R}$, $\mathcal{T}(\mathcal{R})$, corresponds to all the elements of $\mathcal{T}$ that are included in $\mathcal{R}$. In other words, it contains all the elements of $\mathcal{T}$ for which $\mathcal{R}$ is an ancestor.

This terminology is summarized by Fig. 3.

### *2.05.2.2.2 Examples of tree-based representations*

Classical tree-based representation includes the min-tree and max-tree, also known as component trees. Initially proposed by Salembier et al. (1998), these tree-based representations encode the inclusion relationship between the connected components of the upper and lower level sets of the image. More specifically, let $\mathcal{I} : E \rightarrow V \subseteq \mathbb{R}$ be a gray-scale image. Its upper and lower level sets, for a threshold value $h$, are defined by $\mathcal{I}^h = \{x \in E | \mathcal{I}(x) \geq h\}$ and $\mathcal{I}_h = \{x \in E | \mathcal{I}(x) \leq h\}$, respectively. $\mathcal{I}^h$ and $\mathcal{I}_h$ are binary images, composed of connected components, where each connected component $C^h$ (respectively, $C_h$) corresponds to a set of connected pixels whose value is above (respectively, below) the threshold $h$. By varying the threshold $h$, one then obtains a hierarchical decomposition of the image into a set of connected components. The min-tree represents this hierarchical decomposition by encoding the inclusion relationship between the connected components of the lower level sets of the image. The leaves of the min-tree are the regional minima of the image. Conversely, the max-tree encodes the inclusion between the connected components of the upper level set decomposition, and has the local maxima as leaves. Examples of min-tree and max-tree are displayed by Fig. 4. The main limitation of component trees is that they handle bright and dark components separately, which can be an issue when some object of interest appears brighter than the background in some parts of the image, and darker in some other parts.

To handle bright and dark components in a self-dual way, several authors have introduced the notion of shapes, which have led to the definition of the tree of shapes (ToS) (also called inclusion tree in Monasse and Guichard (2000)). Instead of considering the connected components of the upper and lower level set decompositions, the ToS encodes the inclusion relationship between the level lines (i.e., the topological boundaries of the connected components). More particularly, a shape is defined as a connected component with holes filled, and the ToS of an image can be viewed as a merging between the min-tree and max-tree of this image. All leaves of the ToS correspond to some regional minima and maxima of the image, as shown in Fig. 4.

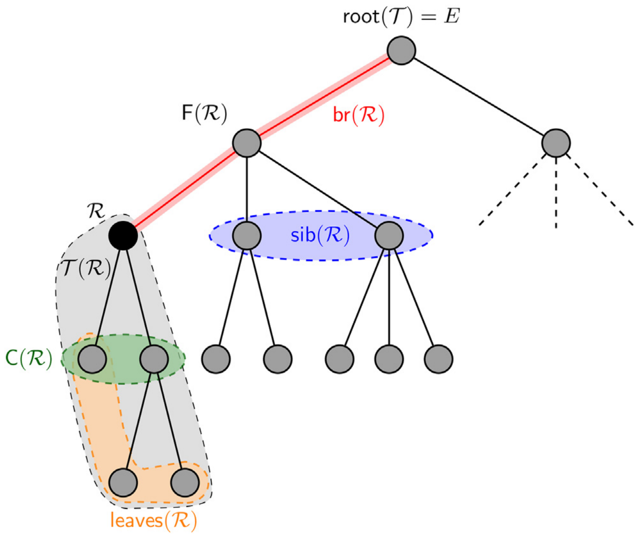


**Fig. 3** Tree-based representation terminology.

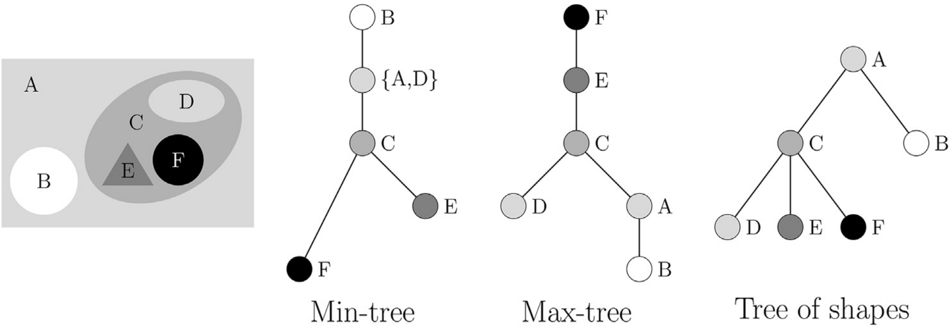


**Fig. 4** Examples of tree-based representations of the *gray*-scale image on the *left*: min-tree, max-tree and tree of shapes.

Component and inclusion trees find numerous image-processing applications, such as image filtering (Salembier and Serra, 1995; Xu et al., 2012b), image segmentation (Cardelino et al., 2006; Jones, 1999) of object recognition (Pan et al., 2009). In remote sensing data analysis, component and inclusion trees are employed for very-high-resolution image filtering through the use of attribute profiles (Dalla Mura et al., 2010) and self-dual attribute profiles (Dalla Mura et al., 2011), respectively. An extensive review is provided in Cavallaro et al. (2015). However, both structures rely on the absolute pixel scalar values of the image and the natural total ordering for this set of scalar values, hence making their extension from gray-scale to multivalued image challenging. In addition, there is no guarantee that the objects of interest in the analyzed image can be appropriately described only by their own pixel values, let alone the fact that those values may not correspond to extrema in the image. As a matter of fact, an object seems to be of interest if it sufficiently differs from its surrounding. This leads to work on dissimilarities between pixels (or regions) rather than on their absolute values, in particular through the introduction of a distance or dissimilarity function, which is notably the purpose of hierarchies of partitions.

## 2.05.2.3 Hierarchies of Partitions

### 2.05.2.3.1 Generalities on hierarchies of partitions

Hierarchies of partitions are a special case of tree-based image representations. As a matter of fact, the definition of a tree-based representation, as given in section "Tree-Based Image Representations," can be slightly complemented to define a hierarchy of partitions $H$ of $E$, as a collection of regions $H = \{\mathcal{R} \subseteq E\}$ such that

$$\begin{aligned} &- \quad \emptyset \notin H \\ &- \quad E \in H \\ &- \quad \forall \mathcal{R}_i, \mathcal{R}_j \in H, \mathcal{R}_i \cap \mathcal{R}_j \in \{\emptyset, \mathcal{R}_i, \mathcal{R}_j\} \\ &- \quad \forall \mathcal{R} \in H \backslash \mathrm{leaves}(H), \mathcal{R} = \bigcup_{\mathcal{R}_c \in \mathrm{C}(\mathcal{R})} \mathcal{R}_c \end{aligned} \tag{2}$$

In addition to being composed of regions that are pairwise disjoint or nested, the additional requirement for a tree-based representation to be a hierarchy of partitions is that each nonleaf node in the hierarchy can be exactly recomposed from its children. In particular, it means that the whole space $E$ can be retrieved by taking the union of all leaves of the hierarchy (which was clearly not the case for the component and inclusion trees, see Fig. 4). These leaf regions form a partition of $E$, denoted $\pi_0$ and called the leaf partition of $H$.

Alternatively but equivalently, a hierarchy of partitions $H$ of $E$ can be defined as a finite sequence of partitions $\pi_i \in \Pi_E$ ordered by refinement

$$H = \{\pi_i\}_{i=0}^{n} \quad \text{such that} \quad i \le j \Rightarrow \pi_i \le \pi_j \tag{3}$$

It ranges from the leaf partition $\pi_0$ to the root of the hierarchy $\pi_n = \{E\}$. An example of hierarchy of partitions and its associated tree graph is displayed by Fig. 5. Thanks to these two equivalent definitions, it is possible to obtain a hierarchy either by working on the regions (for instance, using some region merging or splitting techniques) or on the partitions directly. Of course, the terminology defined for tree-based representations remains valid for hierarchies of partitions.

Processings that can be applied to hierarchies of partitions are categorized in two classes. Region-based processings aim at exploring the regions of the hierarchy in order to identify the regions of interest that fulfill some predefined criteria (for instance, a given shape, homogeneity, or distance with respect to the neighbors). These strategies are particularly useful to perform object detection and recognition. On the other hand, the goal of partition-based processings is to extract from the hierarchy some specific partitions that conform a given application. One particular way to proceed is through a pruning operation, namely, cutting of some branches of the hierarchy such that the new leaves of the pruned tree achieve the desired partition. In any case, effective and powerful processings can be achieved by combining low-level information extracted from the inherent graph structure of the hierarchical representation with higher level knowledge derived from the regions associated with the graph nodes (Bai et al., 2014; Salembier and Foucher, 2016; Xu et al., 2017).

### 2.05.2.3.2 Examples of hierarchies of partitions

As for tree-based representations, hierarchies of partitions have been widely studied in the literature. A well-known hierarchy of partitions is the quad-tree, proposed by Finkel and Bentley (1974). Starting from the whole image (i.e., the root of the hierarchy), the quad-tree is created by successive region splitting. More particularly, each region, also called quadrant, can be either divided into four subquadrant or left as it is, each quadrant being either square or rectangular. The decision of splitting a region into four subquadrants is often based on some homogeneity considerations: if the region is not homogeneous enough, it is split until it fulfills the desired criterion. Quad-trees have found applications in image segmentation (Spann and Wilson, 1985) and compression (Shusterman and Feder, 1994), notably. However, as each region is rectangular, quad-tree cannot account for irregular contours, and therefore, objects of interest are often split into several nodes. $\alpha$-tree, also known as the hierarchy of quasi flat zones (Meyer and Maragos, 2000), is another well-known hierarchical representation, based on the notion of constrained connectivity (Soille, 2008). More specifically, let $p$ and $q$ be two neighboring pixels, and $d(\mathcal{I}(p), \mathcal{I}(q))$ be the dissimilarity between their respective values for the image I and some distance function $d(\cdot)$. Pixels $p$ and $q$ are said to be $\alpha$-connected if there is a path from $p$ to $q$, namely, a sequence of $(p = x_1, \ldots, x_n = q)$ such that $x_i$ and $x_{i+1}$ are adjacent and $d(\mathcal{I}(x_i), \mathcal{I}(x_{i+1})) \le \alpha$. Following, the $\alpha$-connected component of a pixel $p(\alpha - \mathrm{CC}\ (p))$ is defined as

$$\alpha - \mathrm{CC}(p) = \{p\} \cup \{q \,\text{s.t}\; p \;\text{and}\; q \;\text{are}\; \alpha\text{-connected}\} \tag{4}$$

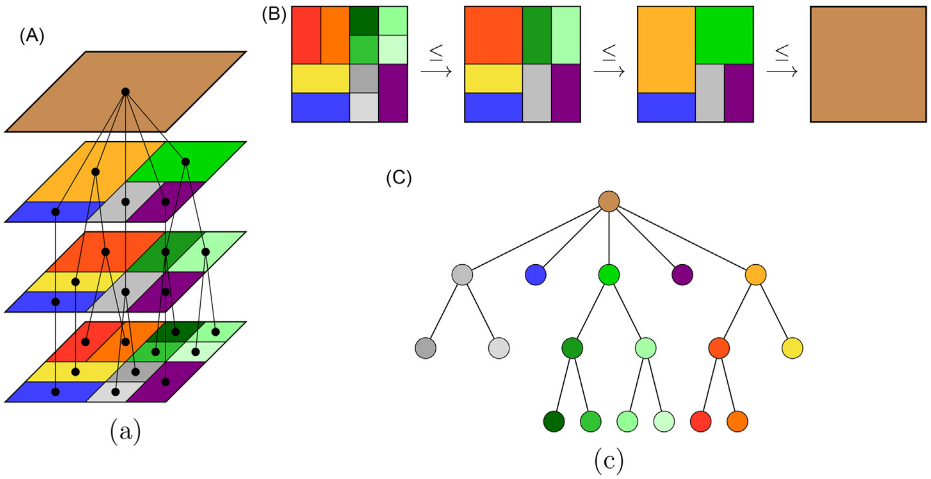


Fig. 5 Example of hierarchy of partitions, obtained as (A) a stack of ordered partitions and (B) the result of a region merging procedure, along with (C) the corresponding tree graph.

Soille (2008) showed that for a given $\alpha$ value, the set of $\alpha-$CC forms a partition $\pi_\alpha$ of $E$ and for two values $\alpha_1 \leq \alpha_2$, $\pi_{\alpha_1} \leq \pi_{\alpha_2}$. Therefore, by using several values $\alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_n$, one induces several partitions ordered by refinement that creates the $\alpha-$ tree hierarchy $H_\alpha = \{\pi_{\alpha_0} \leq \ldots \leq \pi_{\alpha_n}\}$. An example of such hierarchy is displayed by **Fig. 6**.

Based on the well-known hierarchical set-wise optimization procedure, Beaulieu and Goldberg (1989) and Tilton (1999, 2003) proposed the Hierarchical SEGmentation (HSEG) method. In the former, each iteration involves the search for the two adjacent regions that have the lowest pairwise distance. All pairs of regions achieving this distance are then merged. The HSEG algorithm is founded on the same idea, except that the adjacency constraint for regions is partially relaxed. Indeed, a user-chosen proportion of nonadjacent regions can also be merged at each iteration, provided that their distance is less than the minimal distance among all pairs of adjacent regions. For that reason, the HSEG algorithm can be viewed as sequentially alternating between a region growing step and a spectral clustering step. Due to its huge computational load, induced by the important number of pairwise distances that must be evaluated, Tilton (2010) further extended the HSEG algorithm to the recursive HSEG algorithm. This, based on a divide-and-conquer approximation of the HSEG, allows for parallel implementation and computational acceleration.

Finally, a popular hierarchy of partitions is the BPT, as proposed by Salembier and Garrido (2000). Starting from an initial partition $\pi_0$ that defines the leaves of the hierarchy, the BPT is obtained by a bottom-up region merging procedure: pairs of neighboring regions are merged based on their similarity until there is only one region remaining, which is the whole space $E$. The creation of a BPT is bound to the definition of the initial partition as well as the similarity function to assess how close are two neighboring regions. In the last decade, BPTs have proved to be a valuable tool for hierarchical image representation, thanks to the great flexibility of their construction and analysis processes. Consequently, they have found numerous applications in image and video processing such as image segmentation (Valero et al., 2013), filtering (Alonso-González et al., 2012), compression (Salembier and Garrido, 2000) as well as object detection (Vilaplana et al., 2008), and object tracking (Palou and Salembier, 2013). The next section "**The Binary Partition Tree Representation**" is devoted to a more detailed insight of BPTs from a conceptual point of view, while section "**Applications of the BPT Representation in Remote Sensing Data Analysis**" will present some of their applications for remote sensing data analysis.

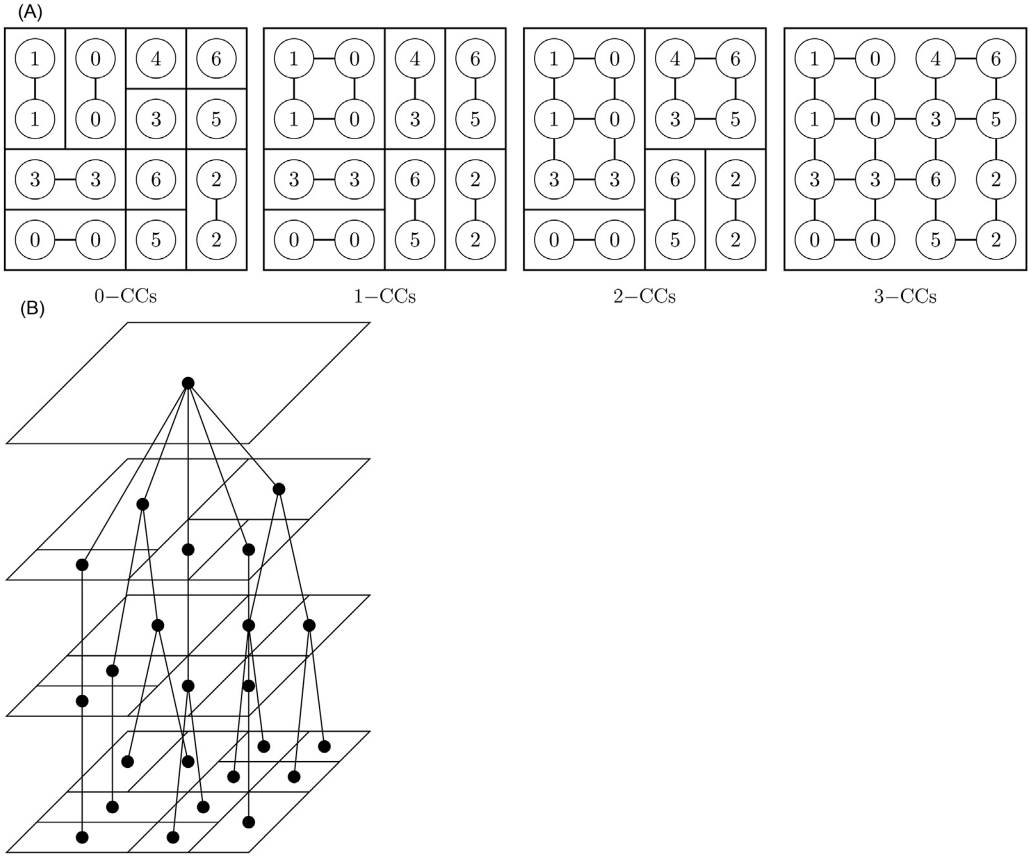


**Fig. 6** Example $\alpha-$ tree hierarchy: (A) $0-$, $1-$, $2-$, and $3-$ connected components of a toy image (with adjacency defined by 4-connectivity) and (B) the corresponding $\alpha-$ tree. In that example, the dissimilarity measure between pixels $p$ and $q$ is $d(\mathcal{I}(p), \mathcal{I}(q)) = |\mathcal{I}(p) - \mathcal{I}(q)|$.

## 2.05.3 The Binary Partition Tree Representation

Although they were not designed to achieve the same tasks, BPT representations rely on the same idea as hierarchical clustering techniques in machine learning (Rokach and Maimon, 2005), in the sense that they iteratively group pairs of regions based on their similarities, and they encode those relationships within a tree structure. This section introduces the key concepts related to the construction of BPTs as well as their most general processing techniques to achieve various image-processing applications such as filtering, segmentation, or object detection, among others.

### 2.05.3.1 Construction of the Binary Partition Tree

The BPT representation relies on the iterative merging procedure of a set of initial regions, which are the leaves of the BPT. The tree is built by keeping track of the merging order. Each region can be merged with only one of its neighbors, resulting in a hierarchy where each region has either two children, or none (in the case of a leaf node). An example of region merging sequence and its corresponding BPT is displayed in Fig. 7.

BPT representations enjoy several desirable properties:

- They decompose an image into a hierarchical set of regions, providing a description of this image at different scales ranging from fine to coarse. This is particularly valuable, since the hierarchical decomposition can serve as an initial support, computed regardless of the application, and its analysis can be performed afterward by tuning the scale according to the desired objective.
- Their construction is based on the merging of similar neighboring regions, and is therefore bound to a user-defined similarity measure (which does not necessarily rely on the absolute pixel values, in opposition to the component and inclusion trees, for example). This introduces a wide flexibility in the construction procedure.
- Even though their construction is rendered flexible by the various possible settings to parameterize the merging procedure, BPTs were intended to be built independently of the underlying application, as a common support basis for all subsequent processing (Salembier and Garrido, 2000). Nevertheless, their application-driven analysis well adapts to a broad range of image processings, and BPTs have been used in an extensive variety of applications in the image and video-processing fields (see, for instance, Alonso-González et al., 2013; Palou and Salembier, 2013; Valero et al., 2013; Vilaplana et al., 2008).

The two parameters of prime importance when building a BPT are the definition of the merging procedure, and the initial partition on which this procedure is applied. While there are numerous options available for those two parameters, some of them have proved to perform consistently well in the literature.

#### *2.05.3.1.1 The initial partition*

The initial partition $\pi_0$ is the parameter on which the region merging procedure is initialized. If a pertinent initial partition does not guarantee a pertinent BPT representation, a poor initial partition does lead to a poor hierarchical decomposition, as all the regions subsequently obtained follow from the initial ones. The simplest configuration for the initial partition is to build to BPT from the pixel level, that is, each individual pixel in the image constitutes a leaf. However, as a BPT built on an initial partition made of $|\pi_0|$ leaves is composed of $(2 \times |\pi_0| - 1)$ regions, the pixel level as an initial partition may lead to a huge BPT structure and this could be problematic from a computational point of view for very large images. Moreover, such BPT would be composed of many small and meaningless regions and this could also slow down the analysis processes further applied on it.

A given partition $\pi_0$ of the image is suited for a BPT representation of this image if it fulfills the following two requirements:

- Its regions should be fine enough (the image should be enough oversegmented) to ensure that the smallest regions of interest within the image are not already merged together in some initial regions of $\pi_0$. Otherwise, those regions of interest would be irremediably lost.

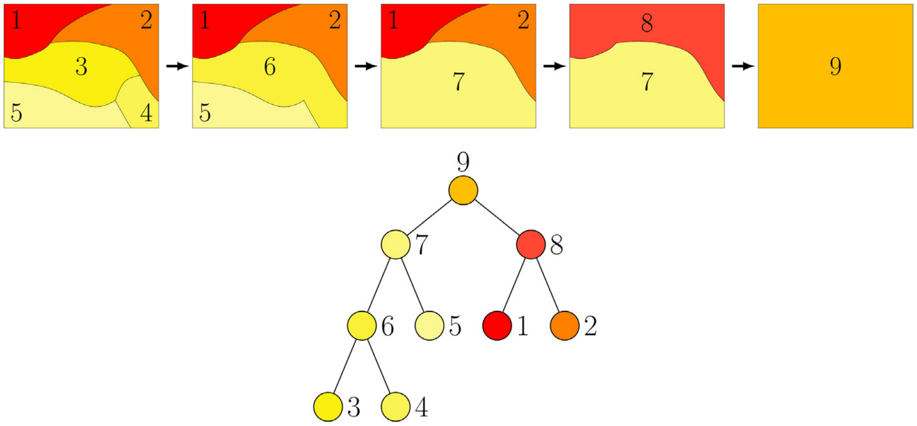


Fig. 7 Construction of a binary partition tree based on an iterative region merging sequence.

- The boundaries of the initial regions should well adhere to the real boundaries of objects of interest in the image, in order to accurately reconstruct these objects of interest (up to a suited definition of the region merging procedure).

For these reasons, efficient segmentation algorithms to design an initial partition include clustering techniques, such as traditional *k*-means (with *k* large enough) or mean shift clustering (Comaniciu and Meer, 2002), or superpixels generation methods such as the watershed algorithm (Vincent and Soille, 1991) (or its multidimensional extension for multivalued images (Tarabalka et al., 2010b)) or SLIC superpixels (Achanta et al., 2012).

### 2.05.3.1.2 The merging procedure

The merging procedure determines in which order the regions should be merged. The BPT is then built following a bottom-up procedure (i.e., starting from the smallest regions) by keeping track of this order. The specification of a merging procedure itself relies on the definition of two inner parameters:

- The region model $\mathrm{M}_{\mathcal{R}}$, which specifies how to mathematically model the regions and their union.
- The merging criterion $\mathcal{O}(\mathcal{R}_i, \mathcal{R}_j)$, which assesses the similarity between neighboring regions $\mathcal{R}_i$ and $\mathcal{R}_j$ by measuring the distance between their region models $d(\mathcal{M}_{\mathcal{R}_i}, \mathcal{M}_{\mathcal{R}_j})$.

The region model $\mathrm{M}_{\mathcal{R}}$ specifies how some region $\mathcal{R}$ of the image is handled within the BPT representation. Therefore, it should be a feature (or a set of features) of this particular region that identifiably describes the properties of this region.

Further, the choice of a particular merging criterion also has an important influence on the BPT representation. As a matter of fact, different merging criteria could lead to significantly different similarity values for the same region model, hence strongly impacting the final BPT structure.

For those reasons, while the BPT representation of an image is intended to be as generic (application independent) as possible, serving only as a support basis for further processing, the definition of the merging procedure should nevertheless take somehow into account the nature and specificities of the handled image. As a matter of fact, relevant definitions for the region model and its associated merging criterion with respect to the processed image should guarantee the consistency of its BPT representation.

First-order parametric region models assume that all regions contained in the BPT representation can be accurately described with a single parameter. In such case, the simplest region model that was initially proposed by Salembier and Garrido (2000) is the mean value/vector within the region:

$$\mathcal{M}_{\mathcal{R}} = \boldsymbol{\mu}_{\mathcal{R}} = (\mu_{\mathcal{R}}(1), \ldots, \mu_{\mathcal{R}}(N)) = \frac{1}{|\mathcal{R}|} \sum_{x \in \mathcal{R}} \mathbf{x} \tag{5}$$

where $\mathbf{x} = \mathcal{I}(x)$ is a $N$-uplet with $N$ being the number of channels in the image. This model is based on the major assumption that pixel values are homogeneously distributed within each band of the image, such that they can be well approximated by their mean. While this may be true for small regions, it is rather too simplistic for large ones (typically regions close to the root). Yet, the simplicity of this region model allows the use of basic merging criteria, such as $\mathcal{L}_p$ norms (recall that the $\mathcal{L}_p$ norm of a vector $\mathbf{x}$ for $p \geq 1$ is defined as $\| \mathbf{x}\|_p = \left(\sum_{i=1}^{N} |x(i(|^p)^{\frac{1}{p}}\right)$, possibly ponderated by the size of the regions being compared (for more details, refer Garrido et al. (1998)).

When dealing with a large number of channels (typically several hundreds in the case of hyperspectral images), the $\mathrm{L}_p$ norm is known to suffer from the curse of dimensionality. Two efficient merging criteria, specifically adapted to hyperspectral imagery, have been proposed by Valero et al. (2010), namely, the spectral angle (SAM) and the spectral information divergence (SID) metrics. The SAM between two regions $\mathcal{R}_i$ and $\mathcal{R}_j$ is defined as the angle between their mean spectrum $\boldsymbol{\mu}_{\mathcal{R}_i}$ and $\boldsymbol{\mu}_{\mathcal{R}_j}$:

$$\mathcal{O}_{SAM}(\mathcal{R}_i, \mathcal{R}_j) = \arccos\left(\frac{\left\langle \boldsymbol{\mu}_{\mathcal{R}_i}; \boldsymbol{\mu}_{\mathcal{R}_j} \right\rangle}{\| \boldsymbol{\mu}_{\mathcal{R}_i}\|_2 \; \| \boldsymbol{\mu}_{\mathcal{R}_j}\|_2}\right) \tag{6}$$

where $\langle \cdot ; \cdot \rangle$ denotes the standard Euclidean dot product. This merging criterion is motivated in hyperspectral imagery by the fact that two spectra describing the same material should have similar shapes, thus a small angle between them in the feature space. The SAM is in addition insensitive to scaling effects (since multiplying a vector by a constant only changes its magnitude). The SID, on the other hand, measures the distance between $\boldsymbol{\mu}_{\mathcal{R}_i}$ and $\boldsymbol{\mu}_{\mathcal{R}_j}$ when interpreted as probability density functions (i.e., when previously normalized to sum to one). A common measure of similarity between such probability density functions is the so-called Kullback–Leibler divergence

$$d_{KL}\left(\boldsymbol{\mu}_{\mathcal{R}_i}, \boldsymbol{\mu}_{\mathcal{R}_j}\right) = \sum_{k=1}^{N} \mu_{\mathcal{R}_i}(k) \log\left(\frac{\mu_{\mathcal{R}_i}(k)}{\mu_{\mathcal{R}_j}(k)}\right) \ . \tag{7}$$

The SID merging criterion is then defined as the symmetric Kullback–Leibler divergence between $\boldsymbol{\mu}_{\mathcal{R}_i}$ and $\boldsymbol{\mu}_{\mathcal{R}_j}$:

$$\mathcal{O}_{SID}(\mathcal{R}_i, \mathcal{R}_j) = d_{KL}\left(\boldsymbol{\mu}_{\mathcal{R}_i}, \boldsymbol{\mu}_{\mathcal{R}_j}\right) + d_{KL}\left(\boldsymbol{\mu}_{\mathcal{R}_j}, \boldsymbol{\mu}_{\mathcal{R}_i}\right) \quad . \tag{8}$$

Still belonging to the family of first-order region models, it is also possible to model a region $\mathcal{R}$ by its covariance matrix $\mathbf{C}_{\mathcal{R}}$ rather than its mean vector $\boldsymbol{\mu}_{\mathcal{R}}$. While this has shown little interest for optical (multispectral or hyperspectral) images up to now, it is however a classical choice to handle synthetic aperture radar (SAR) data (Alonso-González et al., 2012, 2013). Denoting by $\mathbf{k}_x = \left[S_{hh}, \sqrt{2}\,S_{hv}, S_{vv}\right]^T$ the vectorization of the scattering matrix

$$S = \begin{bmatrix} S_{hh} & S_{hv} \\ S_{vh} & S_{vv} \end{bmatrix} \tag{9}$$

for the resolution cell at location $x$ (see, for instance, Cloude and Pottier (1996)), it is known that $\mathbf{k}_x$ follows a 3-D, zero mean, complex Gaussian probability density function (Goodman, 1976)

$$p_k(\mathbf{k}_x) = \frac{1}{\pi^3 \det(\mathbf{C})} \exp\left(-\mathbf{k}_x^H C^{-1} \mathbf{k}_x\right) \tag{10}$$

where $(\cdot)^H$ denote the transconjugate operation, and $\mathbf{C}$ is the Hermitian positive-definite covariance matrix

$$\mathbf{C} = \mathbb{E}\left[\mathbf{k}_x \mathbf{k}_x^H\right] = \begin{bmatrix} \mathbb{E}\left[S_{hh}S_{hh}^H\right] & \sqrt{2}\,\mathbb{E}\left[S_{hh}S_{hv}^H\right] & \mathbb{E}\left[S_{hh}S_{vv}^H\right] \\ \sqrt{2}\,\mathbb{E}\left[S_{hv}S_{hh}^H\right] & 2\mathbb{E}\left[S_{hv}S_{hv}^H\right] & \sqrt{2}\,\mathbb{E}\left[S_{hv}S_{vv}^H\right] \\ \mathbb{E}\left[S_{vv}S_{hh}^H\right] & \sqrt{2}\,\mathbb{E}\left[S_{vv}S_{hv}^H\right] & \mathbb{E}\left[S_{vv}S_{vv}^H\right] \end{bmatrix} \tag{11}$$

with $\mathbb{E}[\cdot]$ being the statistical expectation. While $\mathbf{C}$ is in practice the most valuable radar observable (as it fully describes the probability distribution (Eq. 10) of target vector $\mathbf{k}_x$), it needs to be estimated from the data. Therefore, the statistical expectation is replaced by a spatial averaging, leading to the following region model for all regions:

$$\mathcal{M}_{\mathcal{R}} = \widehat{\mathbf{C}}_{\mathcal{R}} = \frac{1}{|\mathcal{R}|} \sum_{x \in \mathcal{R}} \mathbf{k}_x \mathbf{k}_x^H \quad . \tag{12}$$

The estimated covariance matrix $\hat{\mathbf{C}}_{\mathcal{R}}$ is statistically characterized by a complex Wishart distribution (Lee et al., 1994). Note however that, for very small regions (typically one or a few pixels), the covariance matrix $\hat{\mathbf{C}}_{\mathcal{R}}$ may be singular. In such case, a prefiltering step is applied, as described by Alonso-González et al. (2013).

The use of a covariance matrix following Eq. (12) as a region model $\mathrm{M}_{\mathcal{R}}$ allows to define merging criteria in the space of Hermitian positive-definite matrices. Several metrics were compared in particular by Alonso-González et al. (2012, 2013). For instance, the revised Wishart similarity (RWS) is based on a statistical test assuming that the two regions follow a Wishart probability density function, and one of them is known. Thus, it is not symmetric as it depends on which region probability density function is supposed to be known, and the RWS merging criterion is defined as the following symmetric version:

$$\mathcal{O}_{RWS}(\mathcal{R}_i, \mathcal{R}_j) = \left(\mathrm{tr}\left(\widehat{\mathbf{C}}_{\mathcal{R}_i}^{-1}\widehat{\mathbf{C}}_{\mathcal{R}_j}\right) + \mathrm{tr}\left(\widehat{\mathbf{C}}_{\mathcal{R}_j}^{-1}\widehat{\mathbf{C}}_{\mathcal{R}_i}\right)\right)\left(|\mathcal{R}_i| + |\mathcal{R}_j|\right) \tag{13}$$

where $\mathrm{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. Defined as such, the RWS merging criterion considers all elements of the covariance matrices $\widehat{\mathbf{C}}_{\mathcal{R}_i}$ and $\widehat{\mathbf{C}}_{\mathcal{R}_j}$. It can be further modified to account only for the diagonals entries (setting all off-diagonal elements to 0), yielding the following diagonal revised Wishart similarity (DWS) merging criterion:

$$\mathcal{O}_{DWS}(\mathcal{R}_i, \mathcal{R}_j) = \left(\sum_{k=1}^{3} \frac{\left[\widehat{\mathbf{C}}_{\mathcal{R}_i}\right]_{kk}^2 + \left[\widehat{\mathbf{C}}_{\mathcal{R}_j}\right]_{kk}^2}{\left[\widehat{\mathbf{C}}_{\mathcal{R}_i}\right]_{kk}\left[\widehat{\mathbf{C}}_{\mathcal{R}_j}\right]_{kk}}\right)\left(|\mathcal{R}_i| + |\mathcal{R}_j|\right) \tag{14}$$

with $[\mathbf{A}]_{kk}$ being the $k$-th diagonal entry of matrix $\mathbf{A}$. The DWS merging criterion is significantly simpler to implement than the RWS as it does not necessitate any matrix inversion. In addition, it can be useful to estimate the impact of employing only the diagonal elements of the covariance matrix $\hat{\mathbf{C}}_{\mathcal{R}}$ instead of all its entries. Finally, another possible merging criterion between $\hat{\mathbf{C}}_{\mathcal{R}_i}$ and $\hat{\mathbf{C}}_{\mathcal{R}_j}$ is their so-called geodesic similarity (GS), based on the computation of their geodesic distance in the cone of positive-definite Hermitian matrices (Moakher and Zéraï, 2011):

$$\mathcal{O}_{GS}(\mathcal{R}_i, \mathcal{R}_j) = \left\|\mathbf{log}\left(\widehat{\mathbf{C}}_{\mathcal{R}_i}^{-\frac{1}{2}}\widehat{\mathbf{C}}_{\mathcal{R}_j}\widehat{\mathbf{C}}_{\mathcal{R}_i}^{-\frac{1}{2}}\right)\right\|_F + \log\left(\frac{2|\mathcal{R}_i||\mathcal{R}_j|}{|\mathcal{R}_i| + |\mathcal{R}_j|}\right) \tag{15}$$

where $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}\left(\mathbf{A}\mathbf{A}^H\right)}$ is the Frobenius norm of matrix $\mathbf{A}$, and $\mathbf{log}(\cdot)$ and $\log(\cdot)$ stand for the matrix and natural logarithms, respectively.

First-order models assume that each region can be appropriately described by a single parameter. In spite of its simplicity, this hypothesis has two main limitations. First, it lacks descriptive accuracy for very large regions (those being close to the root of the BPT). Indeed, such regions are likely composed of several smaller coherent regions on their own, and the use of a single modeling parameter will result in an averaging effect. Second, it does not consider the possible variability within each region with respect to the model. For instance, the mean-based region model supposes that the region is spectrally homogeneous and does not take into account its possible spectral variability, which can be of interest for certain applications (Tochon et al., 2015). For that purpose, the nonparametric region model, also called histogram-based region model, was introduced by Valero et al. (2010). This model is defined as a set of $N$ histograms:

$$\mathcal{M}_{\mathcal{R}} = \{\mathcal{H}_{\mathcal{R}}^{1}, \ldots, \mathcal{H}_{\mathcal{R}}^{N}\} \tag{16}$$

where each $\mathcal{H}_{\mathcal{R}}^{k}$ is the empirical spatial distribution of the pixel values within region $\mathcal{R}$ for the $k$-th channel. Each histogram $\mathcal{H}_{\mathcal{R}}^{k} = \{\mathcal{H}_{\mathcal{R}}^{k}(a_1), \ldots, \mathcal{H}_{\mathcal{R}}^{k}(a_{N_{bins}})\}$ is composed of $N_{bins}$ bins $a_p,\ \ p = \{1, \ldots, N_{bins}\}$.

The histogram-based region model allows to define for merging criteria some metrics that measure the similarity between histograms. In particular, Valero et al. (2010, 2013) introduced three histogram-based merging criteria. The Battacharyya distance is based on the Battacharyya coefficient (BC) between the normalized histograms $\mathcal{H}_{\mathcal{R}_i}^{k}$ and $\mathcal{H}_{\mathcal{R}_j}^{k}$ of two adjacent regions $\mathcal{R}_i$ and $\mathcal{R}_j$:

$$\mathcal{O}_{BC}(\mathcal{R}_i, \mathcal{R}_j) = \sum_{k=1}^{N} BC\left(\mathcal{H}_{\mathcal{R}_i}^{k}, \mathcal{H}_{\mathcal{R}_j}^{k}\right) \tag{17}$$

where

$$BC\left(\mathcal{H}_{\mathcal{R}_i}^{k}, \mathcal{H}_{\mathcal{R}_j}^{k}\right) = -\log\left(\sum_{p=1}^{N_{bins}} \sqrt{\mathcal{H}_{\mathcal{R}_i}^{k}(a_p)}\sqrt{\mathcal{H}_{\mathcal{R}_j}^{k}(a_p)}\right) \tag{18}$$

is the BC between $\mathcal{H}_{\mathcal{R}_i}^{k}$ and $\mathcal{H}_{\mathcal{R}_j}^{k}$. If the two histograms perfectly overlap, the argument within the logarithm sums to one; hence, a BC is 0. The Battacharyya distance is a bin-to-bin distance, in the sense that it requires both compared histograms to be perfectly aligned. This can be a disadvantage in a situation where two histograms have a similar profile but are not aligned, and one may want to consider those two histograms to be close to each other. The diffusion distance (DIF), initially introduced by Ling and Okada (2006), was proposed to alleviate this issue. This cross-bin distance is based on the idea that the difference between two histograms

$$d_0^k(a_p) = \mathcal{H}_{\mathcal{R}_i}^{k}(a_p) - \mathcal{H}_{\mathcal{R}_j}^{k}(a_p), \quad p = \{1, \ldots, N_{bin}\} \tag{19}$$

can be viewed as a temperature field, and the corresponding distance between those two histograms is the time needed by this field to reach stability via a heat diffusion process, or equivalently, on the state of the temperature field after a given time. More precisely, starting from $d_0^k$, the diffusion process is simulated by convolving the current temperature distribution with a Gaussian kernel

$$d_m^k(a_p) = \left[d_{m-1}(a_p) \times g_\sigma(a_p)\right]\downarrow_2, \quad m = \{1, \ldots, M\} \tag{20}$$

with $g_\sigma$ standing for the Gaussian kernel with variance $\sigma$, $\downarrow_2$ denotes a downsampling by a factor of 2, and $M$ is the number of convolution layers. The final merging criterion between $\mathcal{R}_1$ and $\mathcal{R}_2$ follows by summing over all $N$ bands the $\mathrm{L}_1$ norm of the $M+1$ layers of temperature fields

$$\mathcal{O}_{DIF}(\mathcal{R}_i, \mathcal{R}_j) = \sum_{k=1}^{N}\sum_{m=0}^{M} \parallel d_m^k \parallel_1 \ . \tag{21}$$

Both Battacharyya and diffusion distances are expressed as a sum of individual band-wise terms and do not stress the importance of a particular band over another. When handling hyperspectral images on the other hand, one could make the most of the strong correlations between bands to remove the redundant information contained in each region model. To that extent, Valero et al. (2013) proposed a merging criterion based on multidimensional scaling (MDS) along with a multivariate analysis of variance (MANOVA) statistical test to investigate whether the principal axes of the region models $\mathcal{M}_{\mathcal{R}_i}$ and $\mathcal{M}_{\mathcal{R}_j}$ of two neighboring regions $\mathcal{R}_i$ and $\mathcal{R}_j$ are correlated. More specifically, let $[\boldsymbol{\Delta}_{\mathcal{R}_i}]_{kl} = \mathcal{O}_{DIF}\left(\mathcal{H}_{\mathcal{R}_i}^{k}, \mathcal{H}_{\mathcal{R}_i}^{l}\right)$ be the $N \times N$ distance matrix associated to $\mathcal{R}_i$ whose entries are the pairwise diffusion distances between the various histograms contained in $\mathcal{M}_{\mathcal{R}_i}$. The MDS procedure (Cox and Cox, 2000) first computes the symmetric matrix $\mathbf{B}_{\mathcal{R}_i} = \mathbf{H}\mathbf{D}_{\mathcal{R}_i}\mathbf{H}$ where $[\mathbf{D}_{\mathcal{R}_i}]_{kl} = -\frac{1}{2}[\boldsymbol{\Delta}_{\mathcal{R}_i}]_{kl}^2$ and $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}1111^T$ is the centering matrix. The eigendecomposition of $\mathbf{B}_{\mathcal{R}_i}$ then yields $\mathbf{B}_{\mathcal{R}_i} = \mathbf{U}_{\mathcal{R}_i}\Lambda_{\mathcal{R}_i}^2\mathbf{U}_{\mathcal{R}_i}^T$ and the principal components of $\mathcal{R}_i$ are defined as $\mathbf{P}_{\mathcal{R}_i} = \mathbf{U}_{\mathcal{R}_i}\Lambda_{\mathcal{R}_i}$. Once the same procedure has been applied to $\mathcal{R}_j$, one can then investigate whether the first $D_s$ principal components (columns) of $\mathbf{P}_{\mathcal{R}_i}$ and $\mathbf{P}_{\mathcal{R}_j}$ are correlated. Calling $\widetilde{\mathbf{P}}_{\mathcal{R}_i}$ and $\widetilde{\mathbf{P}}_{\mathcal{R}_j}$ those two $D_s \times N$ matrices, a multivariate linear regression model is formulated:

$$\widetilde{\mathbf{P}}_{\mathcal{R}_j} = \widetilde{\mathbf{P}}_{\mathcal{R}_i}\boldsymbol{\beta} + \mathbf{e} \tag{22}$$

where $\boldsymbol{\beta}$ is the matrix of parameters containing the regression coefficients and $\mathbf{e}$ is the matrix of errors. A MANOVA statistical test (Anderson, 2000) is applied to investigate whether $\boldsymbol{\beta} = 0$ or not; in other words, whether there is no significant relationship between $\mathcal{R}_i$ and $\mathcal{R}_j$ or not. The test that is classically employed for that purpose is the Wilk's lambda likelihood ratio test, that can be written here as $W(\mathcal{R}_i, \mathcal{R}_j) = \det\left(\mathbf{I}_N - \widetilde{\mathbf{P}}_{\mathcal{R}_j}^T\widetilde{\mathbf{P}}_{\mathcal{R}_i}\widetilde{\mathbf{P}}_{\mathcal{R}_i}^T\widetilde{\mathbf{P}}_{\mathcal{R}_j}\right)$. $0 \le W(\mathcal{R}_i, \mathcal{R}_j) \le 1$, with $W(\mathcal{R}_i, \mathcal{R}_j) = 0$ meaning that the hypothesis $\boldsymbol{\beta} = 0$ is false, and thus $\mathcal{R}_i$ and $\mathcal{R}_j$ are highly correlated. Therefore, the MDS merging criterion can be defined as

$$\mathcal{O}_{MDS}(\mathcal{R}_i, \mathcal{R}_j) = W(\mathcal{R}_i, \mathcal{R}_j) = \det\left(\mathbf{I}_N - \overset{\sim T}{\mathbf{P}}_{\mathcal{R}_j}\widetilde{\mathbf{P}}_{\mathcal{R}_i}\widetilde{\mathbf{P}}_{\mathcal{R}_i}^T\widetilde{\mathbf{P}}_{\mathcal{R}_j}\right) \ . \tag{23}$$

The reader is referred to Valero et al. (2013) for practical implementation details.

There is no clear answer to the question related to which couple region model/merging criterion would yield the most consistent BPT representation, as it depends upon not only the nature of the handle data but also on the user's interest and expectation with respect to this data (the mean-based region model would be suitable, for instance, if one wants to retrieve some spectrally homogeneous regions, but would not be a good candidate to emphasis the intraregion spectral variability).

In addition, the merging procedure can lead to small and insignificant regions remaining in the last merging iterations of the construction when the region size does not intervene in the definition of the merging criterion. To overcome this issue, Calderero and Marques (2010) proposed to use a priority rule: all regions whose size is less than a given threshold (typically set to 15%) of the mean size of the regions standing in the current merging iteration are given the merging priority, regardless of their distance with respect to their neighbors.

### 2.05.3.2 Processing of the BPT

Contrary to its construction, which should be (as much as possible) done regardless of the application, further processing of the BTP must be adapted to the pursued goal. Nevertheless, a typical BPT processing can be decomposed into two steps, which are illustrated in Fig. 8A:

1. The population of the tree: During this stage, the user defines some features or attributes, which are then evaluated for each region $\mathcal{R}$. Those computed features are stored in a set $\Omega_{\mathcal{R}}$: the tree is "populated."
2. The decision step: It subsequently evaluates, given a decision rule, if each node should be retained or discarded according to its previously computed set of features. This decision step involves a decision function F that is applied on each region to take the decision whether to keep this node or not. $\mathcal{R}_{\mathcal{F}}$ is the set of nodes of the BPT structure that have been retained by the decision function $\mathcal{F}$.

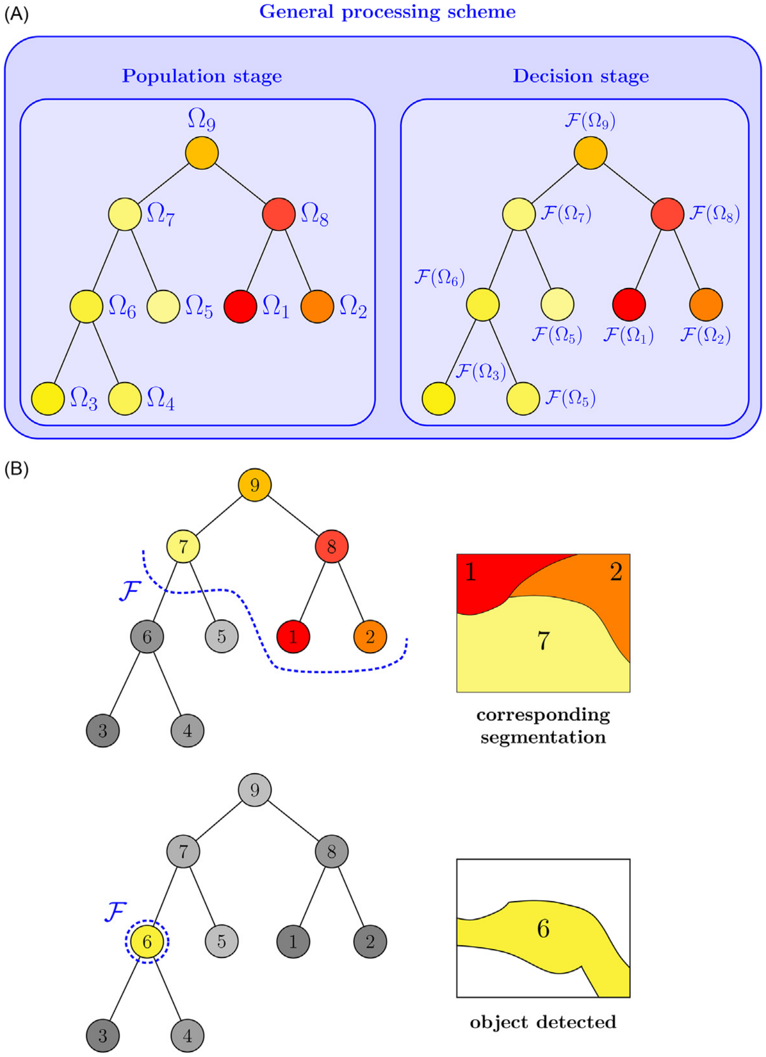


**Fig. 8** (A) General scheme of a BPT processing step and (B) examples of BPT processing results.

The BPT processing procedures can be divided into two classes, depending on the structure of $\mathcal{R}_\mathrm{F}$. Either it does not constitute a partition of $E$ (the spatial support of the image I on which is built the BPT representation), in which case, the processing is termed node selection/identification, or $\mathcal{R}_\mathrm{F}$ forms a partition of the space $E$, in which case, it is called a pruning. A typical example for the former category is the object detection process, that is, the identification in the tree of regions that match some predefined criteria (see, for instance, Salerno et al. (2004 and Vilaplana et al. (2008) and section "**Object Recognition in Urban Hyperspectral Scene**"). For the pruning case, the output $\mathcal{R}_\mathrm{F}$ is also called a cut $\pi \in \Pi_E(H)$, where $\Pi_E(H)$ denotes the set of all cuts of the BPT $H$. Equivalently, a cut $\pi$ can be defined as a partition of $E$ whose all regions belong to the BPT $H$. To obtain a cut, one can work directly in $\Pi_E(H)$ (which has a lattice structure, easily allowing the use of operations such as the refinement suprema and infima for instance). The alternative way to proceed is to analyze the various branches of the BPT and select one node in each, with the additional constraint that each leaf node should have one and only one selected node among its ancestors. Fulfilling this condition ensures to produce some partition of $E$, hence a cut $\pi \in \Pi_E(H)$. From a graphical point of view, a cut is a path that intersects each branch of the BPT representation at most once, and the regions composing the cut are those located directly above it, defining the leaves of the pruned tree. An example of pruning and node selection processes is displayed in **Fig. 8B**. As a simple example to illustrate this processing chain, consider an application where one wants to smooth an image by removing small and inhomogeneous regions. A possible strategy to achieve such goal would be to filter out from the BPT representation all regions whose size $|\mathcal{R}|$ is below a predefined threshold $\delta$. As a consequence, the attribute that should be retrieved for each region is its area, defining a feature set $\Omega_\mathcal{R} = \{|\mathcal{R}|\}$. The decision rule being to remove a node if its size is below the threshold, the decision function becomes $\mathcal{F}(\Omega_\mathcal{R}) \geq \delta$, and all nodes, which do not satisfy this decision, are removed from the BPT, producing a pruned tree where all new leaves have a size greater than or equal to the size threshold $\delta$.

This decision function, applied on the region area, is a particular case of increasing decision: a decision is said to be increasing if $\mathcal{R}_1 \subseteq \mathcal{R}_2 \Rightarrow \mathcal{F}(\Omega_{\mathcal{R}_1}) \leq \mathcal{F}(\Omega_{\mathcal{R}_2})$. In such a case, if a node has to be retained, then so have to be all its ancestors. Conversely, if it is decided that a node should be discarded, it is also the case for all its descendants. When the decision is not increasing, then some more sophisticated strategies have to be used, such as the minimum, maximum, or Viterbi decision rules. The minimum decision rule states that a region is preserved if and only if all its ancestors also have to be preserved. The maximum decision is the opposite, namely, a node is removed if and only if all its descendants also have to be removed. The Viterbi decision strategy, on the other hand, associates to each node a cost reflecting how much impact it would have to change the decision for this node (for instance, how much would it cost to remove a node that was decided to be retained). It then tries to minimize this cost function in order to make the decision function increasing. As an example, if it has been decided that all nodes in a whole branch should be retained except for one, it is less costly to take the decision to retain all nodes of the branch (so inverting the decision for only one node) rather than removing all nodes (see Salembier and Garrido (2000) for more details).

## 2.05.4 Applications of the BPT Representation in Remote Sensing Data Analysis

This section presents some typical applications of the BPT representation for various remote sensing data analysis tasks. Each subsection features the detailed review of one or several related and recent works on the topic and exhibits the results detailed in the referred references.

### 2.05.4.1 Hyperspectral Image Segmentation

Image segmentation, that is, the division of an image into a set of coherent and nonoverlapping regions, is an important operation in image processing. As a matter of fact, pixels, despite being the elementary bricks of an image, cannot be semantically interpreted and a major challenge of pixel-based processing methods is to somehow incorporate some information related to the spatial context within the image. On the other hand, region-based processing approaches can take advantage of the knowledge brought by the spatial structure of the regions and thus allow for high-level processing (such as object recognition). Image segmentation is however an ill-posed problem, as a given image may be segmented at various levels of details, depending on the object of the objects of interest it contains. Fortunately, those objects are very often organized in a hierarchical manner, making hierarchical representation very suitable for segmentation purposes: one can build a hierarchical representation of the image, and an adequate segmentation can then be chosen by browsing the tree structure at the desired scale. The selection of a proper scale to extract a segmentation from the tree structure is however not straightforward. A possible solution is to address it through an energetic minimization framework (Tarabalka et al., 2010a). The definition of an objective function (i.e., the energy function) related to the underlying application allows to associate some numerical criterion to each partition that can be extracted from the hierarchical representation, and one then seeks for the partition whose energy is minimal. This strategy has been successfully investigated for the segmentation of hyperspectral images by Valero (2011) and Veganzones et al. (2014b), which are both reviewed in the following.

#### *2.05.4.1.1 Energy minimization over a hierarchical structure*

The energy minimization framework for image segmentation purposes has already been largely addressed in the literature. In their famous paper, Mumford and Shah (1989) propose to carry out the segmentation process with a variational scheme. They define the best approximation of a given image $\mathcal{I}_0$ as the minimizer of an objective function achieving a trade-off among a misfit term between $\mathcal{I}_0$ and its piece-wise smooth approximation $\mathcal{I}$, a term enforcing smoothness for $\mathcal{I}$ and a term promoting simplicity by regularizing the total length of the boundaries in the segmentation. An alternative well-known approach was presented by Boykov et al. (2001), where segmentation is handled as a labeling problem and tackled using graph-cuts, with a minimized energy written as the sum of a data fitting term and a regularization term. Markov random fields (Li, 2009) are another instance of energetic segmentation method, where the maximum a posteriori estimation of the best data labeling (i.e., segmentation) is reached through an energetic minimization scheme. For a general point a view, an energy function can be viewed as a mapping $\mathscr{E}: \Pi_E \to \mathbb{R}^+$ from the set of partitions of $E$ to real nonnegative numbers (with the intuition that the lower the energy, the better the corresponding partition). However, in many cases, the energy function is first evaluated over the regions composing the partition, which are then somehow assembled into the final energy of the partition. Therefore, the following general definition of energy functions is rather considered:

$$\mathscr{E}(\pi) = \mathop{\mathfrak{D}}_{\mathcal{R}_i \in \pi} \mathscr{E}(\mathcal{R}_i) \tag{24}$$

where $\mathscr{E}: \mathcal{P}(E) \to \mathbb{R}^+$ maps any region $\mathcal{R} \subseteq E$ to the set of real nonnegative numbers, $\mathcal{P}(E)$ is the set of all subsets (i.e., possible regions) of $E$, and $\mathfrak{D}$ is a particular composition rule to express the energy of the partition $\pi$ with respect to the energies of the regions $\mathcal{R}_i \in \pi$ composing this partition. In this formalism, $\mathfrak{D}$ can be arbitrary, but the most common case is to express the energy of a partition as the sum of the energies of its regions (in other words, $\mathfrak{D} \equiv \sum$), the energy being called separable in such occurrence. However, finding the minimizer $\pi^\star = \text{argmin}_{\pi \in \Pi_E} \mathscr{E}(\pi)$ of the energy function is not straightforward, either because the minimization problem is nonconvex and the convergence to a global optimum is not guaranteed, or because this global optimum cannot be reached in an acceptable computational time. These limitations find their source in the nature of the space of all possible partitions $\Pi_E$ of the set $E$, namely, its huge cardinality and its lack of structure. Conducting such energy minimization in the space of cuts $\Pi_E(H)$ of a hierarchical representation $H$, in other words $\pi^\star = \text{argmin}_{\pi \in \Pi_E(H)} \mathscr{E}(\pi)$, allows to alleviate those issues.

Let $H$ be some hierarchy of partitions (not necessarily a BPT representation) and let $H(\mathcal{R})$ be the subhierarchy of $H$ rooted at $\mathcal{R}$, and let $\pi(\mathcal{R})$ be a cut of $H(\mathcal{R})$. $\pi(\mathcal{R})$ is called a partial partition of $\mathcal{R}$. Denoting $\pi^\star(\mathcal{R}) = \text{argmin}_{\pi(\mathcal{R}) \in \Pi_E(H(\mathcal{R}))} \mathscr{E}(\pi(\mathcal{R}))$ the partial partition of $\mathcal{R}$ whose energy is minimal, and $\mathscr{E}^\star(\mathcal{R}) = \mathscr{E}(\pi^\star(\mathcal{R}))$ standing for this optimal energy, Guigues et al. (2006) showed that the following Bellman's dynamic program was holding for all regions $\mathcal{R} \in H$:

$$\mathscr{E}^\star(\mathcal{R}) = \min\left\{ \mathscr{E}(\mathcal{R}), \sum_{r \in \mathrm{C}(\mathcal{R})} \mathscr{E}^\star(r) \right\} \tag{25}$$

$$\pi^\star(\mathcal{R}) = \begin{cases} \{\mathcal{R}\} & \text{if } \mathscr{E}(\mathcal{R}) \le \sum\limits_{r \in \mathrm{C}(\mathcal{R})} \mathscr{E}^\star(r) \\ \coprod\limits_{r \in \mathrm{C}(\mathcal{R})} \pi^\star(r) & \text{otherwise} \end{cases} \tag{26}$$

with $\sqcup$ denoting disjoint union between regions (concatenation). Eqs. (25) and (26) mean that the optimal energy of any region $\mathcal{R} \in H$ is given by comparing the proper energy $\mathscr{E}(\mathcal{R})$ of the region against the sum of the optimal energies of its children, and by picking the smallest of the two. The optimal cut of $\mathcal{R}$ is then given either by itself $\{\mathcal{R}\}$ or by the disjoint union of the optimal cuts of its children. This dynamic program procedure is illustrated by Fig. 9: the optimal cut $\pi^\star(\mathcal{R})$ of $\mathcal{R}$ (in red) is either given by $\{\mathcal{R}\}$ or by $\pi^\star(\mathcal{S}_1) \sqcup \pi^\star(\mathcal{S}_2)$, depending on which has the lowest energy.

In practice, it is possible to obtain the optimal cut of $H$ by applying Eqs. (25) and (26) over each region of the hierarchy, scanned in an ascending pass. The optimal cut $\pi^\star$ of $H$ is given by the one of the root note. The dynamic program procedure illustrates that the optimal cut is obtained by taking advantage of the inclusion relationship holding on the regions of a hierarchy, thus emphasizing the benefit of conducting the energy minimization operation over such hierarchical structures instead of the unconstrained set of partitions $\Pi_E$. This result has been lately generalized by Kiran and Serra (2014) to wider classes of energies (that are not necessarily composed by a sum).

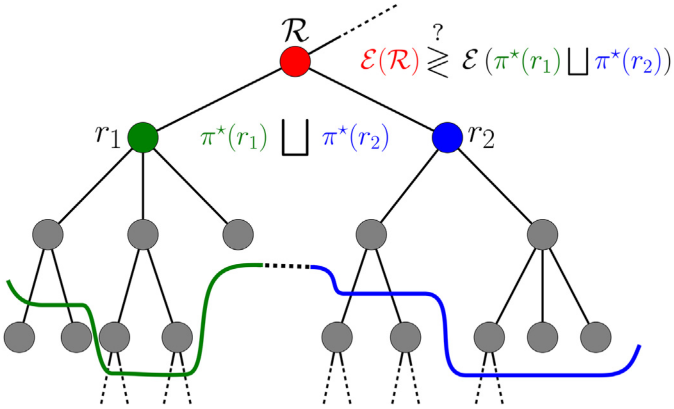

$\mathcal{R}$
$\mathcal{E}(\mathcal{R}) \overset{?}{\gtrless} \mathcal{E}\left(\pi^\star(r_1) \bigsqcup \pi^\star(r_2)\right)$
$r_1$
$\pi^\star(r_1) \bigsqcup \pi^\star(r_2)$
$r_2$

Fig. 9 Illustration of a Bellman's dynamic program step to retrieve the optimal cut of a hierarchy of partition.

Energies in the literature often depend in practice on a positive real-valued parameter $\lambda$ that acts as a trade-off between a data fitting term $\mathscr{E}_\varphi$ (i.e., leading to oversegmentation) and a regularization term $\mathscr{E}_\rho$ (i.e., promoting simplicity, thus favoring undersegmentation) , so that the energy of a partition $\pi$ can be written as

$$\mathscr{E}_\lambda(\pi) = \sum_{\mathcal{R}\in\pi} \left(\mathscr{E}_\varphi(\mathcal{R}) + \lambda\mathscr{E}_\rho(\mathcal{R})\right) \tag{27}$$

In that context, there is no longer one optimal cut $\pi^\star$ for a given hierarchy $H$ and some energy $\mathscr{E}_\lambda$ parameterized by $\lambda$, but rather a family of them $\{\pi^\star_\lambda\}$ in turn indexed by this parameter $\lambda$. Guigues et al. (2006) showed that under the assumption of subadditivity for the regularization term $\mathscr{E}_\rho$ (i.e., for any two disjoint regions $\mathcal{R}_1$ and $\mathcal{R}_2$, $\mathscr{E}_\rho(\mathcal{R}_1 \cup \mathcal{R}_2) \leq \mathscr{E}_\rho(\mathcal{R}_1) + \mathscr{E}_\rho(\mathcal{R}_2)$), then the family of optimal cuts $\{\pi^\star_\lambda\}_{\lambda\in\mathbb{R}^+}$ could be ordered by refinement, that is:

$$\forall\lambda_1, \lambda_2, 0 \leq \lambda_1 \leq \lambda_2 \Rightarrow \pi^\star_{\lambda_1} \leq \pi^\star_{\lambda_2} \ . \tag{28}$$

This property (which has also been extended to wider classes of energy function by Kiran and Serra (2014)) notably allows to transform some hierarchy $H$ into its persistent version $H^\star$, composed of all the optimal cuts $\pi^\star_\lambda$ of $H$ when $\lambda$ spans $\mathbb{R}^+$. The important consequence of this property in practice is that it avoids the delicate tuning of $\lambda$ by instantaneously giving access to all optimal cuts of $H$. One can then browse $H^\star$ for the optimal cut whose number of regions is close to the requested one. This strategy is much convenient to operate as the desired number of regions is an easier input than the associated numerical value of the regularization parameter $\lambda$. The reader is referred to Guigues et al. (2006) for more practical implementation details.

#### 2.05.4.1.2 Application to hyperspectral image segmentation

We now present the application of the previously described hierarchical energy minimization framework to the segmentation of hyperspectral images. The task of accurately segmenting hyperspectral images has mostly been addressed in the literature for classification purposes (Bilgin et al., 2011; Li et al., 2012; Noyel et al., 2007), where a first preliminary oversegmentation is often used as an input for a subsequent classification stage, the obtained classification map serving as the segmentation result. The two limitations of this strategy are that it only allows to define regions of interest that match the classification classes (which may not necessarily be the objective for which the hyperspectral image is being segmented) and it does not exploit the multiscale nature of the image.

The energy minimization framework is convenient because of its flexibility: the defined energy function is a mathematical tool that reflects the desired objective for the segmentation. Therefore, two different energies can lead to two very distinct optimal segmentations. Therefore, the energy must be carefully designed to not only meet the property formulated in the previous section "**Energy minimization over a hierarchical structure**," but also to ensure that its optimal cut achieves what is expected for the underlying application. To obtain homogeneous regions with relatively simple contours, a common choice is to use the piecewise constant Mumford Shah energy, which can be written as

$$\mathscr{E}^{MS}_\lambda(\pi) = \sum_{\mathcal{R}\in\pi} \left(\mathscr{E}_\varphi(\mathcal{R}) + \frac{\lambda}{2}\left|\partial\mathcal{R}\right|\right) \tag{29}$$

where $\mathscr{E}_\varphi(\mathcal{R}) = \sum_{x\in\mathcal{R}} \parallel \mathbf{x} - \boldsymbol{\mu}_\mathcal{R}\|^2_2$ is the data fitting term, measuring the error between the pixel spectra $\mathbf{x}$ in the region $\mathcal{R}$ and their average $\boldsymbol{\mu}_\mathcal{R}$, and the regularization term $\mathscr{E}_\rho$ is half the length of the boundary $|\partial\mathcal{R}|$ of $\mathcal{R}$ (Guigues et al., 2006). The $L_2$ norm is however not well suited to work with hyperspectral data due to the large dimensionality of the pixel spectra. In addition, the regularization term being the total length of all boundaries in the segmentation map makes the parameter $\lambda$ not really intuitive to tune. As an alternative, Valero (2011) proposed a novel energy definition to segment hyperspectral images into spectrally homogeneous regions:

$$\begin{aligned}\underset{\mathscr{E}_\lambda}{\sum} \mathcal{SID}(\pi) &= \sum_{\pi\in\mathcal{R}} \left(\mathscr{E}_\varphi(\mathcal{R}) + \lambda\right) \\ &= \sum_{\pi\in\mathcal{R}} \mathscr{E}_\varphi(\mathcal{R}) + \lambda|\pi|\end{aligned} \tag{30}$$

where $|\pi|$ is the number of individual regions in the partition $\pi$ and where the data fitting term $\mathscr{E}_\varphi(\mathcal{R})$ is defined as

$$\mathscr{E}_\varphi(\mathcal{R}) = \sum_{x\in\mathcal{R}} \mathcal{O}_{SID}(\mathbf{x}, \boldsymbol{\mu}_\mathcal{R}) + \begin{cases} 0 \ \text{ if}|\mathcal{R}_l|\text{and}|\mathcal{R}_r| \leq \tau, \\ \sum\limits_{x\in\mathcal{R}_l} \mathcal{O}_{SID}\left(\mathbf{x}, \boldsymbol{\mu}_{\mathcal{R}_r}\right) + \sum\limits_{x\in\mathcal{R}_r} \mathcal{O}_{SID}\left(\mathbf{x}, \boldsymbol{\mu}_{\mathcal{R}_l}\right)\text{o/w} \end{cases} \tag{31}$$

where $\mathcal{O}_{SID}$ denotes the SID measure, as defined by (Eq. 8), $\boldsymbol{\mu}_\mathcal{R}, \boldsymbol{\mu}_{\mathcal{R}_l}$ and $\boldsymbol{\mu}_{\mathcal{R}_r}$ are the mean spectra of regions $\mathcal{R}$, $\mathcal{R}_l$, and $\mathcal{R}_r$, the latter two being the left and right children of $\mathcal{R}$. The first term of (Eq. 31) measures the error committed when replacing all pixel spectra in region $\mathcal{R}$ by their mean value $\boldsymbol{\mu}_\mathcal{R}$, thus penalizing spectrally inhomogeneous regions. The second term evaluates the error of replacing each pixel spectrum of the child region $\mathcal{R}_l$ by the mean spectrum of its sibling $\mathcal{R}_r$ and vice versa, in order to regularize the case where the region $\mathcal{R}$ has a child, which is much larger than the other one (the contribution of the small child being negligible in the first error term, even if spectrally different from $\boldsymbol{\mu}_\mathcal{R}$). In practice, the second term is added to $\mathscr{E}_\varphi(\mathcal{R})$ if the two children have a size greater than a predefined threshold $\tau$ (set to 3 pixels in Valero (2011)) in order to make this estimation reliable.

This concept of defining a suited energy function for a given task can be adapted to various remote sensing tasks, such as spectral unmixing. The goal of spectral unmixing is to retrieve the spectrally pure constituents, called endmembers, in each pixel spectrum, as well as their respective proportions, termed fractional abundances (Bioucas-Dias et al., 2012). Often, the linear mixing model is assumed, allowing to write this decomposition as a linear combination $\mathbf{x} = \sum_{i=1}^{p} \varphi_i \mathbf{e}_i + \boldsymbol{\eta}$ between the endmembers $\mathbf{e}_i, i = 1, \ldots, p$ (where $p$ can be manually set or estimated beforehand) and their associated fractional abundances $\varphi_i$ subject to positivity and sum-to-one constraints. However, as this decomposition is not perfect, the unmixing error is classically measured as the root mean square error (RMSE) between the true pixel spectrum $\mathbf{x}$ and the reconstructed one $\widehat{\mathbf{x}}$:

$$\varepsilon(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{1}{\sqrt{N}} \| \mathbf{x} - \widehat{\mathbf{x}} \|_2 \tag{32}$$

where $\widehat{\mathbf{x}} = \sum_{i=1}^{p} \varphi_i \mathbf{e}_i$. Note that the endmembers and fractional abundances are estimated by considering all pixels in the image in the usual scenario. In this context, Veganzones et al. (2014b) proposed several energy functions to provide a segmentation map with minimal unmixing reconstruction error, thus optimal with respect to the spectral unmixing operation. They propose, for instance, the following definition:

$$\begin{aligned} \sum_{\mathscr{E}_\lambda} \max(\pi) &= \sum_{\mathcal{R}\in\pi} \left( \frac{|\mathcal{R}|}{|E|} \max_{x\in\mathcal{R}} \varepsilon_{\mathcal{R}}(\mathbf{x}, \widehat{\mathbf{x}}) + \lambda \right) \\ &= \frac{1}{|E|} \sum_{\mathcal{R}\in\pi} |\mathcal{R}| \max_{x\in\mathcal{R}} \varepsilon_{\mathcal{R}}(\mathbf{x}, \widehat{\mathbf{x}}) + \lambda|\pi| \end{aligned} \tag{33}$$

where $\epsilon_{\mathcal{R}}(\mathbf{x}, \widehat{\mathbf{x}})$ denotes the RMSE of pixel spectrum $\mathbf{x}$ with respect to the endmembers and fractional abundances that have been estimated using only the spectral information available within region $\mathcal{R}$ it belongs to, and no longer with respect to the whole image. The data fitting term of energy (Eq. 33) can be seen as the weighted average of the maximum RMSE of the regions in the partition, and the regularization term is again the number of regions in the partition. The objective of this energy function is thus to provide some segmentation where each region has the lowest possible maximum reconstruction error.

The used data for segmentation is a crop of Hyperspectral Digital Imagery Collection Experiment (HYDICE) Washington, DC, mall composed of 200 × 200 pixels and comprising 191 bands. **Fig. 10A** shows an color composition of the scene, which features an urban environment with several buildings and roads, as well as some lawns and trees. A very-high version of the scene is depicted by **Fig. 10B** (https://goo.gl/maps/8Tr91fnwcdo). To demonstrate the impact of the choice of a particular energy on the segmentation results, the same setting is used for the BPT construction in all three cases (the Mumford–Shah energy (Eq. 29), the energy proposed by Valero (2011) (Eq. 30) and the one introduced by Veganzones et al. (2014b) (Eq. 33)). The initial partition is obtained by a mean shift clustering procedure (Comaniciu and Meer, 2002) applied over the RGB composition, and leads to 2339 initial regions. The BPT is built using the mean spectrum region model (Eq. 5) and the spectral angle (Eq. 6) merging criterion.

This BPT is then pruned to achieve the optimal segmentation for all three investigated energy functions. Note that this search for an optimal cut well falls into the processing scheme described in section "**Processing of the BPT**": the population step of the tree structure is achieved by evaluating for each region $\mathcal{R}$ its proper energy $\mathscr{E}_\lambda(\mathcal{R}) = \mathscr{E}_\rho(\mathcal{R}) + \lambda \mathscr{E}_\varphi(\mathcal{R})$. The decision function F of the following decision step is realized by application of the dynamic program Eqs. (25) and (26), which decide whether the region $\mathcal{R}$ belongs to the optimal cut $\pi_\lambda^\star$ or not.

The obtained results are presented in **Fig. 11**. The first row displays the optimal segmentations (where each region has been filled by its mean color) for the Mumford–Shah energy (**Fig. 11A**), the energy proposed by Valero (2011) (**Fig. 11B**), and the

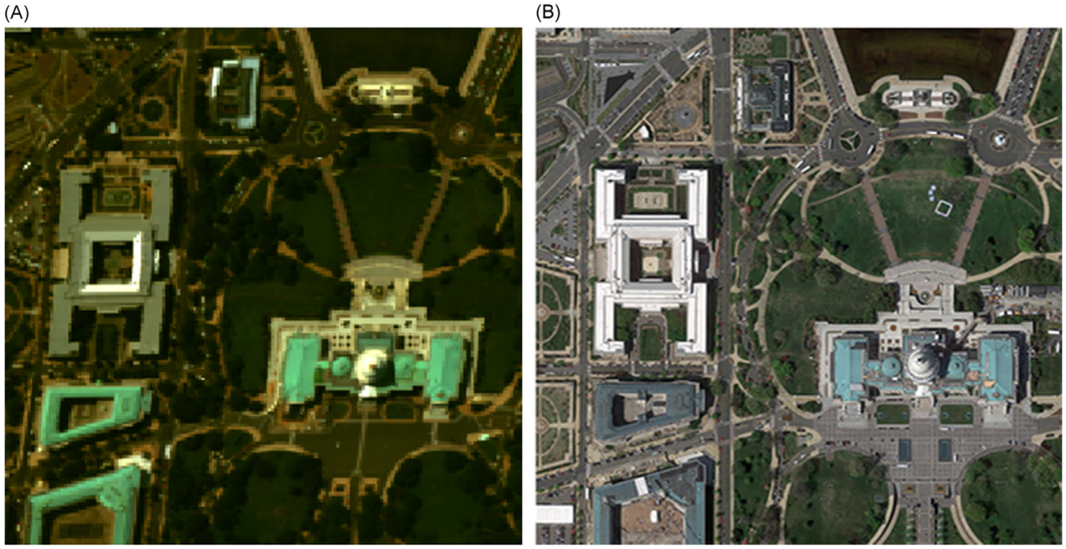


**Fig. 10** (A) 200 × 200 crop of the HYDICE Washington, DC, mall hyperspectral image and (B) very-high resolution version of the same scene.

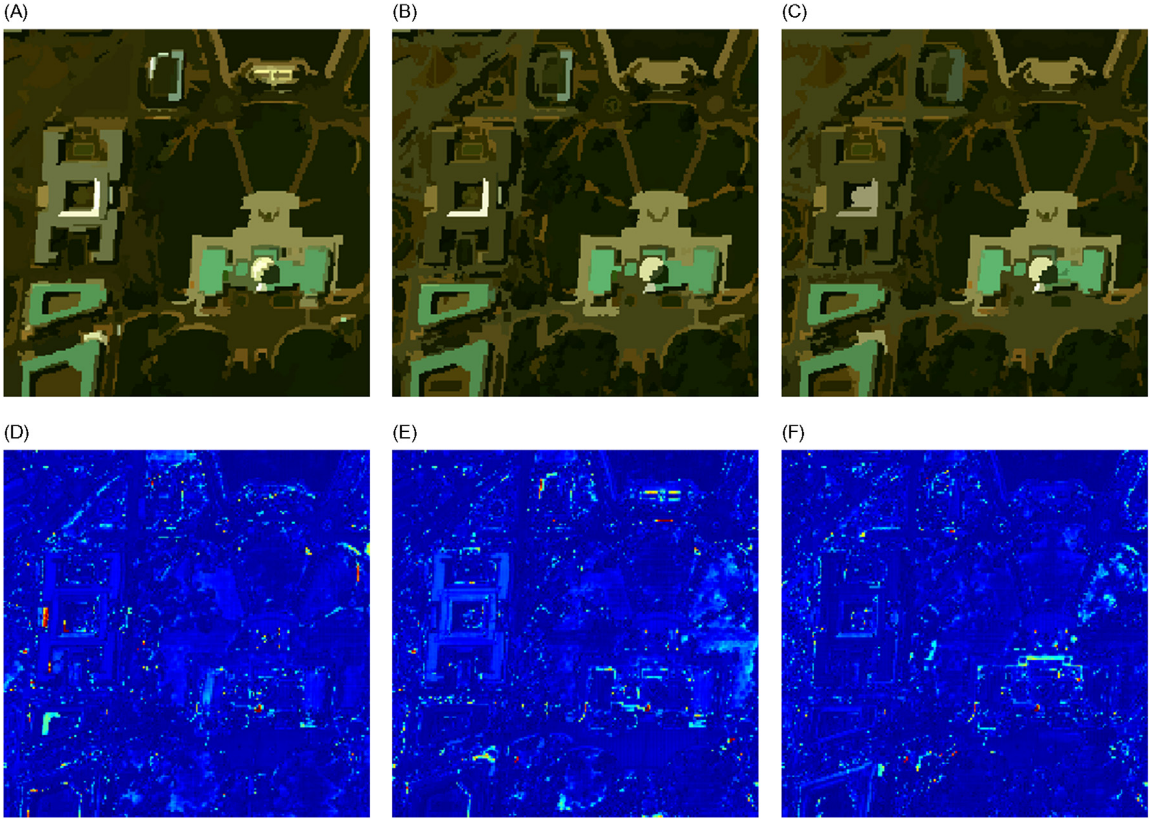


**Fig. 11** *Top row*: obtained optimal segmentation for (A) the Mumford–Shah functional and the energies proposed by (B) Valero (2011) and (C) Veganzones et al. (2014b). *Bottom row*: associated reconstruction error maps (D–F), where the scale has been saturated between 0 (in *blue*) and 0.04 (in *red*).

energy introduced by Veganzones et al. (2014b) (**Fig. 11C**). Those are composed of 400, 401, and 436 regions, respectively, when 400 regions were requested as input (when no optimal cut has the exact number of expected regions, the selected one is as close as possible from the target number). The second row presents the corresponding unmixing reconstruction errors of those three optimal segmentations. One can remark several differences regarding the structure and content of the various regions composing the segmentations in **Fig. 11A** and **B**, whose underlying minimized energies both aim at producing partitions with homogeneous regions. Overall, the regions in **Fig. 11A** are more heterogeneous in terms of size and less complex in terms of shape. This is due to the nature of the regularizing term in the Mumford–Shah energy (Eq. 29), which penalizes regions with complicated boundaries and thus favors either large regions with simple shapes (such as the lawns of the building) or very small regions (such as the various gradations in the cupola of the US capitol building for instance). In the meantime, the optimal segmentation for the energy of Valero (2011) presents more thin details such as the road network on the top left corner of the image, or the various walkways in the lawns. This energy is also less sensitive to gradations as the SID measure is not affected by scaling factors (since the spectra are normalized to compute the measure). Regarding the optimal segmentation with respect to the energy of Veganzones et al. (2014b) (**Fig. 11C**), its differences are rather minor from a visual point of view with respect to the other two strategies. Analyzing the reconstruction error maps on the other hand (bottom row of **Fig. 11**), one can see that the errors appear smaller for this approach (for the gray building on the left, for instance). This is confirmed by **Table 1**, which presents the overall average and maximum reconstruction errors for all three approaches. While the average error values remain comparable, it is worth remarking that the overall maximum errors have been notably decreased by the energy of Veganzones et al. (2014b), which was exactly its purpose. Finally, it should be noted that, as reported in Veganzones et al. (2014b), the used setting for the BPT construction is not the optimal one for unmixing purposes. This also explains why all three segmentations overall look similar in terms of regions and contents, since the energy minimization procedure was conducted each time on the exact same BPT structure.

As a conclusion for this first application of BPT representations for the analysis of remote sensing data, the energy minimization framework is really convenient for segmentation purposes. Its high flexibility, coupled with strong theoretical properties, makes it suited for a wide range of applications where segmentation has a key role to play. For instance, the hierarchical energy minimization procedure, coupled with the unmixing-based functional (Eq. 33), has recently been used as an input by Veganzones et al. (2016) for hyperspectral superresolution purposes, as well as for local spectral unmixing considerations in Tochon et al. (2016).

**Table 1** Average and maximum spectral reconstruction errors for all three investigated approaches

| | *Mumford-Shah* | Valero (2011) | Veganzones et al. (2014b) |
|---|---|---|---|
| mean RMSE | 0.0028 | 0.0029 | **0.0026** |
| max RMSE | 0.0776 | 0.0762 | **0.0523** |

Lowest values are in bold.

### 2.05.4.2 Hyperspectral Image Classification

The classification task aims, given an image, at assigning a label to each pixel, such that pixels sharing the same label (which define a class) have some common properties. In the case of hyperspectral images, classes very often are designed to group pixels made of the same materials (such as grass, soil, concrete, water, and so on), such that the resulting classification map gives a precise insight on the composition of the scene. This ability to identify the materials and their extent on ground is very valuable for land use and land cover mapping, for instance, which is why the classification task is one of the most studied hyperspectral processing applications.

Processing each hyperspectral pixel based only on its spectral content, without taking into account the information provided by its neighboring area, leads to the well-known salt-and-pepper effect: pixel-wise classification techniques suffer due to this phenomenon. To remedy this issue, several spectral–spatial classification methods have been introduced in the past few years (see, for instance, Fauvel et al. (2012); Li et al. (2012); Tarabalka et al. (2009). A review of such methods can also be found in Fauvel et al. (2013). Nevertheless, those approaches only consider the image at its global scale and thus do not take into account the hierarchical dependencies between the various objects composing the scene. For this reason, a hyperspectral classification method relying on the BPT representation was presented by Valero et al. (2013), and this section is devoted to its review.

Since a classification map is a partition of the image (as it divides the image into a set of nonoverlapping but possible nonconnected classes), it can be achieved by a BPT pruning operation and is therefore conducted following the procedure exposed in section "**Processing of the BPT**." More specifically, let $\mathcal{C}_i, i = 1, \ldots, C$ be the $C$ class labels to classify the input data. The most important information to decide whether a node $\mathcal{R}$ in the BPT should be retained or discarded during the pruning operation is its class probability distribution $\mathcal{P}_{\mathcal{R}} = (\mathcal{P}_{\mathcal{R}}(\mathcal{C}_1), \ldots, \mathcal{P}_{\mathcal{R}}(\mathcal{C}_C))$, where $\mathcal{P}_{\mathcal{R}}(\mathcal{C}_i) = \text{proba}(\mathcal{R}$ is classified in $\mathcal{C}_i)$. Therefore, each node in the BPT has its class probability distribution computed during the population step. This can be easily done in a supervised way by using a multiclass classifier. For that purpose, Valero et al. (2013) proposed to use the well-known support vector machine (SVM) (Cortes and Vapnik, 1995), as they have already proved to be well suited for high-dimensional classification of hyperspectral images (Bruzzone et al., 2006), with a standard Gaussian kernel whose parameters are computed with the classical cross-validation strategy. Having built the BPT from the pixel level, the SVM training step is done by selecting some leaf nodes, hence single spectra, for which the true class membership is known (based on some reference ground truth data). Once the kernel function is constructed, it is used to classify all regions $\mathcal{R}$ in the BPT by assigning them their class probability distribution $\mathcal{P}_{\mathcal{R}}$. During the classification stage, each node is modeled by its mean spectrum, as the classifier input is a single spectrum.

In the following, the misclassification (MC) rate of each node is evaluated based on its class probability distribution. The MC rate can be seen as the probability of classifying the region $\mathcal{R}$ into the wrong class, and can be expressed as

$$\mathcal{M}\mathrm{C}_{\mathcal{R}} = 1 - \max_{i=1,\ldots,C} \mathcal{P}_{\mathcal{R}}(\mathcal{C}_i) \ . \tag{34}$$

However, the MC rate as defined by (Eq. 34) suffers from two important issues. Assuming that node $\mathcal{R}$ has $\mathcal{R}_l$ and $\mathcal{R}_c$ for children, with one (say $\mathcal{R}_l$) being significantly bigger than the other. If $\mathcal{R}_l$ belongs to class $\mathcal{C}_i$ and $\mathcal{R}_r$ to class $\mathcal{C}_j$, their union being $\mathcal{R}$ will belong to $\mathcal{C}_i$ since $\mathcal{R}_l$ will have a higher contribution to the MC rate than $\mathcal{R}_r$. Thus, the reliability of the SVM classifier for the $\mathcal{R}$ will not significantly change even if both regions belong to two different classes. The second issue comes from the presence of mixed pixels within the hyperspectral image (i.e., pixels that actually are a combination of several materials belonging to different ground truth classes). Those pixels do not consequently have a high probability of belonging to a specific class, and an important MC rate can result for regions containing such mixed pixels. To solve those problems, the MC rate was modified by Valero et al. (2013) as follows:

$$\mathcal{M}\mathrm{C}_{\mathcal{R}} = 1 - BC\left(\mathcal{P}_{\mathcal{R}_l}, \mathcal{P}_{\mathcal{R}_r}\right) \tag{35}$$

where $BC\left(\mathcal{P}_{\mathcal{R}_l}, \mathcal{P}_{\mathcal{R}_r}\right)$ is the BC between the probability class distributions of $\mathcal{R}_l$ and $\mathcal{R}_r$ as defined by (Eq. 18). However, this expression cannot be used for leaf nodes as they do not have any children. Therefore, Valero et al. (2013) proposed to use the first MC rate (Eq. 34) for leaf nodes, and the modified one for all other nodes in the BPT structure. Note that the usage of the modified MC rate solves the issue arising when two sibling have very different sizes. But as it is also very sensitive to small regions, a node that is wrongly classified and composed of a single or a very few pixels can yield a high MC rate. To solve this weakness, Valero et al. (2013) declared that regions with size below a certain threshold (set to 3 pixels in practice) were considered as unreliable and could not cut the BPT branches.

Once the BPT has been populated with the classification information, the decision step evaluates whether a node can be removed or retained. In the latter case, it means that this node is able to well represent all its descendants from a classification accuracy point of view. Therefore, the decision function F was defined by Valero et al. (2013) as:

$$\mathrm{F}(\mathcal{M}\mathrm{R}_{\mathcal{R}}) = \mathcal{M}\mathrm{R}_{\mathcal{R}} - \frac{1}{|\mathsf{leaves}(\mathcal{R})|}\sum_{r\in \mathsf{leaves}(\mathcal{R})} \mathcal{M}\mathrm{R}_r \qquad (36)$$

where $\mathsf{leaves}(\mathcal{R})$ is the set of leaf nodes handing under region $\mathcal{R}$. Therefore, $\mathrm{F}(\mathrm{MR}_{\mathcal{R}})$ evaluates the difference between the MR rate of node $\mathcal{R}$ with respect to the average MR rate of all its leaves. If this difference increases too much, then $\mathcal{R}$ cannot accurately account for the classification information contained by the leaf nodes from which it is the ancestor, hence the introduction of a threshold $\alpha_{\mathcal{C}}$: if $\mathcal{F}(\mathcal{M}\mathrm{R}_{\mathcal{R}}) < \alpha_{\mathcal{C}}$, then $\mathcal{R}$ can represent its subtree, which can then be removed from the original BPT. On the other hand, $\mathcal{R}$ cannot be a leaf of the pruned tree. Therefore, the threshold $\alpha_{\mathcal{R}}$ controls the size of the pruned BPT, thus the larger the threshold, the smaller the pruned tree.

The data set used to evaluate the described classification pruning methodology corresponds to Pavia University hyperspectral image from reflective optics System imaging spectrometer (ROSIS) sensor. The image is composed of 610 × 340 pixels and contains 103 channels. A false color composition is displayed by **Fig. 12A**. **Fig. 12B** shows the ground truth classification map, composed of nine different classes, in which 20% of pixels for each class were used to train the SVM classifier (see **Fig. 12C**). In order to use some leaf nodes as training pixels, the BPT is initialized at the pixel level (where each single pixel is an individual region of the initial segmentation map). The histogram-based region model (Eq. 16) (using $N_{bins} = 256$) and MDS merging criterion (Eq. 23) (with a number of principal components being $D_s = 2$) are selected to parameterize the merging procedure (a quantitative comparison with other settings can be found in Valero et al. (2013).

After the completion of the BPT construction, its population step is conducted following the exposed procedure: each node $\mathcal{R}$ has its class probability distribution $\mathcal{P}_{\mathcal{R}}$ assigned to it in order to compute its MC rate $\mathrm{MC}_{\mathcal{R}}$. The decision stage then requires the definition of the threshold $\alpha_{\mathcal{C}}$ to compare the own MC rate of each node against the average one of all its leaves. Several values ranging between 0 and 0.4 were investigated by Valero et al. (2013), considering that values of $\alpha_{\mathcal{C}}$ above 0.4 were associated with a too high MC error.

In addition to the classical pixel-wise procedure, obtained results are also compared with the spectral–spatial approach proposed by Fauvel et al. (2012). In this work, two kernels functions are associated to combine both spectral and spatial information within the SVM classification process by means of a morphological area filtering. **Fig. 13** shows the classification maps obtained by the pixel-wise classification on **Fig. 13A**, the spectral–spatial approach (Fauvel et al., 2012) on **Fig. 13B** and the results obtained after applying BPT pruning strategy on **Fig. 13C**. It can be observed that the BPT-based classification map is composed of quite homogeneous regions, and that the classification salt-and-pepper noise is well reduced with respect to the pixel-wise strategy. As a matter of fact, the use of the BPT structure guarantees not to introduce any fake edge (which would not be the case of a classical postprocessing such as the Markovian regularization (for instance, Tarabalka et al. (2010c)), since the regions in the classification map actually correspond to real regions as determined during the construction of the BPT. Moreover, the final classification map can contain both small (but meaningful) and large regions, since the hierarchical structure allows them to be selected at different scales. In order to compare the spectral–spatial and BPT-based approaches, **Table 2** presents the class-specific and global classification accuracies

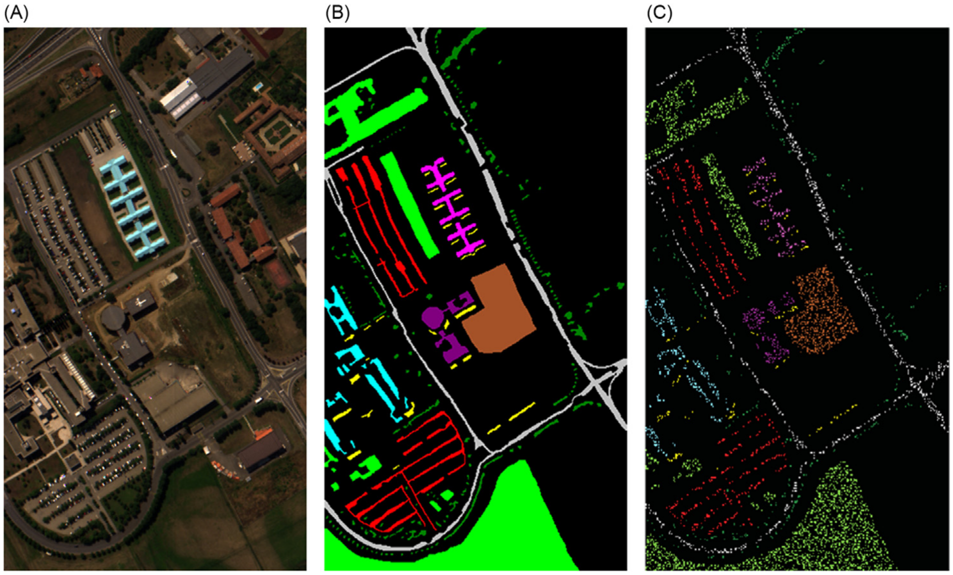


**Fig. 12** (A) RGB composition of ROSIS Pavia university scene, with (B) the classification ground truth data and (C) the training data used in this experiment.

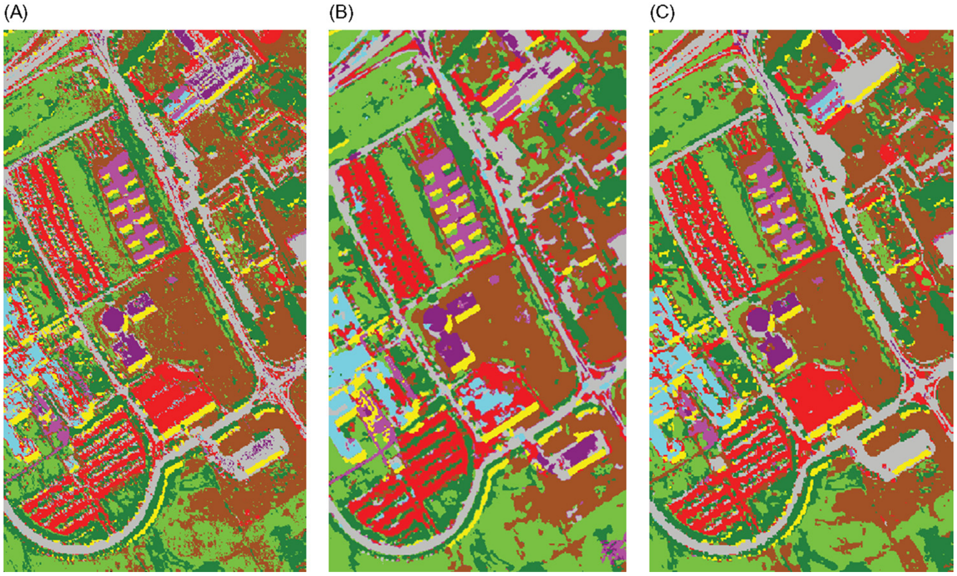


**Fig. 13** (A) pixel-wise classification, (B) spectral–spatial classification following (Fauvel et al., 2012) and (C) BPT-based classification (Valero et al., 2013).

**Table 2** Class specific accuracy for Pavia university data set for the pixel-wise SVM classification, the spectral–spatial classification method proposed by Fauvel et al. (2012) and the BPT-based classification method presented in Valero et al. (2013)

| *Class* | *Simple SVM* | *Spectral–spatial approach* | *BPT-based approach* |
|---|---|---|---|
| 1 | 85.93 | 83.6 | **88.84** |
| 2 | 76.66 | **77.9** | 71.69 |
| 3 | 70.46 | 82.9 | **91.95** |
| 4 | **97.55** | 96.7 | 95.14 |
| 5 | **99.55** | 98.7 | 98.81 |
| 6 | 91.99 | 95.2 | **97.08** |
| 7 | 92.48 | 94.0 | **99.02** |
| 8 | 92.31 | 95.0 | **98.13** |
| 9 | 89.26 | **97.4** | 95.99 |
| Overall | 88.58 | 91.26 | **92.96** |

Highest values are in bold.

(highest values are in bold). The BPT-based classification strategy outperforms the simple SVM and the spectral–spatial approach for most classes, as well as for the overall accuracy. This confirms the ability of the hierarchical structure to efficiently describe the data at various scales, as well as the soundness of the classification pruning process proposed by Valero et al. (2013). Nevertheless, thanks to the flexibility of the BPT framework, the classification information could be integrated directly during the construction step, and other pruning strategies could be also proposed.

### 2.05.4.3 Object Recognition in Urban Hyperspectral Scene

Object detection is a computer vision application, which aims at recognizing and extracting some object of interest from a given image. Designing an automated object detection process can be more or less complex depending on the level of details expected for the sought object (whether the goal is to state if the object is present in the scene or to retrieve its precise boundaries for instance). Assuming that the object is only local with respect to the whole image, and can be discriminated using an appropriate set of features, object detection procedures have very quickly evolved from pixel-based methods to sliding windows (Viola and Jones, 2004) and further to segmentation maps (Russell et al., 2006). As a matter of fact, Malisiewicz and Efros (2007) showed that the use of a so-called soup of segment was greatly enhancing the object detection performances by readily providing a restrained set of candidate objects. Hierarchical image representations are suitable candidates to provide this soup of segments, as they naturally decompose the image into a set of relevant regions across the image support and at various scales. In addition to directly providing a finite number of candidate regions (supported by the node of the tree structure), the candidates they propose are already meaningful (at least with respect to the criterion, which was adopted to construct the hierarchical representation). There are several instances of object detection procedures supported by hierarchical representations in the literature. Akcay and Aksoy (2008) use, for instance, the set of hierarchical connected components generated using morphological profiles for building retrieval. Methods based on the ToS have also been investigated in Xu et al. (2012a) and Carlinet and Géraud (2015), notably for text detection in video sequences. Finally, Salerno et al. (2004) and Vilaplana et al. (2008) have proposed to use BPT representations for faces and road signs detection, whereas Valero et al. (2015), from which we review the work in this section, implement it for building retrieval in urban hyperspectral scenes. From a general point of view, the key aspect of object detection processes is the suited definition of the reference feature set, as they must be discriminating enough to ensure that the object will be identified with respect to its background. Therefore, this stage has to be done beforehand, using any prior information on the object to be detected. More formally, the object detection procedure can also be decomposed following the two steps of a classical BPT processing (see section "**Processing of the BPT**"). Let $H = \{\mathcal{R} \subseteq E\}$ be the considered hierarchical representation accounting for the soup of segment, and let $\Omega^{\text{ref}} = \{\omega_i^{\text{ref}}\}$ be a given set of reference features corresponding to the object of interest, where each $\omega_i^{\text{ref}}$ is an individual feature. For the population step, the object detection process retrieves for each region $\mathcal{R} \in H$ its set of features $\Omega^{\mathcal{R}} = \{\omega_i^{\mathcal{R}}\}$ in the image. Following, the decision step evaluates the closeness between $\Omega^{\text{ref}}$ and $\Omega^{\mathcal{R}}$ for the decision function F, that is, $\mathcal{F}(\Omega^{\mathcal{R}}) = d(\Omega^{\text{ref}}, \Omega^{\mathcal{R}})$, where $d(\cdot,\cdot)$ is a distance (respectively, a similarity) function defined according to the application. The selected region from the hierarchical representation $H$ is the one minimizing (respectively, maximizing) this function. Alternatively, all regions below (respectively, above) some threshold can be retained.

For building detection purposes, Valero et al. (2015) proposed to use different spectral and spatial features, such as

– The spectral class probability distribution $\omega_1^{\mathcal{R}} = \mathcal{P}_{\mathcal{R}}$, where $\mathcal{P}_{\mathcal{R}}(\mathcal{C}_i) = \text{proba}(\mathcal{R} \text{ is classified in } \mathcal{C}_i)$ is the probability that region $\mathcal{R}$ is classified in class $\mathcal{C}_i, i = 1, \ldots, C$. $\mathcal{P}_{\mathcal{R}}$ is the output of a classification stage (typically using a probabilistic SVM classifier) performed on the node $\mathcal{R}$ (see previous section "**Hyperspectral Image Classification**" for more details). The corresponding feature similarity function is simply the probability that $\mathcal{R}$ belongs to the target class $\mathcal{C}_{target}$ (to which belongs the object to detect), that is,

$$d\left(\omega_1^{\mathcal{R}}, \omega_1^{\text{ref}}\right) = \mathcal{P}_{\mathcal{R}}\left(\mathcal{C}_{target}\right) \quad . \tag{37}$$

– The spectral class membership homogeneity $\omega_2^{\mathcal{R}}$, which evaluated whether a given region $\mathcal{R}$ is homogeneous in terms of class membership. This term is important in the BPT context, as nodes close to the root of the hierarchy are likely to represent regions combining many different classes. The associated similarity function is defined as

$$d\left(\omega_2^{\mathcal{R}}, \omega_2^{\text{ref}}\right) = BC\left(\mathcal{P}_{\mathcal{R}_l}, \mathcal{P}_{\mathcal{R}_r}\right) \tag{38}$$

where $\mathcal{P}_{\mathcal{R}_l}$ and $\mathcal{P}_{\mathcal{R}_r}$ are the class probability distributions of the left and right child nodes of $\mathcal{R}$, and $BC(\cdot,\cdot)$ denotes the BC (Eq. 18). Note that if two sibling nodes have similar class probability distributions, their union will also have a similar distribution, that is, the object is in the process of being formed.

– The region area $\omega_3^{\mathcal{R}} = |\mathcal{R}|$. The goal of this feature is to prevent the detection of too small (close to leaf nodes) or too large (close to the root) meaningless regions. Assuming that the object to detect has a size comprised between $\mathcal{A}_{min}$ and $\mathcal{A}_{max}$, the feature similarity is a simple hard thresholding with respect to the interval $[\mathcal{A}_{min}; \mathcal{A}_{max}]$:

$$d\left(\omega_3^{\mathcal{R}}, \omega_3^{\text{ref}}\right) = \mathbf{11}_{[\mathcal{A}_{min};\mathcal{A}_{max}]}(|\mathcal{R}|) \tag{39}$$

where $11_X$ is the indicator function of set $X$.

– The area of the smallest oriented bounding box $\omega_4^{\mathcal{R}} = |\mathcal{BB}(\mathcal{R})|$, in order to incorporate some knowledge on the expected shape of the object. The associated feature similarity is defined as the ratio between the area of the bounding box $\mathscr{BB}(\mathcal{R})$ and the one of $\mathcal{R}$,

$$d\left(\omega_4^{\mathcal{R}}, \omega_4^{\text{ref}}\right) = \frac{|\mathcal{BB}(\mathcal{R})|}{|\mathcal{R}|} \tag{40}$$

It measures the compactness of region $\mathcal{R}$, as buildings are expected to be of rectangular shapes (and thus highly compact).

The overall similarity between the set of features $\Omega^{\mathcal{R}}$ of the candidate region $\mathcal{R}$ and the reference set $\Omega^{\text{ref}}$ is finally defined by Valero et al. (2015) as

$$d\left(\Omega^{\mathcal{R}}, \Omega^{\text{ref}}\right) = \prod_{i=1}^{4} d\left(\omega_i^{\mathcal{R}}, \omega_i^{\text{ref}}\right) \tag{41}$$

It is interpreted as the likelihood of region $\mathcal{R}$ to be the sough object. Therefore, a high value (theoretically, 1) for $d(\Omega^{\mathcal{R}}, \Omega^{\text{ref}})$ means that region $\mathcal{R}$ is very likely to be an instance of the object to retrieve, and region $\mathcal{R}$ should be retained. The major drawback of this approach is that small regions close to the leaves of the hierarchy could receive a high overall value by means of (Eq. 41) even though the object is still being formed. A more robust approach proposed by Valero et al. (2015) is to evaluate the evolution of $d(\Omega^{\mathcal{R}}, \Omega^{\text{ref}})$. Plotting this curve from a leaf node to the root reveals three different behaviors: starting from low values (for nodes very close to the leaf), the first interesting point corresponds to a noticeable increase in terms of similarity value $d(\Omega^{\mathcal{R}}, \Omega^{\text{ref}})$, meaning that the object is in process of being formed. Then a stable range is observed for $d(\Omega^{\mathcal{R}}, \Omega^{\text{ref}})$, before it suffers from a sharp drop at a node whose region in the image corresponds to the reunion of the sought object of interest with a neighboring region. Therefore, the node that should be picked to represent the object is the one located prior to this drop. This procedure was automated

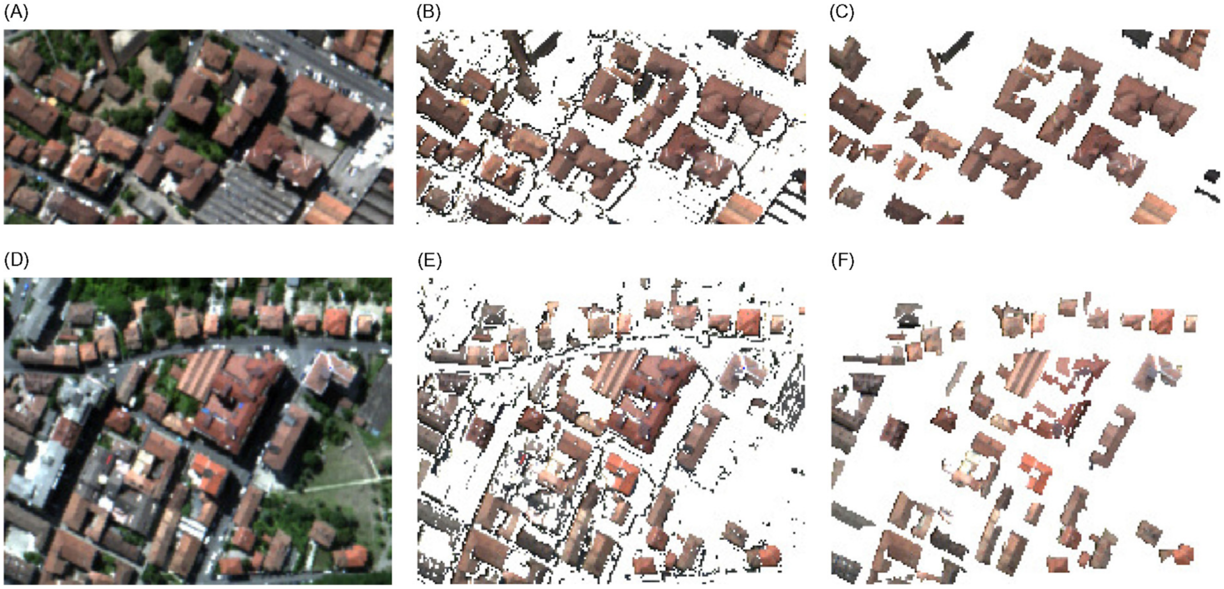


**Fig. 14** (A) and (D): RGB composition of a subset of Pavia city center data set. (B) and (E) Corresponding pixel-wise SVM classification. (C) and (E) Corresponding BPT-based building detection.

by Valero et al. (2015) by selecting the node $\mathcal{R}^{\star}$ whose similarity value is high enough, and whose difference in similarity value with respect to its father is the largest, that is,

$$\mathcal{R}^{\star} = \underset{\substack{\mathcal{R} \in \mathrm{br}_{\mathcal{R}} \\ d\left(\Omega^{\mathcal{R}}, \Omega^{\mathrm{ref}}\right) \geq \delta}}{\operatorname{argmax}} \left( d\left(\Omega^{\mathcal{R}}, \Omega^{\mathrm{ref}}\right) - d\left(\Omega^{\mathcal{F}(\mathcal{R})}, \Omega^{\mathrm{ref}}\right) \right) \tag{42}$$

where $\mathrm{br}_{\mathcal{R}}$ denote the branch containing $\mathcal{R}$ that is being analyzed (that is, one of the branches linking one of the leaves of $\mathcal{R}$ and passing through $\mathcal{R}$ to reach the root), $\delta$ is a threshold ensuring that $\mathcal{R}$ is likely enough to be the sought object, and $\mathrm{F}(\mathcal{R})$ is the father of $\mathcal{R}$. Note that, as $\mathcal{R}$ belongs to several branches (as many as the number of leaves hanging under it), it will be analyzed several times, but may not however be picked by all its leaves. To solve any potential conflictual case, it was decided in Valero et al. (2015) to select the closest node from the root if the criterion (Eq. 42) highlighted several nodes in the same branch, as the region analysis may be more reliable for large regions.

The BPT-based building retrieval methodology previously described is evaluated onto two different subsets of ROSIS Pavia city center data set, as presented in **Fig. 14A and D**. The BPT is built from the pixel level with a histogram-based region model and MDS merging criterion (where $D_s$ is also set to 2). As in section "**Hyperspectral Image Classification**," the probability class distribution $\mathcal{P}_{\mathcal{R}}$ is obtained by making use of a probabilistic SVM classifier, allowing to compute features $\omega_1^{\mathcal{R}}$ and $\omega_2^{\mathcal{R}}$. The target class $\mathcal{C}_{target}$ is set to be the tile class provided by the ground truth data (http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenesredirect=no). The target area range for $\omega_3^{\mathcal{R}}$ is set to be $[\mathcal{A}_{min}, \mathcal{A}_{max}] = [30; 1000]$, and the threshold $\delta$ stating whether a node is likely to be an instance of the object or not is fixed to $\delta = 0.65$. The obtained detection results are displayed in **Fig. 14C and F**, and are compared with the tile class from the pixel-wise SVM classification (**Fig. 14B and E**). Overall, the BPT-based detection results outperform the pixel-wise SVM classification. The use of a hierarchical region-based representation of the image is clearly beneficial for the object detection procedure, as it leads to more consistent and less noisy detected objects than the pixel-based SVM. The presented results are also comparable with those obtained by Akcay and Aksoy (2008). Regarding the results, the only downside of the proposed methodology concerns the detection of aggregated buildings. While they may all be individually rectangular, their shape might become more complex if they are all spatially adjacent to each other. This can either lead to detect them as a whole structure (and not a stack of different buildings), or to partial detection only, as it can be seen in the middle of **Fig. 14f**. Nevertheless, this issue could be solved by the use of more complex spatial features.

### 2.05.4.4 Anomaly Detection in Hyperspectral Data

Target detection designates the process of discovering the presence of a specific signal of interest (being the target) among a set of signals. In the case of hyperspectral images, the handled signals are the various pixel spectra, and target detection algorithms basically behave like binary classifiers, labeling every pixel as either being the sought target or belonging to the background (Nasrabadi, 2014). Anomaly detection is a special case of target detection, where no a-priori target is provided. Thus, the goal of anomaly detection is to distinguish observations of unusual materials from typical background signatures without any information related to the nature of the anomalies (Stein et al., 2002). Anomaly and target detection find diverse applications such as surveillance, rare minerals detection, mine detection, and so on (Manolakis et al., 2014). In those scenarios, the background is modeled according to some multivariate statistical distribution, and the anomaly detection process then tests whether each pixel spectrum **x**, considered as the realization of a multivariate random variable, is likely to be drawn from the background distribution (being the null hypothesis $\mathcal{H}_0$) or not:

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{x} = \mathbf{b} \\ \mathcal{H}_1 &: \mathbf{x} = \mathbf{s} + \mathbf{b} \end{aligned} \tag{43}$$

where **b** represents the background, and **s** denotes the presence of an anomalous signal. **x** is an anomaly when $\mathcal{H}_1$ is decided. Assuming that the background statistics can be modeled as a multivariate Gaussian distribution (whose mean and covariance are estimated from the pixels in the image) leads to the benchmark Reed-Xiaoli (RX) anomaly detector (Reed and Yu, 1990). The statistical mean and covariance parameters can be estimated either from the whole image or from a local neighborhood $\mathcal{V}(x)$ of **x**. The latter case is called the adaptive RX anomaly detector. The major advantage of the adaptive RX version over the global one is that it is able to adapt to background nonhomogeneities and can handle the case where the noise independence assumption is violated by adapting to the local specificities of the data. More specifically,

$$\widehat{\boldsymbol{\mu}}_{SMV} = \frac{1}{|\mathcal{V}(x)|} \sum_{y \in \mathcal{V}(x)} \mathbf{y} \tag{44}$$

being the estimated sample mean vector (SMV) of pixel **x** in its neighborhood $\mathcal{V}(x)$, and

$$\widehat{\boldsymbol{\Sigma}}_{SCM} = \frac{1}{|\mathcal{V}(x)|} \sum_{y \in \mathcal{V}(x)} (\mathbf{y} - \widehat{\boldsymbol{\mu}}_{SMV})(\mathbf{y} - \widehat{\boldsymbol{\mu}}_{SMV})^T \tag{45}$$

being the similarly estimated sample covariance matrix (SCM), the generalized likelihood ratio test yields

$$t_{RX}(\mathbf{x}) = (\mathbf{x} - \widehat{\boldsymbol{\mu}}_{SMV})^T \widehat{\boldsymbol{\Sigma}}_{SCV}^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_{SMV}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_{GLRT}. \tag{46}$$

Due to the fact that the mean and covariance matrix are estimated from the data, it is known that the statistic value $t_{RX}(\mathbf{x})$ of **x** follows a noncentral F-distribution with $N$ (being the number of channels in the hyperspectral image) and $|\mathcal{V}(x)| - N + 1$ degrees of freedom when rescaled as (Rencher and Christensen, 2012)

$$\frac{|\mathcal{V}(x)| - N + 1}{N|\mathcal{V}(x)|} t_{RX}(\mathbf{x}) \sim F_{N, |\mathcal{V}(x)| - N + 1}. \tag{47}$$

For large neighborhood $\mathcal{V}(x)$ (i.e., $|\mathcal{V}(x)| > 10 \times N$), the distribution of $t_{RX}(\mathbf{x})$ can be approximated by the classical $\chi^2$ distribution.

The conventional approach to develop an adaptive RX algorithm is by using sliding windows. For a given tested pixel **x**, two different windows, centered on **x**, are defined: a small one with the size of the expected anomalies maximum size, named the guard window, and a larger one, usually termed outer window. The pixels inside the outer window, except those lying inside the guard window, compose the neighborhood $\mathcal{V}(x)$ from which the SMV (Eq. 44) and SCM (Eq. 45) are estimated. The size of $\mathcal{V}(x)$ has to be chosen large enough to ensure the invertibility of the covariance matrix and small enough to justify both spectral homogeneity (stationarity) and spatial homogeneity. The main limitation of the sliding windows approach is that the definition of square or rectangle windows is not adapted to the geometry of the scene. In order to tackle this issue, Veganzones et al. (2014a) propose to make the most of the inclusion relationship holding between the nodes composing a BPT representation to define the guard and outer windows, and thus the adaptive neighborhood $\mathcal{V}(x)$. More specifically, let **x** be the spectrum of the pixel $x$ under test, and denote by $\mathrm{L}_x$ the leaf to which pixel $x$ belongs to in the BPT (i.e., $x \in \mathrm{L}_x$) and $\mathsf{br}_x = \mathsf{br}(\mathrm{L}_x)$ the branch between $\mathrm{L}_x$ and the root. Obviously, $x \in \mathcal{R}$, $\forall \mathcal{R} \in \mathsf{br}_x$. Then, the guard node $\mathfrak{G}_x$ of $x$ is defined as the smallest node in $\mathrm{br}_x$ whose size is above a fixed threshold $\mathcal{A}_{guard}$. Similarly, the outer node $\mathfrak{O}_x$ is defined as the smallest node in $\mathrm{br}_x$ whose size is above a larger fixed threshold $\mathcal{A}_{outer} > \mathcal{A}_{guard}$. The adaptive neighborhood $\mathcal{V}(x)$ of $x$ is then defined as all pixels contained in $\mathfrak{O}_x$ but not in $\mathfrak{G}_x$: $\mathcal{V}(x) = \mathfrak{O}_x \backslash \mathfrak{G}_x$. All pixels in the image are then tested following this methodology, allowing to derive an anomaly detection map. Note that, even though the approach proposed by Veganzones et al. (2014a) defines neither a BPT pruning nor a BPT node selection process (as there is no population and decision stages), this procedure completely relies on the hierarchical representation of the data provided by the BPT, and can thus be considered as a BPT-based application.

The BPT-based adaptive RX anomaly detector methodology is investigated by Veganzones et al. (2014a) on the hyperspectral data set provided by the Target Detection Blind Test project (details are provided in Snyder et al. (2008)). The dataset, collected by the HyMap sensor, includes a high-resolution hyperspectral image with approximately 3 m resolution, spectral libraries of targets in the scene, and the location of targets in the scene. **Fig. 15A** depicts a false color image of the dataset, and **Fig. 15B** shows the location of the targets in the scene, constituted of three civilian vehicles and four small fabric panels for a total of 129 target pixels in the image. The BPT is built using the multidimensional watershed (Tarabalka et al., 2010b) for the initial partition, the mean spectrum (Eq. 5) for the region model, and the SAM distance (Eq. 6) as a merging criterion. The adaptive RX detector is calculated for all pixels following (Eq. 46). For the sliding windows approach, the guard window has size 3 × 3 and the outer window is of size

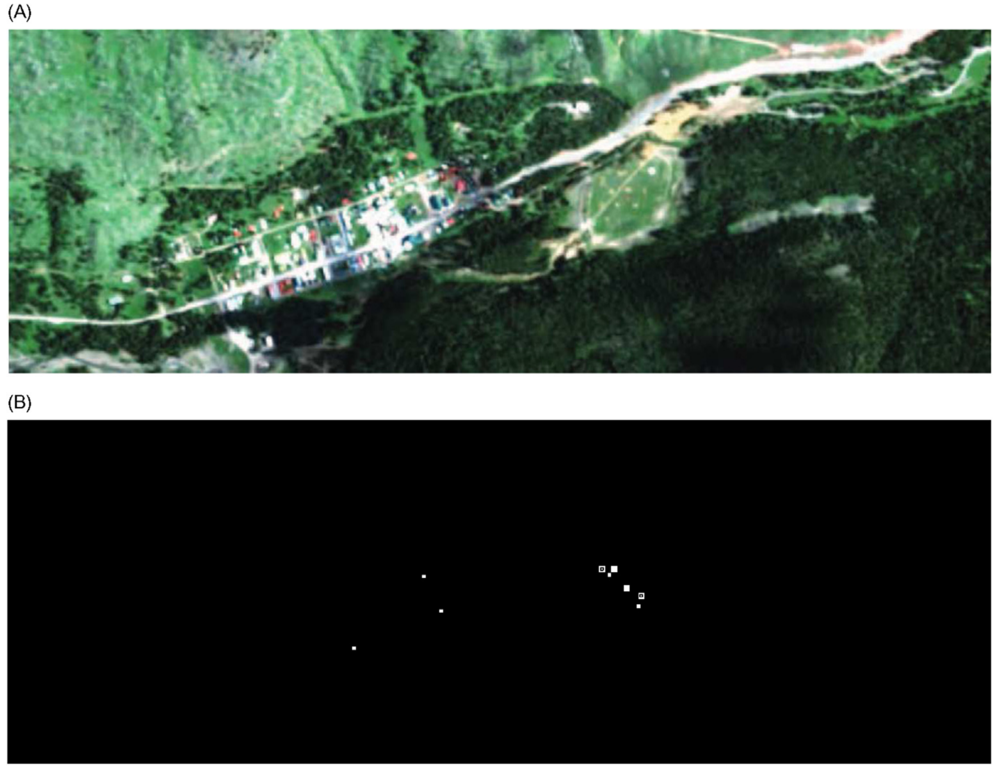


**Fig. 15** (A) False color representation of the Target Detection Blind Test hyperspectral data set, and (B) location of the targets in the scene.

21 × 21. For the BPT-based approach, the minimal sizes for the guard node and outer node are set to $\mathcal{A}_{guard} = 9$ pixels and $\mathcal{A}_{outer} = 400$ pixels, respectively. For both approaches, the probability of detection $p_d$ is calculated for various values of probability of false alarm $p_{fa}$, using as ground-truth the location of the 129 target pixels. Fig. 16A shows the resulting ROC curve for both approaches and it can be appreciated how the BPT-based method outperforms the sliding windows approach in almost all probability of false alarm range. This improvement is due to the definition of more homogeneous regions to estimate the statistic parameters, leading to a more robust test statistic $t_{RX}(\mathbf{x})$ for all pixels in the image. This is confirmed by Fig. 16B–C, which exhibit the detections obtained by the adaptive RX algorithm over the test dataset for a probability of false alarm set to 0.05, using the conventional sliding windows and BPT-based approach, respectively. As a matter of fact, the detected anomalies look more homogeneous and less noisy for the BPT-based method, suggesting that taking into account the geometry of the scene is indeed beneficial to the definition of the adaptive neighborhood $\mathcal{V}(x)$ and estimation of the statistical parameters. Moreover, the BPT-based approach defined by Veganzones et al. (2014a) could be easily extended to other anomaly or target detection techniques.

### 2.05.4.5 Filtering and Segmentation of SAR Images

We now conclude this section devoted to the description of state-of-the-art BPT applications for remote sensing data analysis by presenting some works related to SAR image processing by means of BPT representations. More specifically, we focus in the following on the SAR filtering procedure introduced in Alonso-González et al. (2013) and SAR segmentation approaches as described in Alonso-González et al. (2012, Salembier (2015, and Salembier and Foucher (2016).

#### *2.05.4.5.1 Speckle noise filtering*

SAR imagers are becoming more and more popular for Earth observation purposes because of their ability to work regardless of the day/night light cycle and their insensitivity to the varying weather conditions. SAR sensors are able to localize on-ground targets by illuminating the scene with electromagnetic waves and further recording their echoes. PolSAR is a specific SAR imaging system that considers different polarization states for the transmitted and received waves. Their capacity of exploring the complete space of polarization states, and thus being sensitive to the geometry and dielectric properties of the target, represents one of the most important properties of PolSAR data. To achieve high spatial resolutions, SAR systems repeatedly illuminate the target thanks to the sensors motion, resulting in coherently recorded echoes. The speckle phenomenon follows from this coherent addition of multiple scattered electromagnetic waves. Despite representing a true electromagnetic measurement, the speckle is of such complexity that it has to be modeled from a stochastic point of view and is thus assimilated to a noise term that shall be filtered out from the SAR data to grant access to the information of interest.

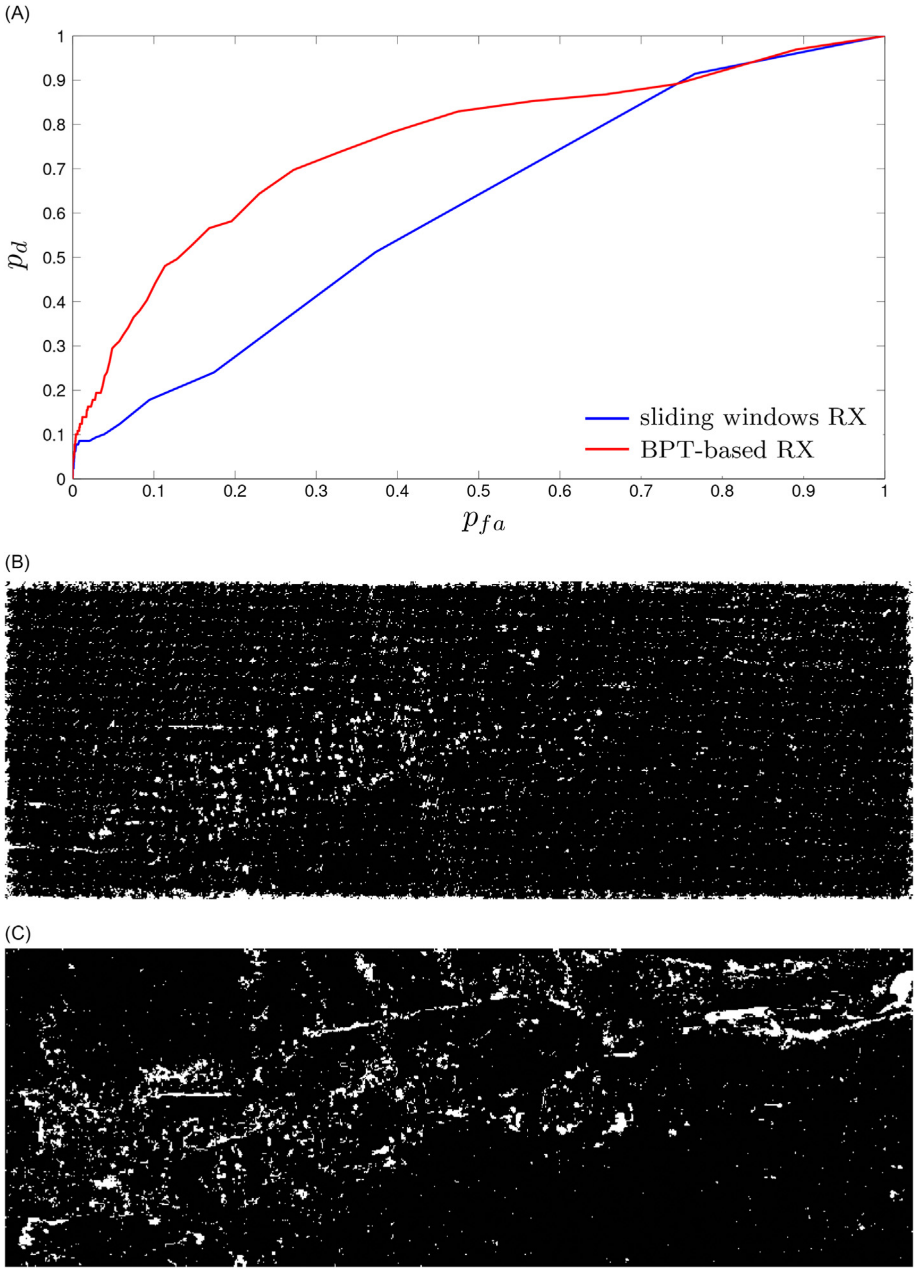


**Fig. 16** (A) ROC curves for the compared sliding windows and BPT-based approaches, and resulting detections of the adaptive RX algorithm ($p_{fa} = 0.05$) for (B) the sliding windows and (C) BPT-based approaches.

SAR speckle filtering has been the concern of numerous studies in the literature, the most well known being the window-based Frost filter (Frost et al., 1982) and Lee filter (Lee et al., 1999), or the more complex adaptive filter presented in Vasile et al. (2006). Those approaches define a neighborhood where the data is assumed to be locally stationary. It raises the usual question about the scale at which those neighborhood should be considered. Therefore, Alonso-González et al. (2012, 2013) proposed to tackle this issue by taking advantage of the hierarchical nature of the BPT representation for SAR speckle noise filtering. More specifically, they designed a BPT pruning strategy to retrieve the most homogeneous regions in terms of covariance matrices. This homogeneity was evaluated for each region $\mathcal{R}$ of the BPT as follows:

$$\begin{aligned}\Phi(\mathcal{R}) &= \frac{1}{|\mathcal{R}|}\sum_{x\in\mathcal{R}}\frac{\|\widehat{\mathbf{C}}_x-\widehat{\mathbf{C}}_{\mathcal{R}}\|_F^2}{\|\widehat{\mathbf{C}}_{\mathcal{R}}\|_F^2}\\ &= \frac{1}{|\mathcal{R}|\,\|\widehat{\mathbf{C}}_{\mathcal{R}}\|_F^2}\sum_{x\in\mathcal{R}}\|\widehat{\mathbf{C}}_x-\widehat{\mathbf{C}}_{\mathcal{R}}\|_F^2\end{aligned} \tag{48}$$

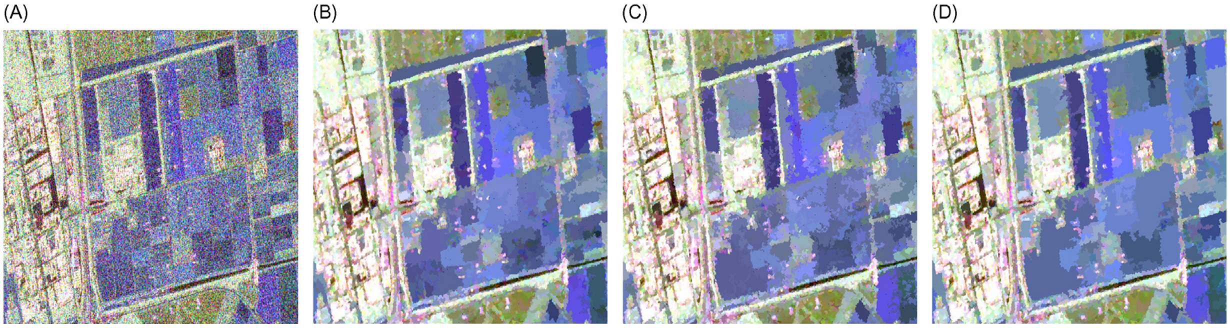


**Fig. 17** (A) 512 × 512 subset of the SAR image, and resulting filtered image with $\delta_p = -4\,\text{dB}$ for (B) the RWS, (C) the DWS and (D) the GS merging criteria.

where $\hat{\mathbf{C}}_{\mathcal{R}}$ stands for the estimated covariance matrix (Eq. 12) for the whole region $\mathcal{R}$, $\hat{\mathbf{C}}_x$ is the covariance matrix associated with pixel $x$, and $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^H)}$ is the Frobenius norm of matrix $\mathbf{A}$. The defined homogeneity criterion (Eq. 48) can be viewed as the mean loss of information occurring when modeling the region $\mathcal{R}$ by its estimated covariance matrix. Therefore, the lower $\Phi(\mathcal{R})$, the more homogeneous the region $\mathcal{R}$. Having populated the BPT following the homogeneity criterion, a minimum rule is then employed by Alonso-González et al. (2012, 2013) for the decision step. Using a top-down approach, each region $\mathcal{R}$ of the BPT (starting from the root and going down to the leaves) is analyzed. If its homogeneity $\Phi(\mathcal{R})$ is above some threshold $\delta_p$, then the region $\mathcal{R}$ is declared not to be homogeneous enough, and its two children are further investigated. The final output of this procedure is therefore some partition $\pi = \{\mathcal{R}_i\}$ where each region $\mathcal{R}_i$ is such that $\Phi(\mathcal{R}_i) \leq \delta_p$. In other words, the regions in the obtained are sufficiently homogeneous to claim that the speckle noise has been filtered out by the pruning process.

The BPT-based speckle filtering technique has been assessed by Alonso-González et al. (2013) on a PolSAR image composed of 5300 × 3100 pixels and acquired over Flevoland, The Netherlands, by the RADARSAT-2 sensor. The scene covers an area of 25 × 25 $\text{km}^2$ with a spatial resolution of 5.2 m in range and 7.6 m in azimuth, and is mainly composed of agricultural fields, sea, and urban areas. Fig. 17A shows a 512 × 512 subset of the original PolSAR image. To evaluate the influence of the merging procedure on the BPT-based speckle filtering, three BPT instances are built using the SCM (Eq. 12) as a region model and the RWS (Eq. 13), the DWS (Eq. 14), and the GS (Eq. 15) as merging criteria over the PolSAR data, filtered with a bilateral filter (Tomasi and Manduchi, 1998) beforehand. Fig. 17B–D present the obtained results when the filtering threshold parameter is set to $\delta_p = -4$ dB. In any case, it can be observed that the BPT approach is able to retain the spatial details that are associated with point targets within the urban areas, while eliminating most of the speckle noise that contaminates the agricultural fields. As a matter of fact, the speckle filtering pruning strategy is able to select large homogeneous regions (close to the root of the BPT structure) as well as very small regions (close to the leaf nodes) at the same time thanks to the hierarchical structure of the BPT that describes the image at different scales. A closer analysis of Fig. 17B–D shows that contours are better preserved when using the RWS merging criterion rather than the DWS one, since the latter only employs the diagonal terms of the covariance matrices of the two regions being compared. Furthermore, it also seems that the use of the GS merging criterion leads to larger regions for the same level of filtering, which would imply that the regions merging during the construction of the BPT were less affected by the speckle noise.

Fig. 18 presents the influence of the filtering threshold $\delta_p$ on the speckle noise removal. When the value of $\delta_p$ increases (from $-7$ dB in Fig. 18A to $-1$ dB in Fig. 18D), larger regions are created since the homogeneity constraint is relaxed. Nevertheless, it should be stressed that no new contours or artifacts are introduced in the process despite the filtering becoming more severe, as the resulting regions remain organized in a hierarchical manner (the obtained pruning partition for $\delta_p = -7\,\text{dB}$ is a refinement of the one for $\delta_p = -5\,\text{dB}$, which is itself finer than $\delta_p = -3\,\text{dB}$, and so on). Therefore, this filtering process fully exploits the hierarchical structure that is offered by the BPT representation by providing filtered regions at various scales (depending on the level of filtering that was required by the user) but which nonetheless preserve the image contours. This property makes the presented filtering strategy part of a largest family of tree processing tools called connected operators (Salembier and Wilkinson, 2009).

### *2.05.4.5.2 Segmentation of PolSAR images*

As already discussed in section "**Hyperspectral Image Segmentation**" for hyperspectral images, BPT representations are suitable tools for segmentation purposes as they condense in their structures all potential regions of interest, at various scales, in the image. In addition, any pruning strategy applied to the BPT structure results in a particular segmentation of the image (this was the case in particular for the speckle noise filtering previous presented). The segmentation of SAR images by means of BPT representations has notably been investigated by Alonso-González et al. (2012), Salembier (2015), and Salembier and Foucher (2016). In the former case, a very simple pruning strategy was employed for coastline segmentation, based on the $N_s$ most dissimilar regions in the image. This pruning technique, called the region-based pruning, provides the last $N_s$ regions standing before the completion of the BPT. Those can easily be obtained by traversing the tree structure in a reversed order with respect to its construction, and stops when the desired number of regions $N_s$ has been reached. As the optimal number of regions $N_s$ is very likely to be unknown in practical scenarios, the region-based pruning strategy is often used only for exploratory purposes, in order to validate whether the BPT

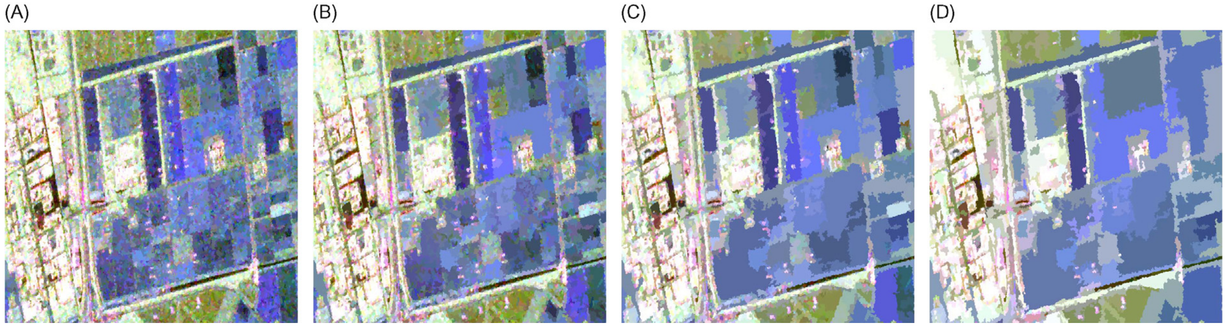


**Fig. 18** BPT-based speckle filtering results when varying the filtering threshold: (A) $\delta_p = -7$ dB dB, (B) $\delta_p = -5$ dB, (C) $\delta_p = -3$ dB and (D) $\delta_p = -1$ dB.

was constructed in a meaningful way or not. Nevertheless, there are some cases where $N_s$ can be deduce from the underlying task. In Alonso-González et al. (2012), for instance, the goal is to detect a coastline, and thus separate the mainland from the water. In such a case, it is clear that the simplest partition would be given by $N_s = 2$, assuming that one region accounts for the land and the other represents the sea.

This approach was validated by Alonso-González et al. (2012) over a RADARSAT-2 PolSAR image acquired over the coastline of Barcelona, in 2008, whose 1500 × 2500 pixels subset is displayed in **Fig. 19A**. The bilateral filter was initially employed to smooth out the image prior to the BPT construction; the latter was parameterized by the estimated covariance region model (Eq. 12) and the geodesic distance merging criterion (Eq. 15). The $N_s = 2$ most dissimilar regions (therefore being the two children of the BPT root node) are displayed by **Fig. 19B**. It can be seen that segmenting the image into the most two dissimilar regions indeed allows to separate the inland from the sea. Thin structures in the coastline, such as breakwaters, are also well preserved. Those two observations give the idea that the BPT has been properly built, so that a simple region-based pruning strategy perfectly achieves the intended task. Nevertheless, a major issue to this approach is that the number of optimal regions must be known in advance, which is seldom true in practice. Considering the coastline detection application, the region-based pruning strategy would become irrelevant if there are an unknown number of islands to additionally segment within the scope of the imaged scene. As already discussed in section "**Hyperspectral Image Segmentation**," a potential solution to this issue is to embed the segmentation procedure into an energy minimization framework, where the energetic function is designed to achieve a trade-off between data fitting, leading to oversegmentation, and simplicity through a regularization term. This approach has been investigated lately by Salembier (2015)

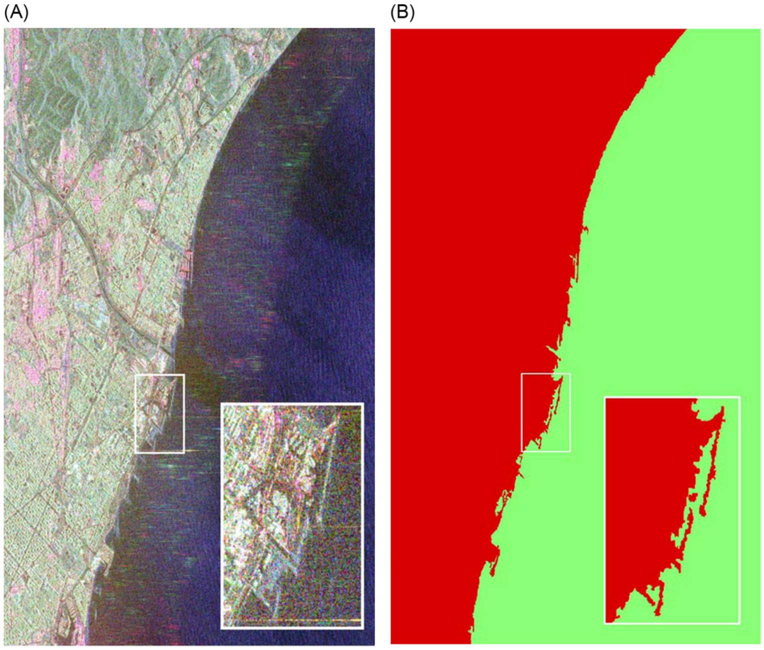


**Fig. 19** (A) C-band Pauli RADARSAT-2 image of Barcelona coastline, and (B) resulting segmentation showing the two most dissimilar regions.

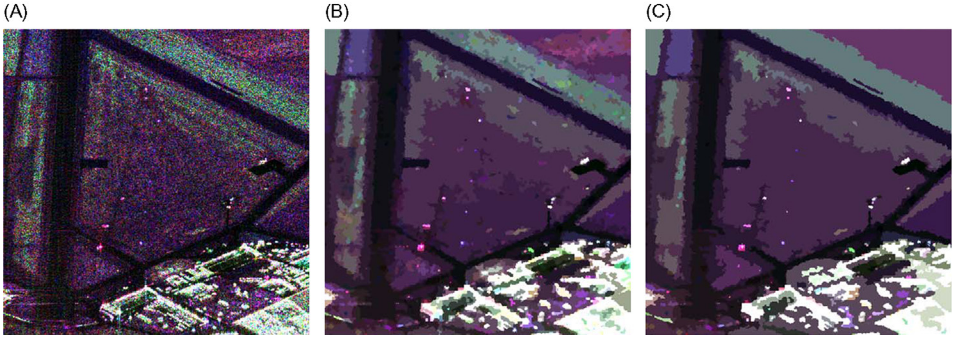


**Fig. 20** (A) original L-band PolSAR data of Oberpfaffenhofen airport, and (B) and (C) optimal segmentation with $\lambda = 10$, 858 regions and $\lambda = 30$, 325 regions, respectively.

and Salembier and Foucher (2016), extending in particular the speckle noise filtering proposed in Alonso-González et al. (2012; 2013). More specifically, Salembier (2015) defined the following energy functional:

$$\begin{aligned}\mathcal{E}_\lambda^{\mathrm{SAR}}(\pi) &= \sum_{\pi\in\mathcal{R}} (|\mathcal{R}|\Phi(\mathcal{R}) + \lambda)\\ &= \sum_{\pi\in\mathcal{R}}\left(\sum_{x\in\mathcal{R}} \frac{\| \widehat{\mathbf{C}}_x - \widehat{\mathbf{C}}_{\mathcal{R}}\|_F^2}{\| \widehat{\mathbf{C}}_{\mathcal{R}}\|_F^2}\right) + \lambda|\pi|\end{aligned} \tag{49}$$

where $\Phi(\mathcal{R})$ is the homogeneity criterion (Eq. 48) that was proposed by Alonso-González et al. (2012; 2013) for speckle noise filtering. As energy (Eq. 49) is separable (i.e., composed by a sum rule), it can further be minimized through the dynamic program Eqs. (25) and (26). In addition, as its regularization term is subadditive, this ensures the various optimal cuts $\pi_\lambda^\star$ to be ordered by refinement $(0 \le \lambda_1 \le \lambda_2 \Rightarrow \pi_{\lambda_1}^\star \le \pi_{\lambda_2}^\star)$. Finding the minimum of energy $\mathcal{E}_\lambda^{\mathrm{SAR}}$ as defined by (Eq. 49) can actually be seen as the unconstrained version of the following minimization problem

$$\min_{\pi\in\Pi_E(H)} \sum_{\pi\in\mathcal{R}} |\mathcal{R}|\Phi(\mathcal{R}) \quad \text{such that } |\pi| \le N^* \tag{50}$$

which translates into finding the partition whose number of regions is not more than $N^*$ and whose such regions are as homogeneous as possible. Therefore, it defines another speckle filtering strategy, where the degree of filtering is no longer controlled by some parameter $\delta_p$ (see the previous section "**Speckle noise filtering**"), but rather by a target number of regions $N^*$, which may be easier to get an intuition of and thus facilitate the tuning of this parameter.

This energy minimization framework was investigated by Salembier (2015) and Salembier and Foucher (2016) over the L-band PolSAR data acquired in 2003 by the Deutsches Zentrum für Luftund Raumfahrt (DLR) ESAR system over the area of the Oberpfaffen airport near Munich, Germany. The image, displayed by **Fig. 20A**, enjoys a spatial resolution of 1.5 m. The initial partition is obtained with the SLIC superpixels generation method (Achanta et al., 2012), applied after a denoising step using the $\sigma$-Lee filter (Lee et al., 2009). The BPT is built using the estimated covariance matrix region model (Eq. 12) and the geodesic distance merging criterion (Eq. 15). **Fig. 20B–C** show two optimal partitions, extracted from the BPT with $\lambda = 10$ and $\lambda = 30$, respectively. The former is composed of 858 different regions, while the latter comprises 325 regions, confirming that the larger the $\lambda$ value, the coarser the optimal partition. In both cases, large and homogeneous regions appear well filtered, while local details remain conserved, as it is the case for the five corner reflectors located at the center of the scene. The successful application of the hierarchical energy minimization framework for the segmentation of SAR images also corroborates the conclusions that were drawn in section "**Hyperspectral Image Segmentation**," namely, the high tunability of this segmentation framework through an appropriate definition for the energy function.

## 2.05.5 Conclusion

In conclusion, this article dealt with an extensive review of the recent advances in utilization of hierarchical representations in remote sensing data analysis. In a first instance, common hierarchical representations were introduced as tools borrowed from the field of mathematical morphology. This overview led to classify such structures as tree-based representations or hierarchies of partitions, the latter category being a special case of the former one. For the more general tree-based representations, the image is divided into a set of nonoverlapping regions such that any two regions are either disjoint or nested. Min-trees and max-trees (also known as the component trees) along with trees of shapes (also called inclusion trees) are common instances of such tree-based representations. Those rely on the absolute order holding on the pixels scalar values and depict the inclusion relationships between the various connected components composing the image. Thus, they are particularly suited for gray-scale image analysis, but their extension to multichannels images remains an active field of research, despite some recent very promising results. The use of tree-based representation structures for remote sensing data analysis has already been investigated and reviewed in the literature, but they were consequently put aside in this article. More specifically, the attention was turned to hierarchies of partitions, as a special case of tree-based representations. In addition to any pair of regions being either disjoint or nested, hierarchies of partitions require in their definition that any node (expect for the leaf ones) is fully recovered as the union of all its children. For this reason, hierarchies of partitions can alternatively (but equivalently) be defined as a collection of partitions of the image support space, ordered by refinement (the ordering holding on the space of partitions). Popular hierarchies of partitions include the $\alpha$-trees and the BPT. This article deeply focused on the latter as both its construction and processing are highly flexible and make BPT structures highly tunable for a wide range of applications. In particular, BPTs have received an increasing attention lately for remote sensing data analysis. In order to review those applications, the key concepts related to the BPT construction and its further processing were introduced. Specifically, various region models and their associated merging criteria were described, as well as possible choices for the initial partition, providing the user with a large range of configurations to parameterize the BPT construction, which is intended to be as much application independent as possible. Its processing, on the other hand, has to be adapted to the underlying goal, and a very general scheme was purposely initially presented. Finally, a detailed review of the most recent works involving BPTs for remote sensing data analysis tasks was presented. The covered topics included hyperspectral and SAR image segmentation, hyperspectral image classification, object recognition in hyperspectral urban environment scenes, anomaly detection, and SAR speckle noise filtering. Each topic comprised the accurate description of one or several related works, going for the general problem statement to a detailed, step-by-step description of the implemented BPT processing method as well as its rationale with respect to the task to achieve. Selected results from the cited references were presented to support the soundness of the described works.

## References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*, 2274–2282.

Akcay, H. G., & Aksoy, S. (2008). Automatic detection of geospatial objects using multiple hierarchical segmentations. *IEEE Transactions on Geoscience and Remote Sensing, 46*, 2097–2111.

Alonso-González, A., López-Martínez, C., & Salembier, P. (2012). Filtering and segmentation of polarimetric SAR data based on binary partition trees. *IEEE Transactions on Geoscience and Remote Sensing, 50*, 593–605.

Alonso-González, A., Valero, S., Chanussot, J., López-Martínez, C., & Salembier, P. (2013). Processing multidimensional SAR and hyperspectral images with binary partition tree. *Proceedings of the IEEE, 101*, 723–747.

Anderson, T. W. (2000). *An introduction to multivariate statistical analysis* (vol. 2). New York: Wiley.

Bai, Y., Dong, L., Huang, X., Yang, W. and Liao, M. (2014). Hierarchical segmentation of polarimetric SAR image via non-parametric graph entropy. In: *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pp. 2786–2789. IEEE.

Beaulieu, J. M., & Goldberg, M. (1989). Hierarchy in picture segmentation: a stepwise optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*, 150–163.

Bilgin, G., Erturk, S., & Yildirim, T. (2011). Segmentation of hyperspectral images via subtractive clustering and cluster validation using one-class support vector machines. *IEEE Transactions on Geoscience and Remote Sensing, 49*, 2936–2944.

Bioucas-Dias, J., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., & Chanussot, J. (2012). Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5*, 354–379.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing, 65*, 2–16.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 1222–1239.

Bruzzone, L., Chi, M., & Marconcini, M. (2006). A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 44*, 3363–3373.

Calderero, F., & Marques, F. (2010). Region merging techniques using information theory statistical measures. *IEEE Transactions on Image Processing, 19*, 1567–1586.

Cardelino, J., Randall, G., Bertalmio, M. and Caselles, V. (2006). Region based segmentation using the tree of shapes. In: *Image Processing, 2006 IEEE International Conference on*, pp. 2421–2424. IEEE.

Carlinet, E., & Géraud, T. (2014). A comparative review of component tree computation algorithms. *IEEE Transactions on Image Processing, 23*, 3885–3895.

Carlinet, E., & Géraud, T. (2015). MToS: a tree of shapes for multivariate images. *IEEE Transactions on Image Processing, 24*, 5330–5342.

Cavallaro, G., Dalla Mura, M., & Benediktsson, J. A. (2015). Analyzing remote sensing images with hierarchical morphological representations. In *Handbook of pattern recognition and computer vision* (pp. 313–330). Singapore: World Scientific.

Cloude, S., & Pottier, E. (1996). A review of target decomposition theorem in radar polarimetry. *IEEE Transactions on Geoscience and Remote Sensing, 34*, 498–518.

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 603–619.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Cox, T. F., & Cox, M. A. A. (2000). *Multidimensional scaling*. Boca Raton: CRC Press.

Dalla Mura, M., Benediktsson, J. A., Waske, B., & Bruzzone, L. (2010). Morphological attribute profiles for the analysis of very high resolution images. *IEEE transactions on Geoscience and Remote Sensing, 48*, 3747–3762.

Dalla Mura, M., Benediktsson, J. A., & Bruzzone, L. (2011). Self-dual attribute profiles for the analysis of remote sensing images. In *Mathematical morphology and its applications to signal and image processing* (pp. 320–330). Heidelberg: Springer.

Fauvel, M., Chanussot, J., & Benediktsson, J. A. (2012). A spatial–spectral kernel-based approach for the classification of remote-sensing images. *Pattern Recognition, 45*, 381–392.

Fauvel, M., Tarabalka, Y., Benediktsson, J. A., Chanussot, J., & Tilton, J. C. (2013). Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE, 101*, 652–675.

Finkel, R. A., & Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta Informatica, 4*, 1–9.

Frost, V. S., Stiles, J. A., Shanmugan, K. S., & Holtzman, J. C. (1982). A model for radar images and its application to adaptive digital filtering of multiplicative noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 157–166.

Garrido, L., Salembier, P., & Garcia, D. (1998). Extensive operators in partition lattices for image sequence analysis. *Signal Processing, 66*, 157–180.

Goodman, J. W. (1976). Some fundamental properties of speckle. *Journal of the Optical Society of America, 66*, 1145–1150.

Gu, Y., Zhang, Y., & Zhang, J. (2008). Integration of spatial–spectral information for resolution enhancement in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing, 46*, 1347–1358.

Guigues, L., Cocquerez, J. P., & Le Men, H. (2006). Scale-sets image analysis. *International Journal of Computer Vision, 68*, 289–317.

Jones, R. (1999). Connected filtering and segmentation using component trees. *Computer Vision and Image Understanding, 75*, 215–228.

Kiran, B. R., & Serra, J. (2014). Global–local optimizations by hierarchical cuts and climbing energies. *Pattern Recognition, 47*, 12–24.

Lang, S. (2008). Object-based image analysis for remote sensing applications: Modeling reality-dealing with complexity. In *Object-based image analysis* (pp. 3–27). Berlin: Springer.

Lee, J. S., Hoppel, K. W., Mango, S. A., & Miller, A. R. (1994). Intensity and phase statistics of multilook polarimetric and interferometric SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing, 32*, 1017–1028.

Lee, J. S., Grunes, M. R., & De Grandi, G. (1999). Polarimetric SAR speckle filtering and its implication for classification. *IEEE Transactions on Geoscience and Remote Sensing, 37*, 2363–2373.

Lee, J. S., Wen, J. H., Ainsworth, T. L., Chen, K. S., & Chen, A. J. (2009). Improved sigma filter for speckle filtering of SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing, 47*, 202–213.

Li, S. Z. (2009). *Markov random field modeling in image analysis* (Vol. 26). London: Springer.

Li, J., Bioucas-Dias, J. M., & Plaza, A. (2012). Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing, 50*, 809–823.

Ling, H. and Okada, K. (2006). Diffusion distance for histogram comparison. In: *Computer Vision and Pattern Recognition, 2006 I.E. Computer Society Conference on*, pp. 246–253. IEEE.

Malisiewicz, T. and Efros, A. A. (2007). Improving spatial support for objects via multiple segmentations. In: *British Machine Vision Conference (BMVC)*. pp. 1–10.

Manolakis, D., Truslow, E., Pieper, M., Cooley, T., & Brueggeman, M. (2014). Detection algorithms in hyperspectral imaging systems: an overview of practical algorithms. *IEEE Signal Processing Magazine, 31*, 24–33.

Meyer, F., & Maragos, P. (2000). Nonlinear scale-space representation with morphological levelings. *Journal of Visual Communication and Image Representation, 11*, 245–265.

Moakher, M., & Zéraï, M. (2011). The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision, 40*, 171–187.

Monasse, P., & Guichard, F. (2000). Fast computation of a contrast-invariant image representation. *IEEE Transactions on Image Processing, 9*, 860–872.

Mumford, D., & Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics, 42*, 577–685.

Najman, L., & Cousty, J. (2014). A graph-based mathematical morphology reader. *Pattern Recognition Letters, 47*, 3–17.

Nasrabadi, N. M. (2014). Hyperspectral target detection: an overview of current and future challenges. *IEEE Signal Processing Magazine, 31*, 34–44.

Noyel, G., Angulo, J., & Jeulin, D. (2007). Morphological segmentation of hyperspectral images. *Image Analysis and Stereology, 26*, 101–109.

Palou, G. and Salembier, P. (2013). Hierarchical video representation with trajectory binary partition tree. In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2099–2106. IEEE.

Pan, Y., Birdwell, J. D., & Djouadi, S. M. (2009). Preferential image segmentation using trees of shapes. *IEEE Transactions on Image Processing, 18*, 854–866.

Reed, I. S., & Yu, X. (1990). Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing, 38*, 1760–1770.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis.* (Vol. 709). Hoboken: John Wiley & Sons.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321–352). New York: Springer.

Russell, B. C., Freeman, W. T., Efros, A., Sivic, J., Zisserman, A., et al. (2006). Using multiple segmentations to discover objects and their extent in image collections. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pp. 1605–1614. IEEE.

Salembier, P. (2015). Study of binary partition tree pruning techniques for polarimetric SAR images. In: *International Symposium on Mathematical Morphology and Its Applications to Signal and Image Processing*, pp. 51–62. Springer International Publishing: Springer.

Salembier, P., & Foucher, S. (2016). Optimum graph cuts for pruning binary partition trees of polarimetric SAR images. *IEEE Transactions on Geoscience and Remote Sensing, 54*, 5493–5502.

Salembier, P., & Garrido, L. (2000). Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Transactions on Image Processing, 9*, 561–576.

Salembier, P., & Serra, J. (1995). Flat zones filtering, connected operators, and filters by reconstruction. *IEEE Transactions on Image Processing, 4*, 1153–1160.

Salembier, P., & Wilkinson, M. H. (2009). Connected operators. *IEEE Signal Processing Magazine, 26*, 136–157.

Salembier, P., Oliveras, A., & Garrido, L. (1998). Antiextensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing, 7*, 555–570.

Salerno, O., Pardàs, M., Vilaplana, V. and Marqués, F. (2004). Object recognition based on binary partition trees. In: *Image Processing, 2004. ICIP'04. 2004 International Conference on*, pp. 929–932. IEEE.

Samet, H. (1984). The quad-tree and related hierarchical data structures. *ACM Computing Surveys (CSUR), 16*, 187–260.

Shusterman, E., & Feder, M. (1994). Image compression via improved quad-tree decomposition algorithms. *IEEE Transactions on Image Processing, 3*, 207–215.

Snyder, D., Kerekes, J., Fairweather, I., Crabtree, R., Shive, J. and Hager, S. (2008). Development of a web-based application to evaluate target finding algorithms. In: *IGARSS 2008-2008 I.E. International Geoscience and Remote Sensing Symposium*, pp. 915–918. IEEE.

Soille, P. (2008). Constrained connectivity for hierarchical image partitioning and simplification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, 1132–1145.

Spann, M., & Wilson, R. (1985). A quad-tree approach to image segmentation which combines statistical and spatial information. *Pattern Recognition, 18*, 257–269.

Stein, D. W., Beaven, S. G., Hoff, L. E., Winter, E. M., Schaum, A. P., & Stocker, A. D. (2002). Anomaly detection from hyperspectral imagery. *IEEE Signal Processing Magazine, 19*, 58–69.

Tarabalka, Y., Benediktsson, J. A., & Chanussot, J. (2009). Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques. *IEEE Transactions on Geoscience and Remote Sensing, 47*, 2973–2987.

Tarabalka, Y., Chanussot, J., & Benediktsson, J. (2010). Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics, 40*, 1267–1279.

Tarabalka, Y., Chanussot, J., & Benediktsson, J. A. (2010). Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition, 43*, 2367–2379.

Tarabalka, Y., Fauvel, M., Chanussot, J., & Benediktsson, J. A. (2010). SVM- and MRF-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters, 7*, 736–740.

Tilton, J. C. (1999). A recursive PVM implementation of an image segmentation algorithm with performance results comparing the HIVE and the Cray T3E. In: *Frontiers of Massively Parallel Computation, 1999. Frontiers' 99. The Seventh Symposium on the*, pp. 146–153. IEEE.

Tilton, J. C. (2003). Analysis of hierarchically related image segmentations. In: *Advances in Techniques for Analysis of Remotely Sensed Data, 2003 I.E. Workshop on*, pp. 60–69. IEEE.

Tilton, J. C. (2010). *Split-remerge method for eliminating processing window artifacts in recursive hierarchical segmentation.* US Patent 7,697,759.

Tochon, G., Féret, J., Valero, S., Martin, R., Knapp, D., Salembier, P., Chanussot, J., & Asner, G. (2015). On the use of binary partition trees for the tree crown segmentation of tropical rainforest hyperspectral images. *Remote Sensing of Environment, 159*, 318–331.

Tochon, G., Drumetz, L., Veganzones, M. A., Dalla Mura, M. and Chanussot, J. (2016). From local to global unmixing of hyperspectral images to reveal spectral variability. In: *8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2016).*

Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In: *Computer Vision, 1998. Sixth International Conference on, IEEE*. pp. 839–846.

Valero, S. (2011). *Hyperspectral image processing and representation using Binary Partition Trees.* Ph.D. thesis, Gipsa-Lab, Department of Images and Signals, Grenoble Institute of Technology, Grenoble.

Valero, S., Salembier, P. and Chanussot, J. (2010). Comparison of merging orders and pruning strategies for binary partition tree in hyperspectral data. In: *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 2565–2568. IEEE.

Valero, S., Salembier, P., & Chanussot, J. (2013). Hyperspectral image representation and processing with binary partition trees. *IEEE Transactions on Image Processing, 22*, 1430–1443.

Valero, S., Salembier, P., & Chanussot, J. (2015). Object recognition in hyperspectral images using Binary Partition Tree representation. *Pattern Recognition Letters, 56*, 45–51.

Vasile, G., Trouvé, E., Lee, J., & Buzuloiu, V. (2006). Intensity-driven adaptive-neighborhood technique for polarimetric and interferometric SAR parameters estimation. *IEEE Transactions on Geoscience and Remote Sensing, 44*, 1609–1621.

Veganzones, M. A., Tochon, G., Dalla-Mura, M., Plaza, A. J., & Chanussot, J. (2014a). Hyperspectral image segmentation using a new spectral unmixing-based binary partition tree representation. *IEEE Transactions on Image Processing, 23*, 3574–3589.

Veganzones, M. A., Frontera-Pons, J., Pascal, F., Ovarlez, J. P. and Chanussot, J. (2014a). Binary partition trees-based robust adaptive hyperspectral RX anomaly detection. In: *2014b IEEE. International Conference on Image Processing (ICIP)*, pp. 5077–5081. IEEE.

Veganzones, M. A., Simoes, M., Licciardi, G., Yokoya, N., Bioucas-Dias, J. M., & Chanussot, J. (2016). Hyperspectral super-resolution of locally low rank images from complementary multisource data. *IEEE Transactions on Image Processing, 25*, 274–288.

Vilaplana, V., Marques, F., & Salembier, P. (2008). Binary partition trees for object detection. *IEEE Transactions on Image Processing, 17*, 2201–2216.

Vincent, L., & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*, 583–598.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision, 57*, 137–154.

Xie, H., Pierce, L. E., & Ulaby, F. T. (2002). SAR speckle reduction using wavelet denoising and Markov random field modeling. *IEEE Transactions on Geoscience and Remote Sensing, 40*, 2196–2212.

Xu, Y., Géraud, T. and Najman, L. (2012a). Context-based energy estimator: Application to object segmentation on the tree of shapes. In: *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pp. 1577–1580. IEEE.

Xu, Y., Géraud, T. and Najman, L. (2012b). Morphological filtering in shape spaces: Applications using tree-based image representations. In: *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 485–488. IEEE.

Xu, Y., Carlinet, E., Géraud, T., & Najman, L. (2017). Hierarchical segmentation using tree-based shape spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 457–469.

Yokoya, N., Grohnfeldt, C. and Chanussot, J. (2016). Hyperspectral and multispectral data fusion: a comparative review. *Geoscience and Remote Sensing Magazine*, IEEE (submitted).