Algorithm Theoretical Basis Document

CHRIS/PROBA
Cloud Screening Module

Version 2.0, April 2008

Luis Gómez-Chova[1], Luis Guanter[2], Luis Alonso[1], Javier Calpe[1], Jose Moreno[1]

[1] *University of Valencia, Dr. Moliner 50, 46100, Burjassot–Valencia (Spain)*

[2] *GeoForschungsZentrum Potsdam, Telegrafenberg 14473, Potsdam (Germany)*

Contents

Abstract

This document describes the theoretical basis of the modules developed for the BEAM toolbox to mask clouds present in images acquired by the Compact High Resolution Imaging Spectrometer (CHRIS).

Accurate and automatic detection of clouds in satellite scenes is a key issue for a wide range of remote sensing applications. With no accurate cloud masking, undetected clouds are one of the most significant source of error in both sea and land cover biophysical parameter retrieval. The objective of the algorithms presented in this document is to detect clouds accurately providing probability and cloud abundance rather than binary flags. For this purpose, the method exploits the special characteristics of the CHRIS/PROBA instrument, such as the high number of spectral bands, the spatial resolution, and the specific band locations.

Acronyms and Abbreviations

ATBD Algorithm Theoretical Basis Document

ATGP Automated Target Generation Process

BEAM Basic ERS & Envisat (A)ATSR and MERIS

CCD Charge Coupled Device

CHRIS Compact High Resolution Imaging Spectrometer

CWV Columnar Water Vapor

DEM Digital Elevation Model

DOY Day Of Year

EM Expectation–Maximization

ESA European Space Agency

FCLSU Fully Constrained Linear Spectral Unmixing

FZA Fly-by Zenith Angle

HDF Hierarchical Data Format

LSU Linear Spectral Unmixing

MAP Maximum a posteriori Probability

MDL Minimum Description Length criterion

NDVI Normalized Difference Vegetation Index

NIR Near–infraRed

PROBA PRoject for On-Board Autonomy

ROI Regions Of Interest

SZA Sun Zenith Angle

TOA Top Of Atmosphere

VIS Visible

VISAT Visualisation and Analysis Tool

VNIR Visible and Near InfraRed

VZA View Zenith Angle

1   Introduction

In remote sensing images acquired by instruments working in the visible and near-infrared range (VNIR) of the electromagnetic spectrum, undetected clouds are the most significant source of error for true ground reflectance estimation (Simpson, 1999). Therefore, the operational use of these images can be hampered by the presence of clouds.

The simplest approach to cloud detection in a scene is the use of a set of static thresholds (reflectance, temperature) applied to every pixel in the image, which provides a cloud flag (binary classification). These methods can fail for several reasons, such as subpixel clouds, high reflectance surfaces, illumination and observation geometry, sensor calibration, variation of the spectral response of clouds with cloud type and height, transparency of clouds, etc. Therefore, cloud screening approaches heavily depend on the characteristics of each sensor. Obviously, its spectral and spatial resolution, or the spectral range are critical. In particular, the Compact High Resolution Imaging Spectrometer (CHRIS) (Barnsley et al., 2004) provides us with a number of channels in the visible (VIS) and near infrared (NIR) range with enough spectral resolution to increase the cloud detection accuracy (Diner et al., 1999): it covers a spectral range from 400 nm to 1050 nm with a maximum spatial resolution of 17 or 34 m at nadir depending on the acquisition mode (more information on http://earth.esa.int/proba/).

CHRIS is mounted on board the European Space Agency (ESA) small satellite platform called PROBA (*Project for On Board Autonomy*). Thanks to the PROBA platform pointing capabilities, the acquisition plan tries to avoid acquisitions with cloud coverage, but occasionally images are partially affected by clouds. In these cases, users that requested the acquisition have a special interest on both cloud location and contribution to the registered spectra.

In this context, the main objective of this module is to develop and validate a method for cloud detection using self-contained information provided by CHRIS data. The method should be capable of: detecting clouds accurately providing probability and cloud abundance instead of flags. For this purpose, the method exploits the characteristics of CHRIS, such as the high number of spectral bands, the spatial resolution, and the specific band locations. Basically, it evaluates the information coming from the spectral signature and from several specific features characterizing clouds in order to generate a probabilistic cloud mask, which indicates the cloud abundance per pixel. Later it can be converted into a binary cloud mask by setting thresholds that can be tuned depending on the application. As a consequence, by masking only the image areas affected by cloud covers, the whole image is not necessarily discarded, thus making multitemporal and multiangular studies possible.

2   Cloud Screening Algorithm

In this section, we present a cloud detection procedure which is constituted by the following steps (Fig. 1):

1. *Image Pre-Processing*: a pre-processing stage of the CHRIS data to correct illumination effects is necessary for their proper analysis.

2. *Feature extraction*: physically-inspired features are extracted to increase separability of clouds and surface.

3. *Image clustering*: an unsupervised clustering is performed on the extracted features in order to separate clouds from the ground-cover.

4. *Cluster labeling*: resulting clusters are subsequently labeled into geo-physical classes (identification of cloud clusters).

5. *Spectral unmixing*: a spectral unmixing is applied to the segmented image in order to obtain an abundance map of the cloud content in the cloud pixels.

In the following, we analyze each of these components in detail.

2.1   Image Pre-Processing

A multispectral image consists of two spatial dimensions (along-track and across-track) and one spectral dimension (wavelength). The image is registered by the instrument in a data-cube where the along-track dimension, y, corresponds to the image lines; the across-track dimension, x, is associated to the pixel line; and the spectral dimension, λ, represents the image bands. The
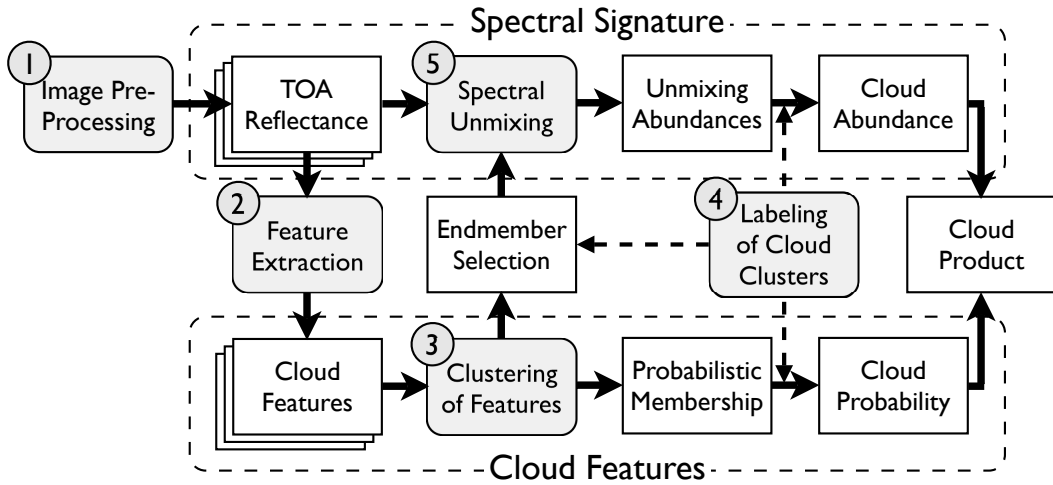

Figure 1: Scheme of the cloud screening algorithm constituted by the preprocessing (TOA reflectance derivation) and four main processing steps (grey boxes).

size of the multispectral data-cube can be written in the form $l \times p \times b$, where l is the number of image lines, p is the number of pixels per line, and b is the number of spectral bands.

CHRIS products are provided in *top of the atmosphere* (TOA) radiance (radiometrically calibrated data), and each pixel k is defined by $\{L_k(\lambda)\}_{k=1}^n$, where the number of pixels is $n = l \times p$ and $\mathbf{L}_k \in \mathbb{R}^b$ is the spectral signature sampled at b spectral bands of the VNIR spectral region $\{\lambda_i\}_{i=1}^b \in [400 - 1050]$, which location varies depending on the acquisition mode (see Fig. 3). The cloud screening module requires as input the CHRIS image corrected of noises (see noise reduction module ATBD for more details (Gómez-Chova, 2007)): *drop-outs* cannot be processed since they present wrong values in some bands, and the *vertical stripping* introduces differences between bands that affect features extracted from the spectra. This corrected radiance must be pre-processed in order to estimate TOA reflectance. This allows us to remove in practice the dependence on particular illumination conditions (day of the year and angular configuration) and illumination effects due to rough terrain (cosine correction), since the method is intended to work under many situations. TOA apparent reflectance is estimated according to:

$$\rho(x, y, \lambda_i) = \frac{\pi L(x, y, \lambda_i)}{\cos(\theta(x, y)) I(\lambda_i)}, \tag{2.1}$$

where $L(x, y, \lambda)$ is the provided at-sensor upward radiance at the image location (x, y), $I(\lambda)$ is the extraterrestrial instantaneous solar irradiance, and $\theta(x, y)$ is the angle between the illumination direction and the vector perpendicular to the surface. In the proposed algorithm, $\theta(x, y)$ is approximated by the *Solar Zenith Angle* provided in the CHRIS HDF file attributes since one can assume flat landscape and constant illumination angle for the area observed in a CHRIS image. Finally, the Sun irradiance, $I(\lambda)$, is taken from Thuillier et al. (2003), corrected for the acquisition day, and convolved with the CHRIS spectral channels (Fig.2).

The extraterrestrial solar irradiance $I(\lambda)$ given in Thuillier et al. (2003) is provided from 200 to 2400 nm in mW/m^2/nm. It shall be corrected for the Julian day of year (DOY), J, according to the following approximate formulae:

$$I(\lambda) = \frac{1}{(1 - 0.01673 \cos(0.9856(J - 4)\pi/180))^2} I(\lambda), \tag{2.2}$$

where the day of year can be easily obtained from the *Image Date* metadata of the CHRIS HDF file.

Since the reference extraterrestrial solar irradiance presents a different spectral sampling, it is resampled to the CHRIS spectral channels. A specific CHRIS band, i, consists of the addition of one or more CCD detector pixel elements depending on the band width. Therefore, the spectral response of a CHRIS band, $S_i(\lambda)$, is the sum of the the spectral response $S(\lambda)$ of the corresponding detectors of the CCD array. Then, the mean solar irradiance for a given band, $I(\lambda_i)$, is obtained by integrating the extraterrestrial solar irradiance by its spectral response:

$$I(\lambda_i) = \frac{\int_0^\infty S_i(\lambda) I(\lambda) d\lambda}{\int_0^\infty S_i(\lambda) d\lambda}. \tag{2.3}$$

The theoretical half-width of the instrument line-spread functions correspond to spectral resolutions of 1.25 nm at 415 nm, increasing to 11.25 nm at 1050 nm. One could assume a Gaussian
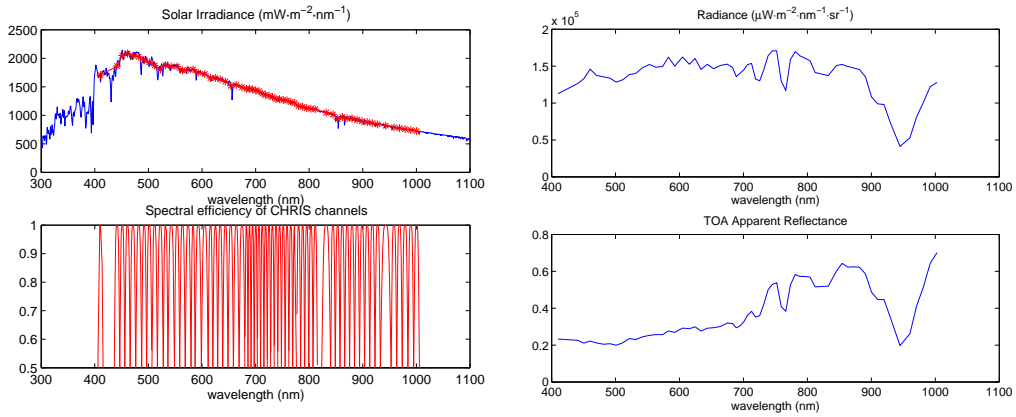
Figure 2: Sun irradiance corrected for the day of the year and the CHRIS channels (left) and TOA apparent reflectance estimated from the at sensor radiance (right).

response function, $S(\lambda)$, for each element using its full-width half-maximum (FWHM), and sum up all the elements of the band. However, in the header of the CHRIS HDF files, although the CCD row number for lower and upper wavelengths of each band is provided, the FWHM of each CCD element is not. Therefore, for the shake of simplicity, the spectral response of a CHRIS band, $S_i(\lambda)$, is defined as a *bell-shaped* function depending on the mid-wavelength, λ_i, and the band width, $\Delta\lambda_i$, of the band, directly:

$$S_i(\lambda) = \frac{1}{1 + |2(\lambda - \lambda_i)/\Delta\lambda_i|^4} \qquad \lambda_i - \Delta\lambda_i < \lambda < \lambda_i + \Delta\lambda_i, \qquad (2.4)$$

where both the mid-wavelength (*WlMid*) and the band width (*BWidth*) values for each channel are included in the CHRIS HDF file.

2.2   Feature Extraction

The measured spectral signature depends on the illumination, the atmosphere, and the surface. Figure 3 shows CHRIS Mode 1 band locations compared with the spectral curve of healthy vegetation, bare soil, and the atmospheric transmittance. The spectral bands free from atmospheric absorptions contain information about the surface reflectance, while others are mainly affected by the atmosphere.

At this step, rather than working with the spectral bands only, physically-inspired features are extracted in order to increase the separability of clouds and surface covers. These features are extracted independently from the bands that are free from strong gaseous absorptions, $\lambda_i \subset B_S$, and from the bands affected by the atmosphere, $\lambda_i \subset B_A$. A detailed analysis of the extracted features follows. For illustration purposes, only the extracted features from a single given CHRIS image (CHRIS-BR-050717-576C-41) are provided.
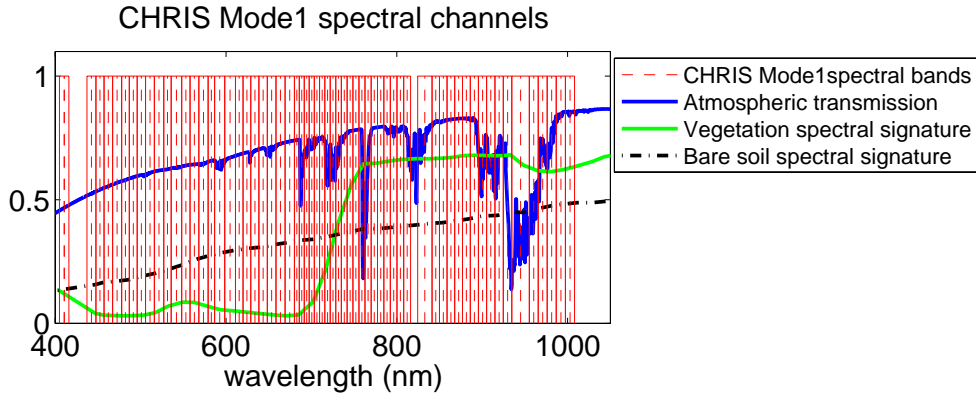
Figure 3: CHRIS Mode 1 band locations (boxes) superimposed to a reflectance spectra of healthy vegetation (dashed), bare soil (dash-dotted), and the atmospheric transmittance (solid).

2.2.1   Surface Features

Regarding the reflectance of the surface, one of the main characteristics of clouds is that they present *bright* and *white* spectra (Fig. 4). We can exploit CHRIS bands for extracting information about the target reflectance, i.e. cloud brightness and cloud whiteness for cloudy pixels:

- A *bright* spectrum means that the intensity of the spectral curve (related to the albedo) should present relatively high values. Therefore, cloud brightness is calculated for each pixel as the integral of spectrum, $f_{Br} = \int \rho(\lambda) \mathrm{d}\lambda$, which is approximated through trapezoidal numerical integration:

$$\hat{f}_{Br} = \frac{1}{\lambda_{max} - \lambda_{min}} \sum_{\lambda_i \subset B_S} \frac{\rho(\lambda_{i+1}) + \rho(\lambda_i)}{2} (\lambda_{i+1} - \lambda_i), \qquad (2.5)$$

  which has the same units as $\rho(\lambda)$, and differs from the average of the spectral bands since it takes into account the distribution of the energy along the spectrum.

- A *white* spectrum means that the spectral signature must be flat along the spectrum. The first derivative of the spectral curve should present low values, but noise and calibration errors may reduce the accuracy in the estimation of the spectrum flatness when computing the spectral derivative in bands with similar wavelengths. Therefore, we compute for each pixel the deviation from the flatness as the (trapezoidal approximate) integral of $e(\lambda) = |\rho(\lambda) - \hat{f}_{Br}|$:

$$f_{Wh} = \frac{1}{\lambda_{max} - \lambda_{min}} \sum_{\lambda_i \subset B_S} \frac{e(\lambda_{i+1}) + e(\lambda_i)}{2} (\lambda_{i+1} - \lambda_i) \qquad (2.6)$$

Further surface features can be obtained by considering independently the VIS ($\lambda_{VIS} \in [400 - 700]$ nm) and NIR ($\lambda_{NIR} \in [700 - 1000]$ nm) spectral ranges, where surface covers present different reflectance properties. Therefore, instead of working with f_{Br} and f_{Wh}, we can obtain $2 + 2$ features from (2.5) and (2.6) respectively: $f_{Br,VIS}$ and $f_{Wh,VIS}$, computed using $\lambda_i \subset (B_S \cap VIS)$; and $f_{Br,NIR}$ and $f_{Wh,NIR}$, computed using $\lambda_i \subset (B_S \cap NIR)$. For example, clouds over land should
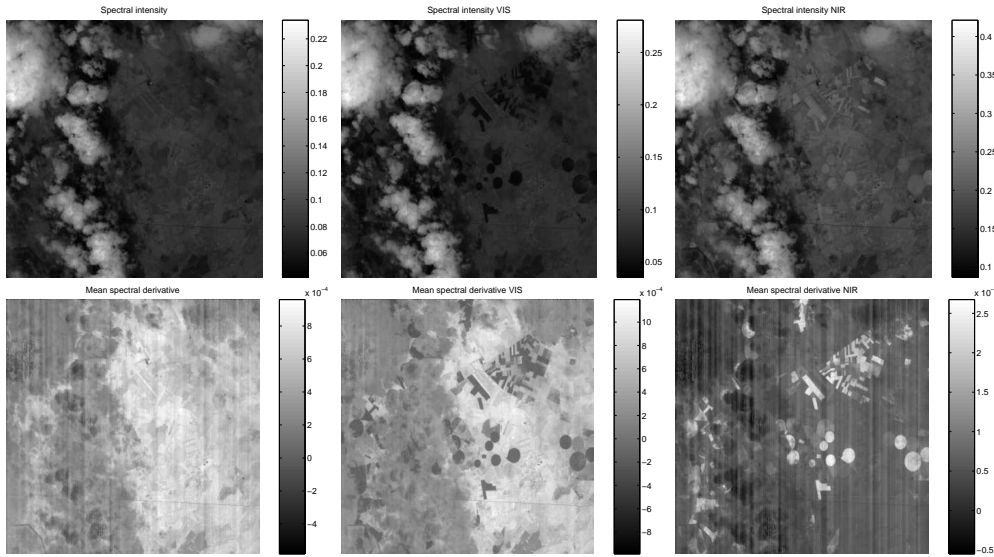
Figure 4: Cloud brightness (*top*) and whiteness (*bottom*) features extracted from the TOA reflectance of the CHRIS-BR-050717-576C-41 image.

be better recognized in $f_{Br,VIS}$ than in $f_{Br,NIR}$ since land covers have less reflectance in the VIS range, while the opposite is true for clouds over sea.

2.2.2 Atmospheric Features

Regarding the atmospheric absorptions, another meaningful feature is the fact that clouds are at a higher altitude than the surface. It is worth noting that atmospheric absorption depends on the atmospheric constituents and the optical path. Since light reflected on high clouds crosses a shorter section of the atmosphere, the consequence would be an abnormally short optical path, thus weaker atmospheric absorption features. Atmospheric oxygen absorption and even water vapor absorption (at 760 nm and 940 nm respectively) are candidate bands to be used in the optical path estimation.

The use of atmospheric absorption in the oxygen-A band to infer cloud pressure, which is related to cloud-top height, has been suggested by several authors (Yamamoto and Wark, 1961; Chapman, 1962). In the case of medium resolution imaging spectrometers, several studies have shown that the oxygen-A band is potentially efficient for determining the cloud-top pressure (Fisher and Grassl, 1991; Buriez et al., 1997; Ramon et al., 2003). All these studies assume that the two spectral channels located at the oxygen-A band (one outside, λ_{out}, and another inside, λ_{in}, the absorption) allow the derivation of an apparent pressure which roughly represents the cloud pressure. In particular, apparent pressure is calculated using an empirical polynomial function of the oxygen transmission derived from the reflectance ratio $\rho(\lambda_{in})/\rho(\lambda_{out})$. However, to obtain reliable estimations of the cloud-top height is still a challenging problem affected by the instrument radiometric and spectral resolution, the influence of ground reflectance, and the need of a reliable surface pressure reference. These difficulties explain the little attention paid to this helpful feature
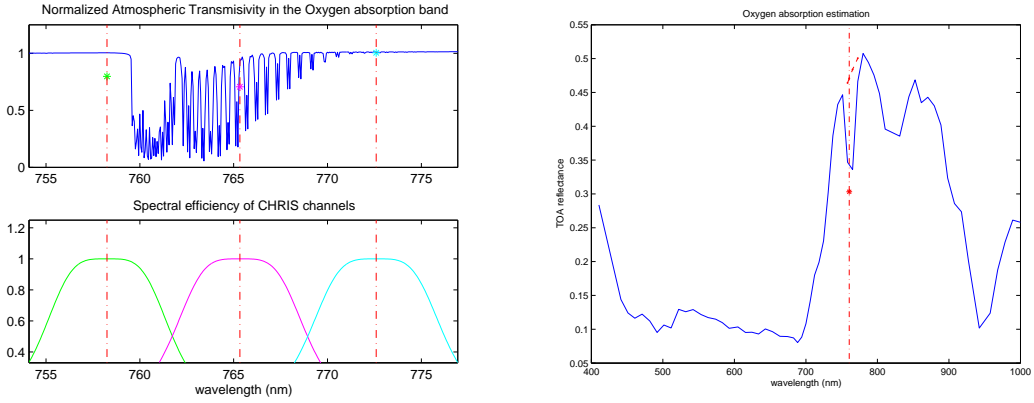
Figure 5: Estimation of the optical path, δ, from the O_2 absorption band. *Left*: Effective atmospheric vertical transmittance, $\exp(-\tau_{atm}(\lambda))$, estimated from a high resolution curve (*top*) taking into account the for the spectral response (efficiency) of CHRIS channels (*bottom*). *Right*: the interpolated, $\rho_0(\lambda)$, and measured, $\rho(\lambda)$, reflectance inside the oxygen band, and the estimated reflectance at the maximum absorption.

in cloud screening. In the case of CHRIS, Modes 1 and 5 present bands in the O_2-A absorption, which makes the inclusion of this atmospheric feature in the cloud screening scheme possible. In the following paragraphs, we show the formulation proposed to extract an atmospheric feature directly related with the optical path.

The light transmitted through a non-dispersive medium can be expressed using the Bouguer-Lambert-Beer law:

$$L(\lambda) = L_0(\lambda) \exp\left(-\frac{\tau(\lambda)}{\mu}\right), \tag{2.7}$$

where $L_0(\lambda)$ is the light entering into the medium, the term $\exp(-\tau(\lambda)/\mu)$ is the transmittance factor, $1/\mu$ is the optical mass obtained from the illumination and observation zenith angles, and $\tau(\lambda)$ is the atmospheric optical depth. Since most of the radiation measured by the sensor has been reflected by the surface, (2.7) can not be used to model the at-sensor radiance. However, it provides a physical basis for the definition of a non-dimensional parameter that accounts for atmospheric absorptions in typical remote sensing scenarios. In our case, the reference radiance $L_0(\lambda)$ will be the radiance outside the absorption feature, calculated by interpolating the nearby channels that are unaffected by absorptions, and $L(\lambda)$ will be the radiance affected by gaseous absorptions after crossing the TOA-surface-sensor path. The inversion of (2.7) provides $\tau(\lambda)$, which is a measure of the strength of the gaseous absorptions in a certain spectral range. The assumption is that variations in $\tau(\lambda)$ are driven by sharp changes in elevation as those due to transitions between cloud-free and cloud-covered areas. Horizontal variations in the atmospheric state are considered a second-order effect compared to cloud-to-surface elevation changes. An equivalent atmospheric transmittance parameter could be calculated as the ratio $L(\lambda)/L_0(\lambda)$, but the contribution of illumination and observation geometries would not be normalized.
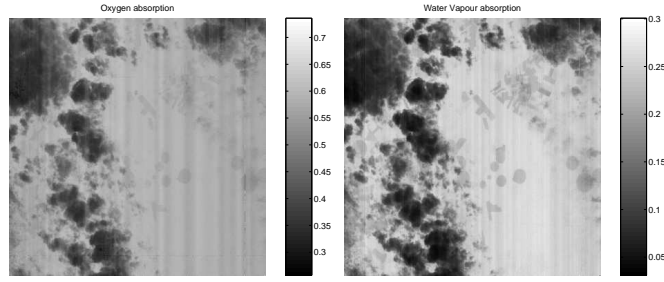
Figure 6: Estimation of the optical path from the oxygen absorption (*left*) and water vapor (*right*) bands for the CHRIS-BR-050717-576C-41image.

The atmospheric optical depth is decoupled into two contributions,

$$\tau(\lambda) = \tau_{atm}(\lambda) \cdot \delta \,, \tag{2.8}$$

where τ_{atm} is an optical thickness reference spectrum at sea level for nadir illumination and observation, and δ is a factor accounting for elevated surfaces such as clouds. The reference $\tau_{atm}(\lambda)$ values are estimated for the channels of the instrument from a high spectral resolution curve. The approach followed in this paper for the so-called Oxygen-A band can be devised from Fig. 5, and the extracted feature is derived from (2.7) and (2.8) as:

$$f_{O_2}(x,y) = -\frac{\mu(x,y)}{\tau_{atm}(\lambda_{in})}\ln\left(\frac{L(x,y,\lambda_{in})}{L_0(x,y,\lambda_{in})}\right), \tag{2.9}$$

where the interpolated radiance at the absorption band is estimated from nearby channels, $L_0(\lambda_{in}) = L(\lambda_{out-inf}) + (\lambda_{in} - \lambda_{out-inf})(L(\lambda_{out-sup}) - L(\lambda_{out-inf}))/(\lambda_{out-sup} - \lambda_{out-inf})$.

An additional estimation of the optical path can be obtained from the water vapor absorption in the NIR close to the end of the valid range of the sensor. In this case, the maximum water vapor absorption (940 nm) is also acquired by Modes 1 and 5 only. In addition, the water vapor distribution is extremely variable, thus it is not straightforward to relate this feature to the real altitude. However, it is still valid for relative measurements inside the same image since almost all the atmospheric water vapor is distributed in the first 2-3 km of the atmosphere below most of the cloud types. The same approach than in the O_2 case has been followed to obtain this feature (Fig. 6)

$$f_{WV}(x,y) = -\frac{\mu(x,y)}{\tau_{atm}(\lambda_{in}(x))}\ln\left(\frac{L(x,y,\lambda_{in}(x))}{L(x,y,\lambda_{out-inf}(x))}\right), \tag{2.10}$$

where we assume that $L_0(\lambda_{in}) = L(\lambda_{out-inf})$ since the interpolation at the end of the spectral range of the sensor is sometimes not possible.

It is worth noting that the extracted atmospherical features are not intended to estimate altitude of clouds. They are an estimation of the optical path by taking into account important issues such as the viewing geometry, atmospheric transmission, and sensor calibration. However, this estimation is affected by the background reflectance of the surface and the atmospheric conditions, which change from one image to another. These problems avoid the use of these features in standard approaches based on thresholds or static models.
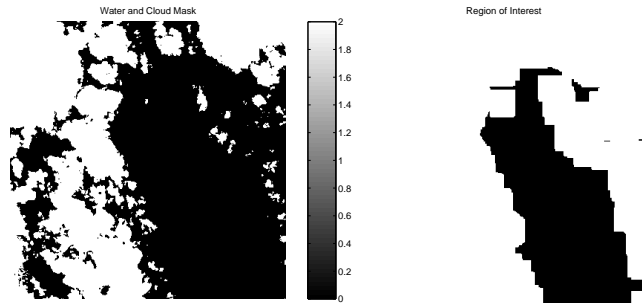
Figure 7: Result of the threshold-based cloud/land/water classification (left) and the growing algorithm (right) for the CHRIS-BR-050717-576C-41image (significant pixels in white).

2.3 Image Clustering

As previously discussed, static thresholds applied to every pixel in the image can fail due to subpixel clouds, sensor calibration, variation of the spectral response of clouds with cloud type and height, etc. In this context, the following step in our methodology considers the use of unsupervised classification methods to find groups of similar pixels in the image. Clustering methods assume that the input data is organized into a number of groups or clusters (Duda and Hart, 1973) according to a given distance measure in some representation space. We use an Expectation–Maximization (EM) algorithm (Dempster et al., 1977) to estimate the parameters of a Gaussian mixture model. The cluster algorithm is only applied to the regions of interest previously determined. The whole image clustering process is further described in the following subsections.

2.3.1 Regions of interest (ROIs)

Before applying a clustering algorithm, we should stress the fact that if clouds were not statistically representative in a given image, clustering methods could not find small clouds or could misclassify clouds as similar classes. Therefore, in addition to using representative features along with the spectral channels, clustering improves when applied over the regions of the image where clouds are statistically representative.

In order to find regions that could potentially contain clouds, we apply hard non-restrictive thresholds to provide a first map of cloud-like pixels. These absolute thresholds were obtained empirically and were applied to well-defined features: the brightness in the VIS and the NIR region, the estimated water vapor absorption, and the NDVI (in order to exclude areas with pure vegetation). Then, a region growing algorithm is carried out, along with a morphological process that dilates cloudy areas. This way, we ensure that all possible clouds and their contiguous areas will be considered in the clustering. The result of this process is far from providing a classification map, but just a *region of interest* (ROI), in which presence of clouds is significant for the purpose of clustering (Fig.7).

2.3.2   Clustering the ROIs

The clustering algorithm is applied to all the pixels in the ROI $\mathcal{X} = \{\mathbf{x}_k\}_{k=1}^n$, where $\mathbf{x}_k \in \mathbb{R}^d$ is the vector of extracted features for each pixel: $\mathbf{x}_k = \{f_{Br,VIS}, f_{Br,NIR}, f_{Wh}, f_{O_2}, f_{WV}\}$. Basically, the aim of the clustering is to associate each input $\mathbf{x}_k$ to one of the clusters ω_j, $j = 1\ldots c$, in order to separate different classes (or at least clouds and ground-cover) present in the scene. We impose the following requirements to the clustering method: 1) taking advantage of all available features (including atmospheric absorptions); 2) considering the full relationship among variables (without applying independent tests to each feature); and 3) providing for each input *soft* association with the clusters (membership or probability) value between zero and one, h_{kj}, with the requirement that the memberships sum to one.

2.3.3   Normal distributions and Expectation–Maximization algorithm

In multispectral image processing, the assumption that the distribution of images can be approximated as a mixture of normally distributed samples is commonly accepted; and we make the same assumption for the extracted features. Therefore, we consider the input as a mixture of normal distributions and use the EM algorithm to obtain the maximum likelihood estimation of the probability density function (pdf) of the Gaussian mixture (Duda and Hart, 1973). The EM algorithm is an iterative procedure that involves two consecutive iterative steps. In the E-step, we compute the posteriors (or membership) h_{kj} of the pixel k associated to the jth Gaussian component of the mixture. In the M-step, we use the obtained h_{kj} to update the mixture coefficient, α_j, the mean, μ_j, and the covariance matrix, Σ_j, for each component of the mixture:

$$\mu_j = \frac{\sum_k h_{kj}\mathbf{x}_k}{\sum_k h_{kj}}$$

$$\Sigma_j = \frac{\sum_k h_{kj}(\mathbf{x}_k - \mu_j)(\mathbf{x}_k - \mu_j)^T}{\sum_k h_{kj}} \tag{2.11}$$

$$\alpha_j = \frac{1}{n}\sum_k h_{kj}$$

The final pdf describes both the class of interest and the ground-cover class, and is worth noting that both heterogeneous classes can be made up of more than one Gaussian component, each representing a different subclass.

2.3.4   Clustering initialization

The clustering algorithm has to be started with initial values for the parameters of the pdf. The k-means algorithm is used to obtain a first approach to the structure of the data in clusters. This algorithm only needs the number of clusters c to be fixed, and minimizes the Euclidean distance of the samples in one cluster to its mean. In k-means, input is associated only with the cluster having the nearest center. The cluster center is the mean of all inputs associated with

that cluster. Once the cluster centers are updated, samples may change to a different cluster so an iterative procedure is followed until centers do not change.

2.3.5   MAP classification

Once we know the Gaussian components of the pdf of the data we perform a Gaussian maximum likelihood classification on the whole image. The algorithm assigns the pixel to the cluster with the maximum *a posteriori* probability (MAP) generating a map with the clusters in the image. The final estimates of the cluster membership for each pixel in the image h_{kj} represent the estimates of the posterior probabilities, which are used to compute the optimal cluster label as:

$$\hat{\omega}_k = \underset{j}{\mathrm{argmax}}\, \{h_{kj}\} \tag{2.12}$$

2.3.6   Remarks on the number of clusters

The proposed image clustering process relies on the key step of selecting the number of clusters, c. At this point, several statistical criteria can be used, such as minimizing the Davies–Bouldin index (Davies and Bouldin, 1979) or the Minimum Descriptive Length (MDL) criterion (Rissanen, 1989) for a number of initialization runs. In our experiments, however, we did not observe a critical behavior in this sense; note that even if a low number of clusters is selected, some of them should correspond to different cloud types, since the (overgrown) ROI is typically well-identified.

Another possibility for the user (not explored in this work) is to initialize the mean of the clusters with the spectral signature of the class of interest using a spectral library. Obviously, the problem of selecting c disappears if a training labeled set is available.

2.4   Cluster Labeling

Once clusters are determined in the previous step, the spectral signature of each cluster, $\mathbf{s}_j(\lambda)$, is estimated as the average of the spectra of the cluster pixels. This step excludes those pixels with abnormally low membership values or posterior probability h_{kj}.

It is important to emphasize that these spectral signatures of each cluster could differ a lot from the spectra obtained when applying the EM algorithm over the image using the spectral bands rather than the extracted features. The extracted features used to find the clusters are optimized to increase separability between clouds, C, and any-other surface type, $\overline{C}$, while in the spectral domain these clusters could present a high degree of overlapping. At this point of the process, the obtained clusters can be labeled into geo-physical classes taking into account four complementary sources of information (Fig. 8): the thematic map with the distribution of the clusters in the scene; the cluster parameters (μ_j, Σ_j) of the extracted features; the spectral signatures of cluster, $\mathbf{s}_j$; and the location in the image of the pixels with the spectral signature closer to $\mathbf{s}_j$. This information can be either analyzed directly by the user, compared to a spectral
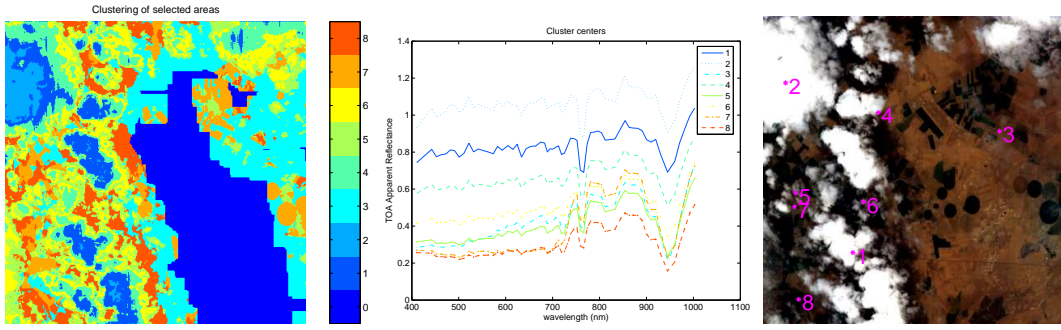
Figure 8: Thematic map with the distribution of the clusters in the scene (*left*), the spectral signatures of the clusters (*center*), and the location in the image of the pixels with the most similar spectra (*right*).

library with representative spectra of all the classes of interest, or applying a set of threshold tests.

During the labeling process, it is possible to reject a given cluster if it contains pixels corresponding to both clouds and ground covers. In this case, after removing this components from the mixture of normal distributions, we perform again a MAP classification on the whole image to obtain the final cluster membership for each pixel in the image, h_{kj}, and its corresponding cluster label $\hat{\omega}_k$.

Once all clusters have been related to a class with a geo-physical meaning (Fig. 9), it is straightforward to merge all the clusters belonging to a cloud type. Since the EM algorithm provides posterior probabilities ($h_{kj} \in [0,1]$ and $\sum_{j=1}^{c} h_{kj} = 1$), a probabilistic cloud index, based on the clustering of the extracted features, can be computed as the sum of the posteriors of the cloud-clusters C:

$$h_{kC} = \sum_{\omega_j \subset C} h_{kj} \tag{2.13}$$

However, if the clusters are well separated in the feature space, the posteriors decrease drastically from one to zero in the boundaries between clusters (Fig. 10). Therefore, this *Cloud Probability* index indicates the probability that one pixel belongs more to a cloud-cluster, C, than to one of the other clusters, $\overline{C}$, found in the image, but it does not give information about the cloud content at subpixel level, which is very important when dealing with thin clouds or partially covered pixels.

2.5   Spectral Unmixing

In order to obtain a cloud abundance map for every pixel in the image, rather than flags or a binary classification, a spectral unmixing algorithm is applied to the multispectral image. The linear spectral unmixing algorithm (LSU) (Chang, 2003) allows decomposing each pixel of the image, $\rho_k(\lambda)$, into a collection of constituent spectra or *endmembers*, and a set of corresponding abundances that indicate the proportion of each endmember in the pixel.
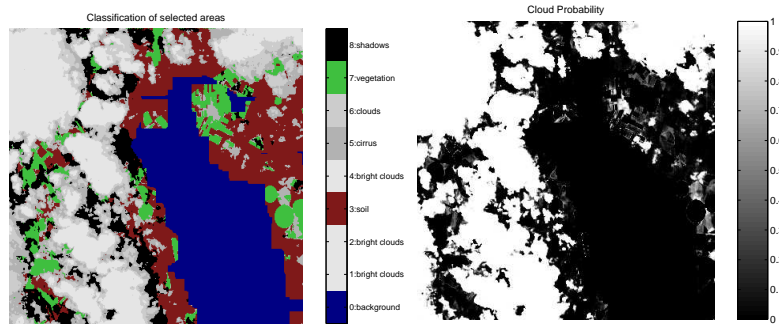
Figure 9: *Left*: Thematic map with the distribution in the scene of the classes of the clusters. *Right*: *Cloud Probability* index computed from the posteriors of the cloud-clusters.

2.5.1 LSU algorithm

The algorithm used to perform the spectral unmixing is the Fully Constrained Linear Spectral Unmixing (FCLSU) (Heinz and Chang, 2001), which guarantees a physical interpretation of the results and can be formalized as follows:

$$\rho_k(\lambda_i) = \sum_{q=1}^{Q} m_q(\lambda_i)a_{kq} + \varepsilon \qquad (2.14)$$

subject to

$$0 \leq a_{kq} \leq 1 \text{ and } \sum_q a_{kq} = 1 \qquad (2.15)$$

where $\rho_k(\lambda_i)$ is the value of the pixel k for the band i, Q represents the number of endmembers that are being unmixed, being the coefficients a_{kq} of this combination the unmixing coefficients, which can be interpreted as the abundance fractions of materials in a pixel. Finally, the term ε represents the residual error per band. Equation (2.14) can be expressed in a matrix form as $\rho_k = \mathbf{M} \cdot \mathbf{a}_k + \varepsilon$, where the spectral signatures of materials, $\mathbf{m}_q$, are expressed in the columns of matrix $\mathbf{M}$. The FCLSU algorithm solves a constrained linear least-squares problem minimizing the norm of $(\mathbf{M} \cdot \mathbf{a}_k - \rho_k)$ where the vector, $\mathbf{a}_k$, of independent variables is restricted to being nonnegative (since it represent the contribution of reflectance signatures $\mathbf{m}_q$) and sum to one (since it is supposed that $\mathbf{M}$ represents all the constituents in the image with at least one pure independent spectrum).

2.5.2 Remarks on endmember extraction for cloud screening

In the literature, there are different approaches to determine the spectra of the different pure constituents in the image (Keshava and Mustard, 2002; Plaza and Chang, 2006). However, in a cloud screening framework, two specific considerations have to be taken into account. First, only one endmember must be selected to represent clouds. This constrain contrasts with the selection of the number of clusters, in which more clusters should model better such an heterogeneous class as clouds. In the classification, most of the cloud-clusters consists of mixed pixels of thin clouds and ground or borders and subpixel clouds. In the LSU method, we assume that clouds represent
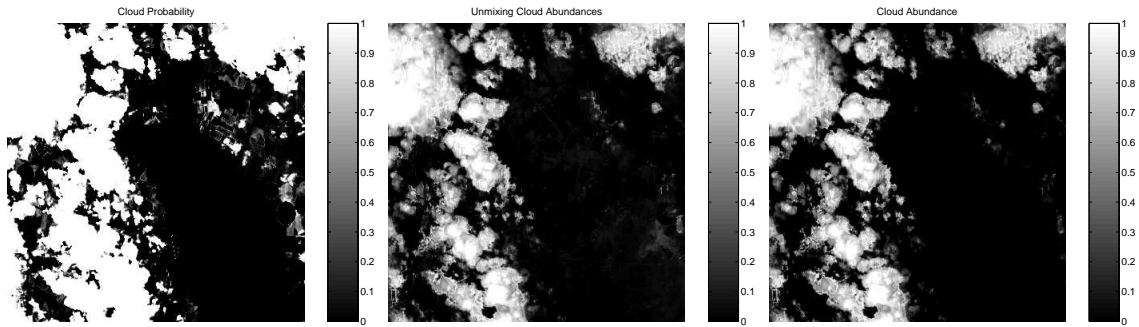
Figure 10: *Left*: *Cloud Probability* index computed from the posteriors of the cloud-clusters. *Center*: *Cloud Abundance* computed from the unmixing coefficients of the cloud clusters. *Right*: Cloud abundance product.

pure constituents and consequently only one endmember must represent them. Some examples of the negative effects of including mixed thin cloud spectra as endmembers were reported in Gómez-Chova et al. (2006). In the proposed algorithm, the *cloud endmember*, $\mathbf{m}_1$, is selected from all the cloud pixels, $\rho_k \subset C$, looking for the brightest one (maximum f_{Br}). The second issue is related to the total number of endmembers. If the value of Q is selected to be too low, then not all constituents will be extracted . On the other hand, if the value of Q is selected to be too high, some extracted endmembers may be unwanted non-pure signatures. However, this does not constitute a critical problem since we are not interested in obtaining accurate abundances for *all* the constituents present in the image, but basically in the cloud abundance. For this reason, obtaining some unpure *ground endmembers*, i.e. those mixture of two or more ground constituents, is not a problem as this will only affect the abundances related to ground endmembers.

2.5.3 Endmember initialization algorithm

Taking into account the previous considerations, we use the Automated Target Generation Process (ATGP) (Ren and Chang, 2003) to select the rest of endmembers, $\{\mathbf{m}_q\}_{q=2}^Q$, from the ground pixels, $\rho_k \subset \overline{C}$. The ATGP finds the endmembers in accordance with an orthogonal subspace projection criteria and it normally outperforms the other common endmember initialization algorithms (Plaza and Chang, 2006). In particular the ATGP is well-suited to our problem since it starts with a initial endmember signature, $\mathbf{m}_1$, then finds the next endmember signature, $\mathbf{m}_2$, looking for the ground pixel with the maximum absolute projection in the space orthogonal to $\mathbf{M} = [\mathbf{m}_1]$, adds the new endmember to $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2]$, and repeats the procedure until a set of Q endmembers $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_Q\}$ is extracted.

2.5.4 Cloud abundance

After the endmember selection, we apply the FCLSU to the image using all the available spectral bands except bands particularly affected by atmospheric absorptions ($\lambda_i \subset B_A$), since the linear mixing assumption is not appropriate at those bands. The FCLSU provides the vector

$\mathbf{a}_k$ of abundances for each sample pixel k. As it happens with the probabilities of the clusters, the abundance $a_{kq} \in [0,1]$ and $\sum_{q=1}^{Q} a_{kq} = 1$. Therefore, the *Cloud Abundance* is the sum of the abundances of the cloud-clusters which, in our case, represents the abundance of the cloud-endmember (Fig. 10):

$$a_{kC} = \sum_{q \subset C} a_{kq} = a_{k1} \tag{2.16}$$

As in the case of the probabilities, a threshold of a_{kC} would give a good cloud mask, but some false detections could appear since the unmixing has been performed on the basis of spectral signatures that could be non-pure pixels or, at least, non completely independent, thus providing relatively high cloud abundances in ground covers with similar spectral signatures.

2.5.5   Cloud product

An improved cloud product map can be obtained when combining the *Cloud Abundance*, a_{kC}, and the *Cloud Probability*, h_{kC}, by means of a pixel-by-pixel multiplication (Fig. 10, right).

$$\phi_k = a_{kC} h_{kC} \tag{2.17}$$

That is, combining two complementary sources of information processed by independent methods: the degree of cloud abundance or mixing (obtained from the spectra) and the cloud probability that is close to one in the cloud-like pixels and close to zero in remaining areas (obtained from the extracted features).

3   CHRIS Acquisition Modes

One of the main advantages of CHRIS instrument is the high configurability of its operation mode. In fact, the number of bands and their nominal wavelength allocations changes to a great extent from one mode to another. However, this advantage is a problem for the detection of clouds since the method has to take into account the number and configuration of the spectral bands for each CHRIS acquisition mode. For example, the oxygen and water vapour atmospheric absorptions or the ice/snow absorption are only present in the spectral region registered by Modes 1 and 5. The spectral coverage of bands acquired by the different modes is depicted in Fig. 11.

The proposed cloud masking algorithm has been designed for CHRIS Modes with full spectral information (Modes 1 and 5). Table 1 shows the available features depending on CHRIS acquisition mode. In the case of the brightness and whiteness, the number of spectral bands used to compute them is lower for modes 2, 3, and 4. Therefore the robustness and the discrimination power of these features will be poorer for these modes. Moreover, the absorption features cannot be computed at all for these modes. In consequence, the proposed algorithm could present a poor performance for Modes 2-3-4 images in critical cloud screening situations, such as over bright surfaces (ice, snow, sand, etc) and around cloud borders or thin semitransparent clouds.

Table 1: Available features depending on CHRIS acquisition mode and number of spectral bands used to compute them.

Feature	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
$f_{Br,VIS}$	22	10	7	5	8
$f_{Br,NIR}$	11	4	6	8	10
f_{Br}	33	14	13	13	18
$f_{Wh,VIS}$	22	10	7	5	8
$f_{Wh,NIR}$	11	4	6	8	10
f_{Wh}	33	10	13	13	18
f_{O_2}	3	–	–	–	3
f_{WV}	2	–	–	–	2

In addition, the cloud screening module is designed for images processed at version 4.1 of the CHRIS HDF files (Cutter and Johns, 2005) in order to take advantage of:

- the improved radiometric calibration which was poor in previous versions (underestimation up to a factor two of the sensor measurements in the NIR) (Cutter, 2004),

- the additional acquisition information contained in the metadata attributes of the CHRIS HDF file (image date, azimuth and zenith angles, etc),

- a quality mask added to the HDF data that includes pixel saturation information and occurrence of errors (used by the noise reduction module (Gómez-Chova, 2007)).
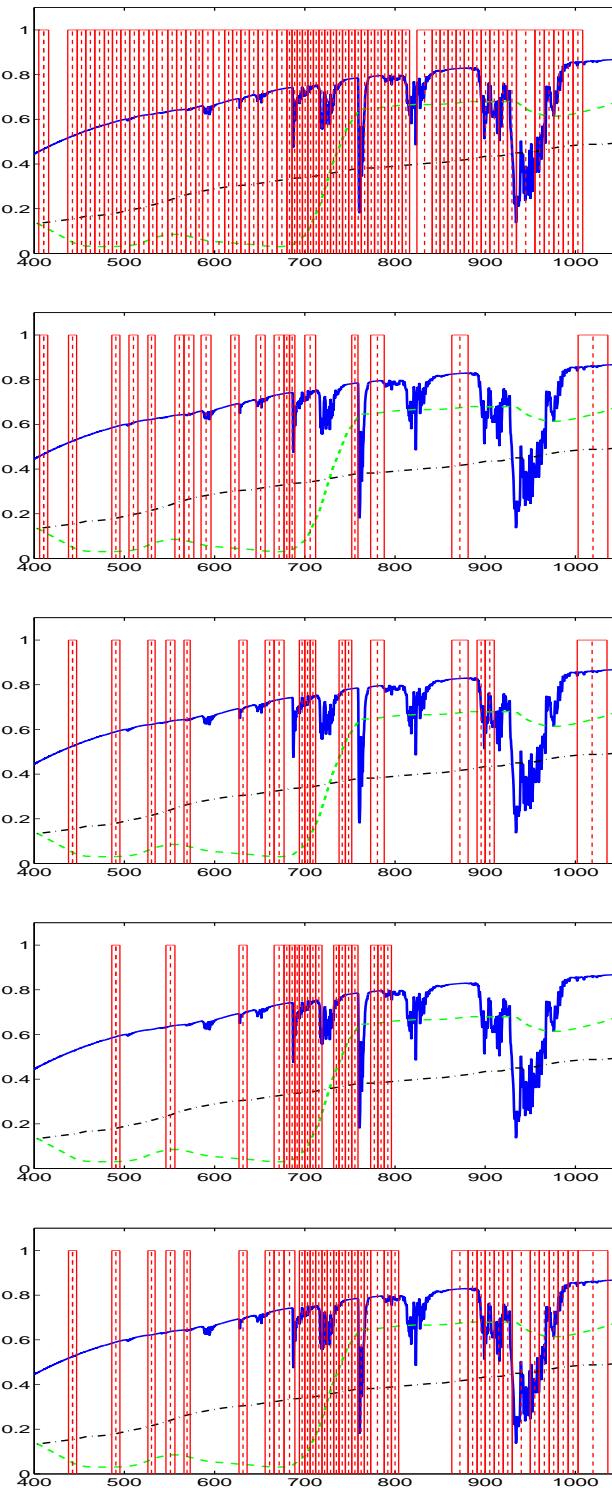
Figure 11: *Top-to-bottom*: Modes 1–5 CHRIS band locations (boxes) superimposed to a reflectance spectra of healthy vegetation (dashed), bare soil (dash-dotted), and the atmospheric transmittance (solid).

4   Conclusions

In this document, a methodology that faces the problem of accurately identifying the location and abundance of clouds in hyperspectral images is described. The algorithm identifies the location of clouds in the image and produces a cloud abundance map in order to quantify how cloud presence affects the measured spectra. The cloud screening algorithm is based on well-funded physical features, which are intended to increase separability between clouds and ground covers, and are extracted from the converted TOA reflectance in order to reduce dependence on illumination and geometric acquisition conditions. We should note that one critical feature introduced in this approach is the use of the atmospheric oxygen and water vapor absorption bands to improve cloud screening results. Afterwards, the main modules of the cloud screening algorithm have been formulated. An operative unsupervised classification is performed based on the extracted features in order to adapt the cloud screening to the given image conditions, mainly the atmospheric conditions, the background, and the present cloud types. This step allows the user to easily discriminate between cloud free and cloudy regions, where the method provides a cloud abundance product, which is based on spectral unmixing algorithm.

Several remarks and conclusions can be drawn. First, the cloud screening method has been proposed as an arrangement of different purpose-designed modules, which have been formulated in terms of simple and operational algorithms that cover the essential requirements for cloud screening process. These modules can be changed by more advanced algorithms or modified to adapt to the characteristics of the sensor.

A second important remark is that the method has been implemented to use self-contained information provided with CHRIS products. The inputs required by the cloud masking algorithm are the CHRIS image corrected of noises, image acquisition parameters extracted from the header information, and a set of classification parameters adjusted for each acquisition mode.

Finally, an important aspect of the proposed method is that it provides a cloud abundance product to the user that estimates the contribution of clouds to the spectra of image pixels. This information can be used to better describe detected clouds (subpixel coverage, transparency, cloud type) and to generate cloud masks with different restrictive levels depending on the application.

Acknowledgement

Bibliography

Barnsley, M., Settle, J., Cutter, M., Lobb, D., and Teston, F. (2004). The PROBA/CHRIS mission: a low-cost smallsat for hyperspectral, multi-angle, observations of the Earth surface and atmosphere. *IEEE Transactions on Geoscience and Remote Sensing*, 42(7):1512–1520.

Buriez, J. C., Vanbauce, C., Parol, F., Goloub, P., Herman, M., Bonnel, B., Fouquart, Y., Couvert, P., and Seze, G. (1997). Cloud detection and derivation of cloud properties from POLDER. *International Journal of Remote Sensing*, 18(13):2785–2813.

Chang, C. (2003). *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Kluwer Academic/Plenum Publishers.

Chapman, R. M. (1962). Cloud distributions and altitude profiles from satellite. *Planetary Space Science*, 9:70–71.

Cutter, M. (2004). CHRIS calibration issues. In (ESA-SP-578), E. P. D., editor, *2nd CHRIS/Proba Workshop*.

Cutter, M. and Johns, L. (2005). CHRIS data products – latest issue. In (ESA-SP-593), E. P. D., editor, *3rd CHRIS/Proba Workshop*.

Davies, D. and Bouldin, D. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-1(2):224–227.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Diner, D., Clothiaux, E., and Girolamo, L. D. (1999). *MISR Multi-angle Imaging Spectro-Radiometer Algorithm Theoretical Basis. Level 1 Cloud Detection.* Jet Propulsion Laboratory, JPL D-13397.

Duda, R. and Hart, P. (1973). *Pattern classification and scene analysis*. Wiley, New York.

Fisher, J. and Grassl, H. (1991). Detection of cloud-top height from backscattered radiances within the Oxygen A band. Part 1: Theoretical study. *Journal of Applied Meteorology*, 30:1245–1259.

Gómez-Chova, L. (2007). CHRIS/PROBA Noise Reduction ATBD. Technical Report 1.0, European Space Agency.

Gómez-Chova, L., Amorós, J., Camps-Valls, G., Martín, J., Calpe, J., Alonso, L., Guanter, L., Fortea, J., and Moreno, J. (2006). Cloud detection for CHRIS/Proba hyperspectral images. In Schafer, K. and Comeron, A., editors, *Remote Sensing of Clouds and the Atmosphere X*, volume 5979, page 59791Q, Bellingham, WA. SPIE.

Heinz, D. and Chang, C.-I. (2001). Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3):529–545.

Keshava, N. and Mustard, J. (2002). Spectral Unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57.

Plaza, A. and Chang, C.-I. (2006). Impact of initialization on design of endmember extraction algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3397–3407.

Ramon, D., Cazier, L., and Santer., R. (2003). The surface pressure retrieval in the MERIS O_2 absorption: validation and potential improvements. In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS'2003*, volume 5, pages 3126–3128.

Ren, H. and Chang, C.-I. (2003). Automatic spectral target recognition in hyperspectral imagery. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1232–1249.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Enquiry*. World Scientific, Singapore.

Simpson, J. (1999). Improved cloud detection and cross-calibration of ATSR, MODIS and MERIS data. In (ESA-SP-479), E. P. D., editor, *ATSR International Workshop on the Applications of the ERS along track scanning radiometer*.

Thuillier, G., Hersé, M., Labs, D., Foujols, T., Peetermans, W., Gillotay, D., Simon, P. C., and Mandel, H. (2003). The solar spectral irradiance from 200 to 2400 nm as measured by the SOLSPEC spectrometer from the ATLAS and EURECA missions. *Solar Physics*, 214:1–22.

Yamamoto, G. and Wark, D. Q. (1961). Discussion of the letter by R. A. Hanel: Determination of cloud altitude from a satellite. *Journal of Geophysical Research*, 66:3596.