Introducing spatial information in k-means algorithm for clouds detection in optical satellite images

L. Beaudoin[a,b], J.M. Nicolas[b], F. Tupin[b] and M. Hückel[a]

[a]Matra Systems & Information, 6 rue Dewoitine
F-78140 Velizy-Villacoublay, France

[b]E.N.S.T., Dep. TSI, 46 rue Barrault,
F-75013 Paris, France

ABSTRACT

Due to restricted visibility time of remote sensing polar platforms from earth reception station, only a limited number of images can be transmitted. In the case of optical images, an in-board cloud cover detection module will allow to transmit only useful (i.e. weakly cloudy) images. In order to derive such a module, we propose a method to detect cloudy areas from subsampled images.

For a pixel ground surface of about 100 x 100 m^2, cloudy areas appear as the highest radiometric value homogeneous areas. The algorithm presented in this paper is based on the k-means method. Its main originality is to improve classical results by introducing isotropic spatial information. Input data are the sorted components of a vector composed of radiometric values for each pixel and its neighbours (4-connexity). Then a classical k-means method with constraints on the cloudy class gravity center is used on these vectors.

We tested the method on a set of 206 subsampled SPOT XS and 138 SPOT P images and their manmade interpretation masks. To evaluate the quality of our results, we used the probability of false alarm (PFA) depending on the number of pixels which have been wrongly declared cloudy, and the probability of non detection (PND) depending on the number of pixels which have been wrongly declared non cloudy. We obtained rather good PFA($< 1\%$) and PND($< 30\%$), and compared these values with results obtained with other methods.

Keywords: Cloud detection, k-means, Spatial information, SPOT, Optical satellite imagering

1. THE CONTEXT OF THE STUDY

In this section, we present the reasons for making an in-board cloud cover estimation in a first time and in a second time we present the physical cloud charateristics that we use in the next sections.

1.1. The interest of in-board cloud cover estimation

Two main operations are made during the flight of the satellite: the first one is the digitalization of the scene, and the second one is the data transmission to the Earth. This second step can only be done when a reception station is seen by the satellite. Due to restricted visibility time of polar platforms, only a limited number of images can be sent. To increase this number, one solution is to improve the transmission rate. Yet, as spatial and spectral resolutions become better, the volume of remote sensing data grows drasticaly. Therefore, an other solution is to develop in board compression methods before data transmission. Since exact reconstruction is needed, compression rates are also limited. An in board cloud cover detection module allows to estimate the cloud cover in high optical resolution images (10 m): by this way, only useful (i.e. non very cloudy) images will be transmitted. An image is commercialy considered very cloudy when more than 20% of its surface is covered with clouds. For the SPOT satellite, about 80% of transmitted scenes are very cloudy[1].

At Matra Systems and Information (MS&I), a specific method has been developed to detect clouds and estimate the cloud cover under real time processing and low available resources constraints. This algorithm, not detailed in

Email: beaudoin@tsi.enst.fr; Telephone: (33) 1 45 818 085; Fax: (33) 1 45 813 794;
Email: nicolas@tsi.enst.fr; Telephone: (33) 1 45 818 129; Fax: (33) 1 45 813 794

this article, has been tested on a SPOT P and XS database*. To evaluate its performance, comparisons have to be done with more classical methods. As we will see later (section 4.3), classical classification algorithms are not completely satisfying. So we propose a modified version of the k-means algorithm which takes into consideration the spatial information. The aim of this paper is to present this original algorithm and to compare its performances to classical methods and the MS&I solution.

1.2. Physical characteristics of clouds

Many clouds are composed of thermal convective cells, which are areas of high steam density. Those cells have a characteristic size between 250 and 750 m, and clouds have a diameter up to few kilometers.[2] In the visible domain, with a pixel ground surface of about 100 x 100 m^2, the cloud reflectance, which depends strongly on the steam density, has low variations, so clouds appear as homogeneous areas. For 1 km^2 pixel size (typical resolution of a meteorological satellite), the reflectance depends on the random density of the convective cells: clouds look rather heterogeneous. In our study (SPOT images), clouds are homogeneous areas and have high radiometric values. Moveover, as snow and ice reflectance have almost the same properties, this can lead to confusion errors in classification schemes.

The snow and ice reflectance depends on two major parameters: the wavelength of the observation and the size of the snow grain radius.[3] Discrimination between reflectance of clouds and snow often uses the $1.6\mu m$[4] and $3.7\mu m$[5] observation wavelengths. But because of the size of fresh snow particles, the discrimination remains be difficult.[6]

About 50% of the Earth surface is cloudy, 30% and 10% of the land surface is covered by snow and ice respectively.[7] Those percentages explain why cloud detection is such an important problem for the remote sensing community. The classical methods use multi-spectral, time and spatial variations of the reflectance to detect cloud cover. The multi-spectral methods,[8–11] use combination of radiometric values of a pixel between 0.4 and 1.3 μm. The time methods[12,13] compare the pixel radiometry and a reference one calculated from a historical and georeferenced data set. The spatial methods[14–16] use relationships between a pixel and its neighbourings. Our adapted version of the k-means algorithm belongs to the spatial methods.

2. THE IMAGE DATA BASE

In this section, we present the chosen performance measures and the image data base used for this study.

2.1. The performance measures

The cloud detection module estimates the cloudy surfaces percentage of the image. Using this information, we will decide whether the image should be or not transmitted to Earth. Two performance measures appear to be important:

- the probability of false alarm (PFA) which is the ratio of the number of ground pixels wrongly declared cloudy to the total number of pixels in the image,
- the probability of non-detection (PND) which is the ratio of the number of cloud pixels not declared cloudy to the total number of pixels in the image.

Let us emphasize that the PFA must remain very weak in order to discard uncloudy images. A relatively high PND is less critical since in the worth case a cloudy image is retransmitted to the ground. Our goals is to obtain PFA about 1% and PND about 30%.

2.2. The image data base

In order to allow a real time in-board processing, we used subsampled SPOT XS and P images called quicklooks. The CNES database includes a set of 206 (XS) and 138 (P) quicklooks and their man made interpretation mask. The ground surface of a pixel is about $120 \times 120 m^2$. Let us remark that the pixel value is a radiometric value and not a luminance one since, in flight, the instrument calibration is not available. Therefore, all the algorithms must use the radiometric information.

The radiometric distribution of cloud and ground pixels for each spectral channel on the whole database is illustrated on figure 1. We conclude from a careful study of these statistics that XS1 and P channels are the most useful for our problem.

*This study has been funded by the CNES.

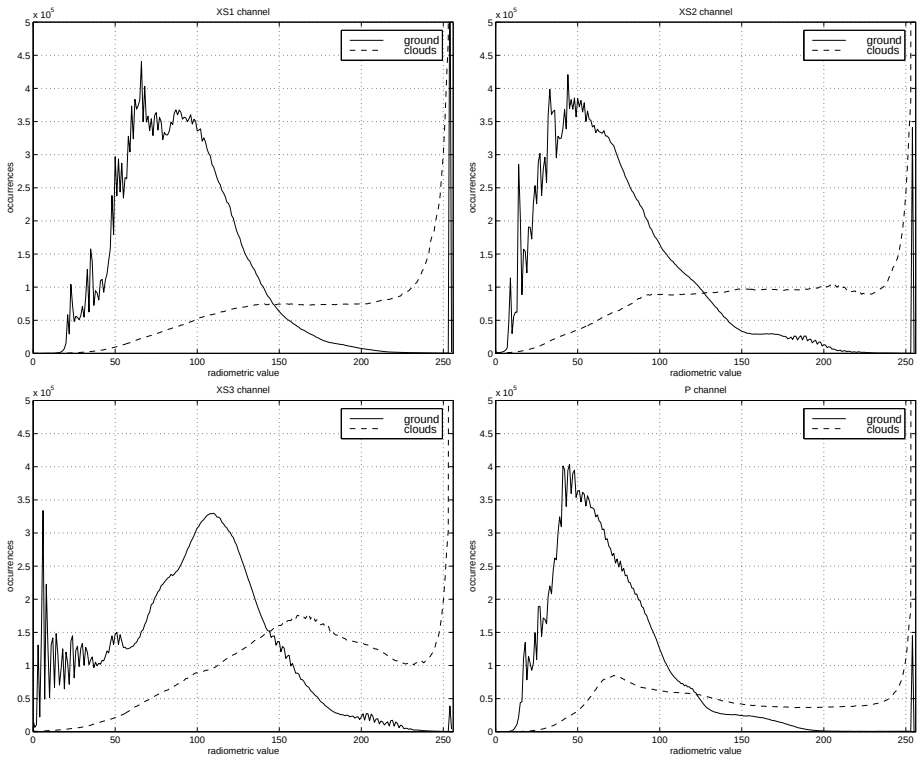

Figure 1. Histograms of ground pixels and cloud pixels on the whole database for each spectral channel.

3. CLASSICAL CLASSIFICATION RESULTS

In this section, we present the results obtained by classical supervised and unsupervised classification methods.

3.1. Supervised classification

For our problem, two classes are looked for: ground pixels class and cloud pixels one. A supervised classification method leads us to determine the optimal radiometric threshold S for cloud/ground discrimination. We note H_g and H_c the hypotheses that a pixel belongs to ground or to cloud respectively. We denote Ω the observations domain ($\Omega = [0, 255]$). To decide between H_g or H_c leads to subdivide Ω in two disconnected domains Ω_g ($\Omega_g = [0, S]$) and Ω_c ($\Omega_c = [S, 255]$). If the pixel radiometry is in Ω_g, then the classification decision δ_g is H_g, otherwise the decision δ_c is H_c.

Four situations are possible, as shown in table 1.

decision	reality	situation
δ_g	H_g	true non-detection
δ_c	H_c	true detection
δ_g	H_c	false non-detection
δ_c	H_g	false alarm

Table 1. Possible situations of a pixel classification.

We define:

- the probability of true detection

$$P_d^{supervised} = P(\delta_c|H_c) = \int_{\Omega_c} p(I|H_c)dI \tag{1}$$

- the probability of false alarm

$$P_{fa}^{supervised} = P(\delta_c|H_g) = \int_{\Omega_c} p(I|H_g)dI \tag{2}$$

We have tested two classical methods to determine S: the first one is based upon the Bayes criterion and the second on the Neyman-Pearson one[17].

Bayesian criterion
Bayesian segmentation minimises the global error risk:

$$P(\omega_i|I) = \frac{P(I|\omega_i).P(\omega_i)}{P(I)} \tag{3}$$

where :

- $P(\omega_i|I)$ is the posterior probability that the pixel belongs to class ω_i if its intensity is I
- $P(I|\omega_i)$ is the probability that the pixel has intensity I if its class is ω_i
- $P(\omega_i)$ is the prior probability of class ω_i
- $P(I)$ is the probability that the pixel has intensity I

Due to high variability between images, we decided that $P(\omega_i) = \frac{1}{k}$ not to favour any of the k classes. In the Bayesian case, for the whole data base we note the absolute threshold S_r^{Bay}. If a pixel radiometry is greater than S_r^{Bay}, this pixel is classified as cloud otherwise it is classified as ground. The values of S_r^{Bay} for each spectral channel are deduced from the intersection between the ground class and the cloud class curves in figure 1. Results are shown in Table 2 (left) where nb_{fa}, nb_{nd}, $\mu_{fa}^{\%}$ and $\mu_{nd}^{\%}$ are respectively the number of images with $PFA > 1\%$, with $PND > 30\%$, the mean percentage of PFA and PND computed on the whole data base. These results are clearly

channel	S_r^{Bay}	nb_{fa}	nb_{nd}	$\mu_{fa}^{\%}$	$\mu_{nd}^{\%}$
XS1	147	81	6	9.1	4.7
XS2	129	57	15	5.8	9.4
XS3	145	104	21	14.8	8.4
P	124	39	19	4.7	12.6

channel	S_r^{Ney}	nb_{fa}	nb_{nd}	$\mu_{fa}^{\%}$	$\mu_{nd}^{\%}$
XS1	228	3	44	0.7	17.8
XS2	219	3	72	0.67	25
XS3	233	3	108	0.12	34.4
P	196	2	33	0.54	20.5

Table 2. Left: results of the Bayesian classification, right: results of the Neyman-Pearson classification

unsatisfacting. It can be explained by the calibration data problem. Even if our learning data base is important, the data pre-processing could not let rigorously compare the radiometries (the comparison could only be possible with luminances). So, we only estimate approximatively S_r^{Bay}.

Neyman-Pearson criterion
The Neyman-Pearson criterion[17] consists in maximizing $P_d^{supervised}$ for a fixed maximal $P_{fa}^{supervised}$. In other words, it considers that the false alarm is the worse error and that it must not exceed a threshold $S_{P_{fa}}$ fixed by an expert. At the same time, it maximizes the detection in order to keep the efficiency of the system. In the case of Neyman-Pearson criterion, we denoted S_r^{Ney} the threshold S. In our case, $P_{fa}^{supervised} = S_{P_{fa}} = 1\%$. Equation (2) and the curves plotted on figure 1 allows to compute P_{fa} as a function of S_r^{Ney}. Table 2 (right) shows the obtained performances.

These results are better than those obtained with the Bayesian approach. Thus the spectral channels XS1 and P seem to be the more adapted to our problem. A lot of images of the data base show a $PND > 30\%$. It could be explained by the fact that, like in the Bayesian case, we use radiometric value and not luminance one, which makes the pixels of dense fog to be wrongly classified as ground when the calibration is not well tuned.

The main problem of supervised methods is its sensitivity to the instrument calibration. This leads to use more adaptative methods like the k-means algorithm.

3.2. Unsupervised k-means method

The k-means algorithm is a very classical non-hierarchic clustering method. The differents steps of this iterative algorithm are:

1. choice of the number of classes
2. initialization of the mass centers m_j of the k classes ω_j
3. at iteration i, the pixel s with a radiometric value I_s is a member of class ω_j if m_j is the nearest mass center from I_s
4. when all the pixels are classified, the new mass centers $m_j(i+1)$ are computed
5. the algorithm stops if all the centers m_j are stable otherwise we go back to step 3

The performances of this algorithm depend on the number of classes (step 1), on the choice on the initialization centers (step 2), and on the data geometric properties. For the initialization of the mass class centers, the center with the higher radiometric value represents the cloud class, the others are equi-distributed between 0 and this center. For the classes number choice, the situation is more complex than for the Bayesian one. In fact, the high data disparity leads to poor classification with only two classes. We tested up to 15 classes with only the class with the highest radiometry taken as the clouds class. The more the number of classes is high, the more the discrimination between clouds and ground and the less the clouds class exhibits false alarms. But on the other hand, the rate of non-detection increases. Figure 2 shows this non-detection rate for the XS1 and P channels.

Though the mean PFA percentage is almost constant (0.8% for XS1 and 0.6% for P) for 5(XS) and 8(P) classes, it is necessary to reach 10(XS) and 11(P) classes in order to obtain the stability for the number of images which have PFA>1%. The results obtained are shown in Table 3. The situations leading wrong classifications could be

channel	nb_{fa}	nb_{nd}	$\mu^{\%}_{fa}$	$\mu^{\%}_{nd}$
XS1	5	38	0.78	17.17
P	5	45	0.49	23.1

Table 3. Results of the k-means classification.

organised in four families:

- images with no or few clouds (figure 3, top left)
- images with a lot of light fog (figure 3, top rigth)
- images with a low contrast (figure 3, bottom left)
- images with snow (figure 3, bottom left)

Figure 4 shows the radiometric distribution of the threshold determined by the k-means algorithm for XS and P data. We can see a dispersion which explains better results than with supervised methods.

The classical classification algorithms show too important $\mu^{\%}_{fa}$. Nevertheless, we can extract some clues in order to resolve our images selection problem in flight, yielding the following conclusions:

- an algorithm which can adapt itself to each image is needed.
- the XS1 and P spectral channels appear to be the best inputs.

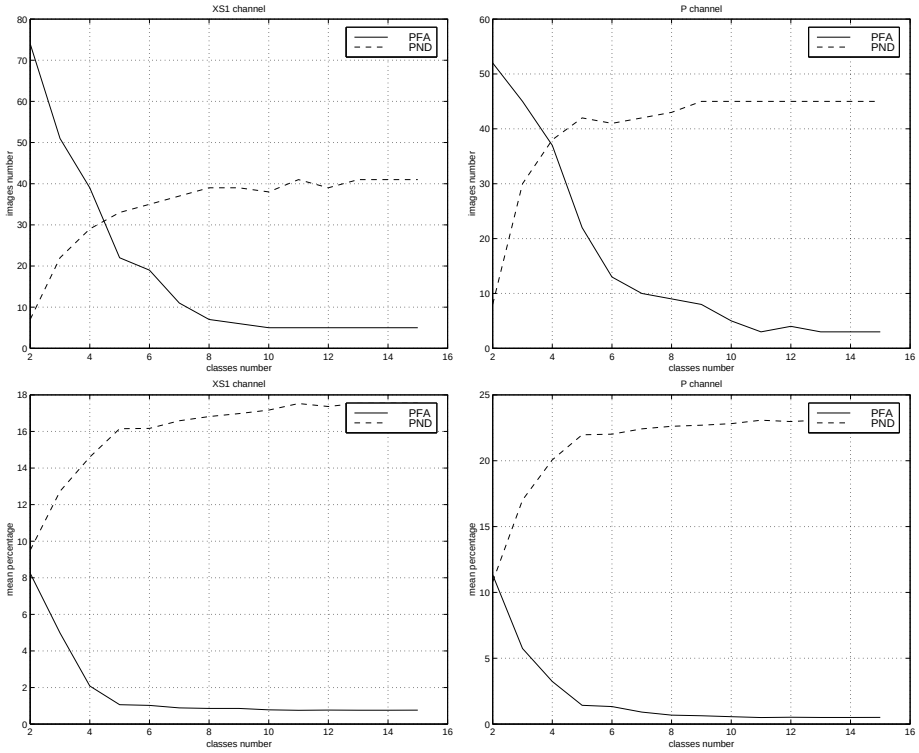

Figure 2. Number of images with PFA>1% and with PND>30% (top), mean PFA and PND percentages computed on the whole data base (bottom) as a function of the number of classes.

4. K-MEANS SEGMENTATION ALGORITHM WITH SPATIAL INFORMATION

As detailed in the section 1.2, the spatial information is important, as we have seen that the cloudy area is homogeneous. In this section, we propose to improve the classical k-means algorithm to take into account this spatial information. Firstly, we detail the algorithm, then we emphasize its differences with the previous ones and we compare the obtained results.

4.1. The algorithm

The basic idea of the algorithm is to modify the classical k-means one in order to take into account the radiometry of a pixel and its 4-connexity neighbours in an anisotropic way. A pixel s is now described by a 5 components vector $\overrightarrow{V_s}$ corresponding to the radiometry of the pixel and its neighbours. By sorting the $\overrightarrow{V_s}$ components from the highest one to the least one, we have lost the spatial localisation information from each pixel to each others, yielding an isotropic algorithm (see figure 5). If the pixel is in an homogeneous area, $\overrightarrow{V_s}$ is colinear to the vector $\overrightarrow{V_{max}}$ which is equal to (255,255,255,255,255). We call the $\overrightarrow{V_{max}}$ direction the homogeneity axis. If a pixel corresponds to an heterogeneous area, its direction can be far from the homogeneity axis. If the pixel s is in a cloudy area, the norm of $\overrightarrow{V_s}$ is great: the mass center of clouds can be initialized as the greatest vector on the homogeneity axis.

Let us detail the different steps of our algorithm:

1. choice of the number of classes
2. initialization of the mass centers $\overrightarrow{V_j^c}(0)$ of the k classes ω_j
3. at step i, the pixel s described by $\overrightarrow{V_s}$ is classified in class ω_j if $\overrightarrow{V_j^c}(i)$ is the nearest center of the vector $\overrightarrow{V_s}$

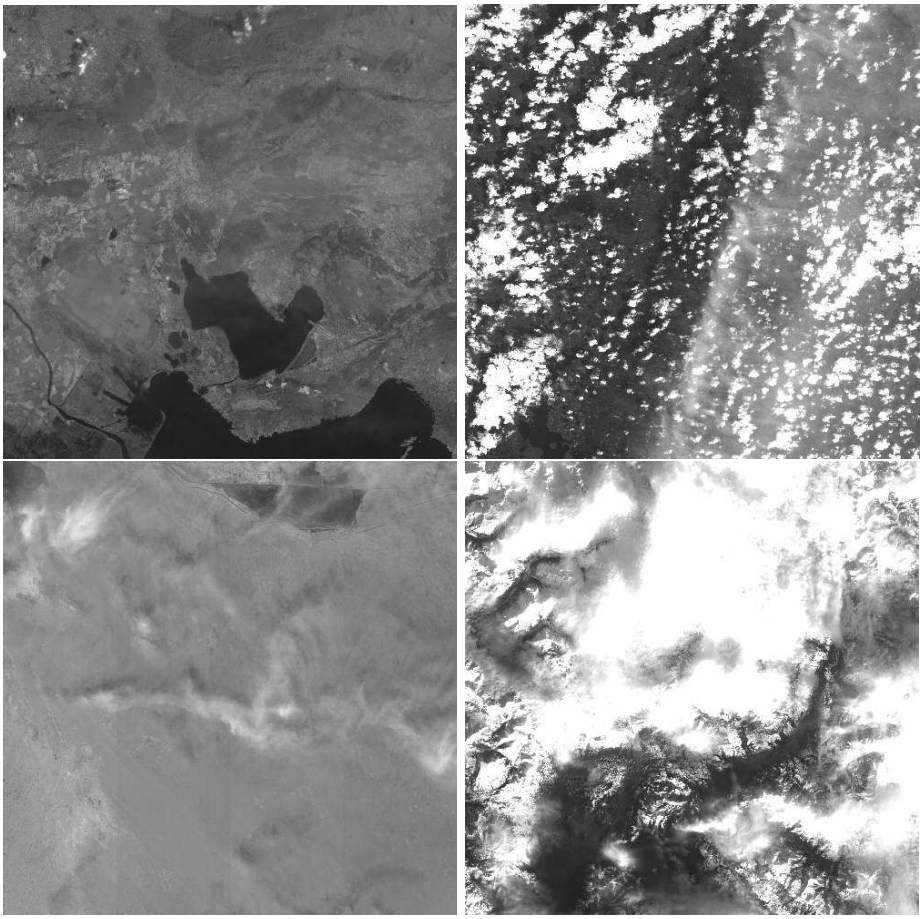

Figure 3. Exemples of situations that lead the classical k-means algorithm to wrong classifications: few clouds (top left), light fog (top right), low contrast (bottom left), snow (bottom right). ©CNES

4. when all the pixels are classified, the new mass centers $\overrightarrow{V_j^c}(i+1)$ are computed. During this step, we constraint the mass center of clouds class to remain on the homogeneity axis. To do so, the new mass center $\overrightarrow{V_{clouds}^c}$ of clouds class is equal to (p, p, p, p, p) where p is the median value of the $\overrightarrow{V_{clouds}^c}$ components
5. the algorithm stops if all the centers $\overrightarrow{V_j^c}$ are stable else we go back to step 3

As for the classical k-means algorithm, the choices of the initial centers and of the number of classes are critical in order to obtain good performances. We propose that the initial centers are equally spaced on the homogeneity axis between a minimal and a maximal value. The choice of the number of classes is, like in the classical case, empiricaly determined. Results are shown in figure 6. A classification in 8 (XS1) and 9 (P) classes seems to be the best compromise.

4.2. The differences with the classical k-means

Let us now emphasize the basic differences between the classical and the adapted k-means algorithms. In the classical one, the goal is to define the radiometric threshold S between the ground and the clouds. In the adapted version, this threshold S does not exist. Indeed, a pixel with high radiometric value (250 for example) but in a non homogeneous area could be classified as non cloud. On the contrary, a pixel with lower radiometric value (243), but in an homogeneous area could be class as cloudy. Figures 7 and 8 give an example of this situation. Figures 7 shows two pixels: A and B, with their neighbours in 4-connexity. To illustrate our algorithm, we keep from the

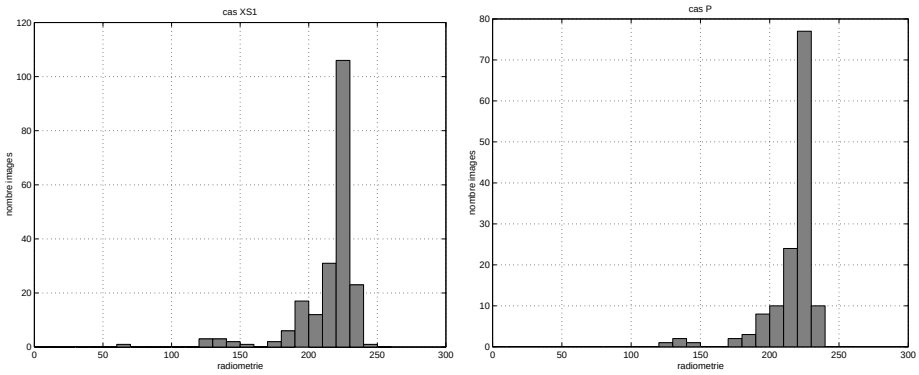

Figure 4. Radiometric threshold distribution determined by the k-means algorithm.

$$\begin{array}{|ccc|}\hline 100 & 250 & 248 \\ 29 & 112 & 107 \\ 30 & 32 & 39 \\ \hline\end{array} \longrightarrow \vec{V_s} = \begin{bmatrix} 250 \\ 112 \\ 107 \\ 32 \\ 29 \end{bmatrix}$$

Figure 5. A pixel with its neighbours and the vector $\vec{V_s}$

5-dimentionnal vectors $\overrightarrow{V_A}$ and $\overrightarrow{V_B}$ only the second and the fourth components. The first vector is equal to (250,235) and the second one to (243,243). Initial centers have been chosen as (250,250) for clouds and as (240,240) for the second class. At the first iteration, only the pixel A is classified as cloudy as we can see on the left of the figure 8. The double arrows line is the separator between the 2 classes. Because of the constraint to clouds mass center to remain on the homogeneity axis, the actual situation is illustrated on the figure 8 right: the pixel B is considered as cloudy, A as non cloudy, even if the B radiometry is lower than the A one.

To summarize, the basics differences with the classical algorithm are:

- the adapted k-means algorithm suppresses from the cloud class the inhomogeneous pixels with an high radiometric value.
- the adapted k-means algorithm adds in the cloud class the homogeneous pixels even if their radiometric value are lower.

4.3. The results

Tested on the whole data base, the modified algorithm suppresses more pixels than it adds in the cloudy class. Figure 9 illustrates this situation. On the left, we can see a XS1 South America scene. On the right, we can see in bright the mask obtained with the classical k-means algorithm minus the one obtained by the k-means with spatial information. As expected, the eliminated inhomogeneous bright pixels are located on clouds boundaries: the PFA is passed from 1.32% to 0.25% and the PND from 10.61% to 16.31%.

Table 4 gives classification results for XS1 scenes (8 classes) and P scenes (9 classes). The comparison with tables 2 and 3 confirms that as well as in nb_{fa} as in $\mu^{\%}_{fa}$, the results are improved.

The algorithm developed by MS&I is available in 2 versions: a mono-spectral(MSI-P) and a multi-spectral one(MSI-XS). Comparisons between all those results are given on figure 10. The k-means algorithm with spatial information seems to match better to our images classification problem than the other classical ones; yet the M&SI algorithm seems better for the XS1 channel.

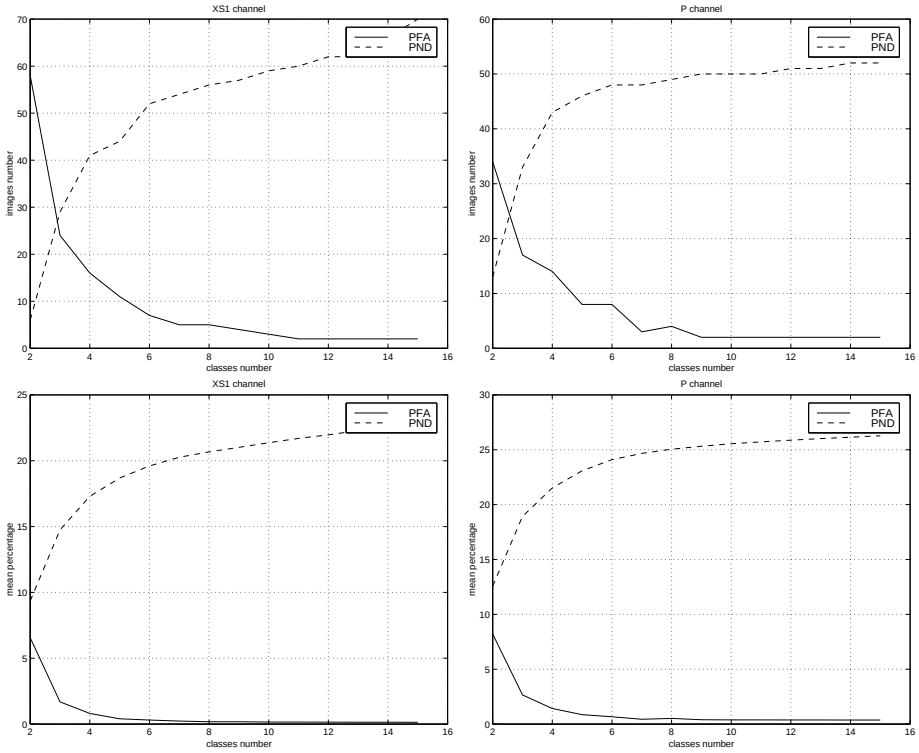

Figure 6. Number of images with PFA>1% and PND>30% (top), mean PFA and PND percentages computed on the whole data base for a k-means classification with spatial information as a function of the number of classes.

250	243
250 250 235	243 243 243
235	243

Figure 7. 2 pixels (left pixel A, right pixel B) and their neighbours in 4-connexity.

channel	nb_{fa}	nb_{nd}	$\mu^{\%}_{fa}$	$\mu^{\%}_{nd}$
XS1	5	56	0.19	20.7
P	2	50	0.4	25.3

Table 4. Results of the k-means classification with spectral information (to be compared with Table 4).

5. CONCLUSION

As, due the restricted visibility time of a remote sensing satellite, a limited number of images could be sent to Earth, in flight estimation of the cloud cover is a major objective.

Because of the calibration errors, supervised methods are inadapted. The classical k-means method gives better results but does not take into account the spatial information. So, a modified version of the classical algorithm is proposed in this article. This version deals with a neighbourhood of each pixels processed to appear as isotropic. After detailed the basics differences with the classical algorithm, it has been shown that the obtained results are better than the classical ones and near from the M&SI algorithm ones which exhibits best results.

Thanks to CNES for the availability of a large image database.

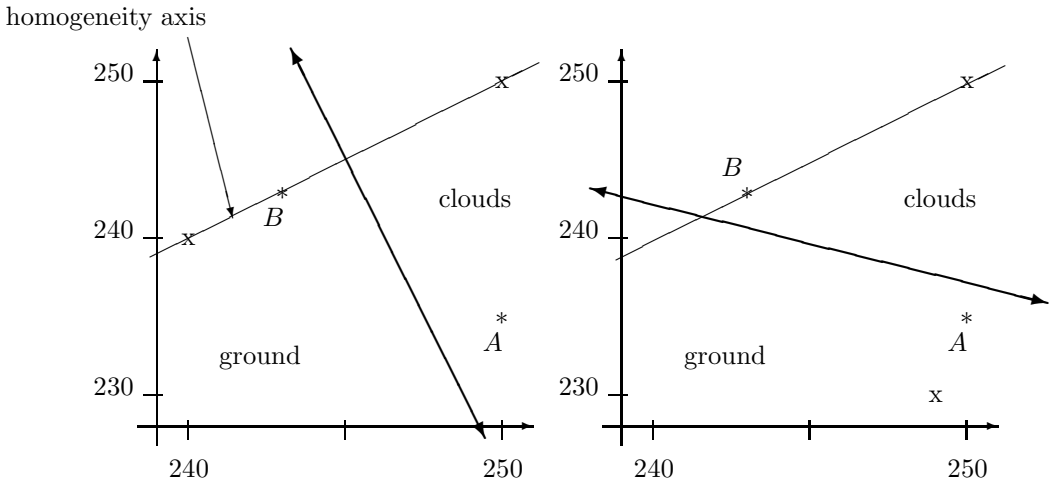

Figure 8. Classification of the pixels A and B. On the left, the end of the first iteration. On the right, the end of the second iteration. The mass centers are represented by x and the vectors associed to the pixels by *. The double arrows line is the separator between the 2 classes.

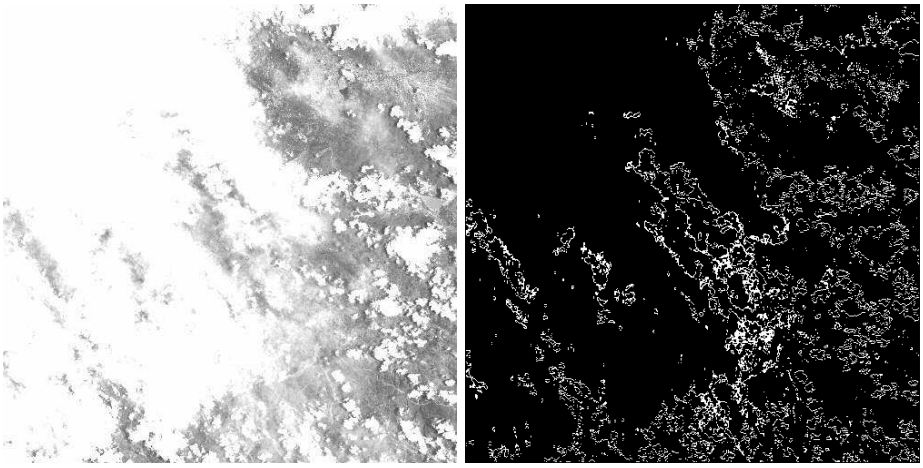

Figure 9. A south America XS1 scene (left) and the mask obtained with the classical k-means algorithm minus the one obtained by the k-means with spatial information.

REFERENCES

1. A. Fontanel, "Commercialisation du service SPOT. un premier bilan après six mois d'exploitation," in *SPOT 1, premiers résultats en vol*, Cépaduès, 1986.
2. R. W. B.A. Wielicki, "Cumulus cloud properties derived using LANDSAT satellite data," in *Journal of climate and applied meteorology*, vol. 25, pp. 261–276, 1986.
3. S. Warren, "Optical properties of snow," in *Reviews of geophysics and space physics*, vol. 20, pp. 67–89, 1982.
4. V. D.K. Hall, G.A. Riggs, "Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data," *Remote sensing of environment* **54**, pp. 127–140, 1995.
5. B. Baum, "A grouped threshold approach for scene identification in AVHRR imagery," *Journal of atmospheric and oceanic technology* **16**(6), pp. 783–800, 1999.
6. D. H. J.P. Ormsby, "Reflectance of fog, clouds and other features over snow and ice," in *IGARSS*, pp. 1717–1720, 1991.

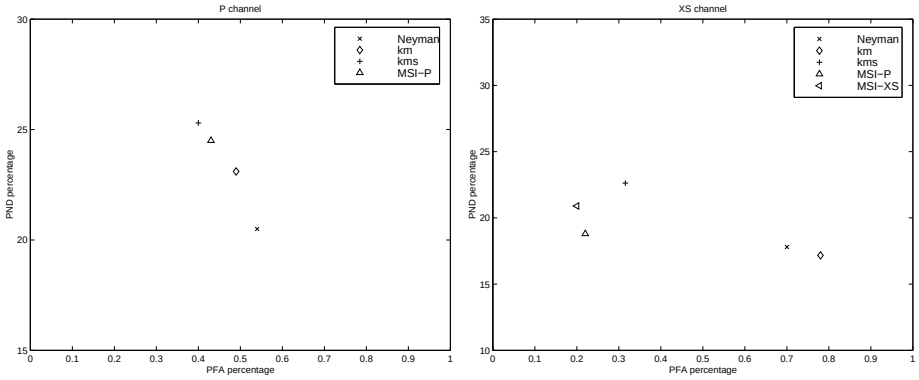

Figure 10. Comparison of all the classification results. *km* means classical k-means, *kms* the adapted version, $MSI - P$ and $MSI - XS$ are for the MS&I algorithm in mono and multi spectral version.

7. R. Crane, "The influence of clouds on climate with a focus on high latitude interactions," in *J. Climatol.*, vol. 4, pp. 71–93, 1984.
8. J. C. A. Arking, "Retrieval of cloud cover parameters from multi-spectral satellite images," in *Journal of climate and applied meteorology*, vol. 24, pp. 322–333, 1985.
9. G. S. M. Desbois, G. Sèze, "Automatic classification of clouds on METEOSAT imagery: application to high-level clouds," in *Journal of applied meteorology*, vol. 21, pp. 401–412, 1982.
10. K. K. R.W. Saunders, "An improved method for detecting clear sky and cloudy radiances from AVHRR data," in *International Journal Remote Sensing*, vol. 9, pp. 123–150, 1988.
11. M. D. L. Wald, G. Sèze, "On some operational procedures for clouds screening in NOAA AVHRR day time imagery," in 11^{th} *Earsel symposium*, 1991.
12. D. Loyola, "A new cloud recognition algorithm for optical sensors," in *IGARSS*, pp. 572–574, 1998.
13. P. H. L. P. A. Meygret, M. Dinguirard, "The SPOT histogram data base," in *SPIE Proceedings*, vol. 2957, pp. 322–331, 1996.
14. F. B. J.A. Coakley, "Cloud cover from high-resolution scanner data: detecting and allowing for partially filled fields of view," in *Journal of geophysical research*, vol. 87, pp. 4917–4932, 1982.
15. K. R. N. Khazenie, "Classification of cloud types based on spatial textural measures using NOAA-AVRHH data," in *IGARSS*, pp. 1701–1705, 1991.
16. P. D. G. Okemba, "Cloud cover estimation by analysing SPOT monochromatic images," in *third conference on image processing and its applications, IEEE*, 1989.
17. C. Desrousseaux, *Utilisation d'un critère entropique dans les systèmes de détection.* PhD thesis, Université de Lille 1, 1998.