Bayesian Segmentation and Clustering for Determining Cloud Mask Images

D. Barreto[a,b], F. Murtagh[a], J. Marcello[b]

[a] School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland
[b] Departamento de Señales y Comunicaciones, U.L.P.G.C, Campus Univ. de Tafira, 35017 Las Palmas de Gran Canaria, Spain

ABSTRACT

We assess both marginal density clustering, and spatial clustering using a Markov random field, on multiband Earth observation data. We use a Bayes factor assessment procedure in all cases. We find that the spatial model leads to better results, although the non-spatial clustering achieves a better false alarm rate.

Keywords: multiband image, Earth observation, principal components analysis, Karhunen-Loève transform, Gaussian mixture model, segmentation, Markov model, Bayes factor, BIC, PLIC

1. INTRODUCTION

A Bayesian assessment framework provides an objective and generally-applicable approach to classification and related decision-making. In this article, we apply a Bayes factor approach to the classification of pixels as being cloud or non-cloud. The Bayes factor, developed by Jefferys in the 1930s, is the posterior odds of one model over another when the prior probabilities of the two models are equal. We describe how approximations to the Bayes factor are used in practice. In particular, we use the Bayes information criterion or BIC, and the pseudo-likelihood information criterion or PLIC. While we employ both of these criteria with Gaussian model fitting, BIC is used in the non-spatial case, and PLIC is used in the spatial case.

2. FINITE GAUSSIAN MIXTURE MODELING AND BIC

In the univariate finite Gaussian mixture model, one-dimensional observations x_i are assumed to be drawn from G groups, each of which is Gaussian distributed. The g-th group has mean μ_g and variance σ_g^2. Given observations $x = (x_1, \ldots, x_n)$, let γ be an unobserved $n \times G$ cluster assignment matrix, where $\gamma_{ig} = 1$ if x_i comes from the g-th group, and $\gamma_{ig} = 0$ otherwise. Our goals are to determine the number of clusters G, to determine the cluster assignment of each pixel, and to estimate the parameters μ_g and σ_g of each cluster.

The probability density for this model is

$$f(x_i|\theta, \lambda) = \sum_{g=1}^{G} \lambda_g f_g(x_i|\theta_g), \tag{1}$$

where $\theta_g = (\mu_g, \sigma_g^2)^T$, $f_g(\cdot|\theta_g)$ is a Gaussian density with mean μ_g and variance σ_g^2, $\theta = (\theta_1, \ldots, \theta_G)$, and $\lambda = (\lambda_1, \ldots, \lambda_G)$ is a vector of mixture probabilities such that $\lambda_g \geq 0$ $(g = 1, \ldots, G)$ and $\sum_{g=1}^{G} \lambda_g = 1$.

We estimate the parameters by maximum likelihood using the EM (expectation-maximization) algorithm.[1,2] This is a procedure for iteratively maximizing likelihoods in situations where there are unobserved quantities and estimation would be simple if these were known. In the clustering case, the unobserved quantities are the cluster assignments given by the matrix γ.

The EM algorithm iterates between the E step and the M step. In the E step, the conditional expectation, $\hat{\gamma}$, of γ given the data and the current estimates of θ and λ is computed, so that $\hat{\gamma}_{ig}$ is the conditional probability

Author contact information: e-mail f.murtagh@qub.ac.uk

that x_i belongs to the g-th group. In the M step, conditional maximum likelihood estimators of θ and λ given the current $\hat{\gamma}$ are computed.

The E step and the M step are both simple, so that the EM algorithm as a whole is also simple. By contrast, direct maximization of the likelihood for the mixture model is complex in general. Although the EM algorithm has some limitations (e.g. it is not guaranteed to converge to a global rather than a local maximum of the likelihood), it is generally efficient and effective for Gaussian clustering problems. This procedure is especially efficient for clustering image pixels using single color bands or grayscale images.

Choosing the Number of Clusters via Bayesian Model Selection

We use Bayesian model selection[3] to choose the number of clusters. We consider a range of candidate numbers of clusters, $G = G_{\min}, \ldots, G_{\max}$. Each possible number of clusters, G, implies a different statistical model for the data, M_G. The model M_G has a vector of unknown parameters, ψ_G, consisting of the G means, the G variances, and the $(G-1)$ independently estimated mixture probabilities: $(3G-1)$ parameters in all. Our prior model probabilities are $p(M_G)$ for $G = G_{\min}, \ldots, G_{\max}$, where $\sum_{G=G_{\min}}^{G_{\max}} p(M_G) = 1$. Often each number of clusters considered is taken to be equally likely *a priori*, so that $p(M_G) = 1/(G_{\max} - G_{\min} + 1)$ for each G. The model parameters ψ_G also have prior distributions $p(\psi_G|M_G)$, which are typically rather diffuse and do not affect the final conclusions unduly. The data produce posterior model probabilities, $p(M_G|x)$, where again $\sum_{G=G_{\min}}^{G_{\max}} p(M_G|x) = 1$.

By Bayes' theorem,

$$p(M_G|x) = \frac{p(x|M_G)p(M_G)}{\sum_{H=G_{\min}}^{G_{\max}} p(x|M_H)p(M_H)}, \quad G = G_{\min}, \ldots, G_{\max}. \tag{2}$$

In (2), $p(x|M_G)$ is the *integrated likelihood* of model M_G, which requires integration over the model's parameter space, as follows:

$$p(x|M_G) = \int p(x|\psi_G, M_G)p(\psi_G|M_G)d\psi_G, \tag{3}$$

by the law of total probability.

The integral (3) is intractable analytically and is not easy to evaluate. However, twice the logarithm of the integrated likelihood can be approximated by the Bayesian Information Criterion, or BIC:

$$\begin{aligned} 2\log p(x|M_G) &\sim 2\log p(x|\hat{\psi}_G, M_G) - (3G-1)\log n \\ &= \text{BIC} \end{aligned} \tag{4}$$

(See Schwarz[4] and Kass and Raftery.[3]) In (4),

$$p(x|\hat{\psi}_G, M_G) = \prod_{i=1}^{n} \sum_{g=1}^{G} \hat{\lambda}_g f_g(x_i|\hat{\theta}_g)$$

is the maximized likelihood. In words, BIC = 2(log maximized likelihood) + (log n)(number of parameters). The BIC measures the balance between the improvement in the likelihood and the number of model parameters needed to achieve that likelihood. While the absolute value of the BIC is not informative, differences between the BIC values for two competing models provide estimates of the evidence in the data for one model against another.

3. SPATIAL CLUSTERING

Background on the approach pursued here can be found in Stanford[5] and Stanford and Raftery.[6]

We consider an unknown, true pixel state, for pixel i, as $X_i \in \{1, 2, \ldots K\}$ for K states. The observed image pixel is Y_i. This can be taken either as a scalar, or instead as a vector for color or multiband images. In this paper, Y_i is a point in a 4-dimensional redshift space. Consider an indicator function, $I(X_i, X_j) = 1$ if $X_i = X_j$ and otherwise $= 0$.

We now use a Markov random field to define spatial structure on X. We take $p(X)$ as being proportional to $\exp(\phi \sum_{i,j} I(X_i, X_j))$. This is a Potts or Ising model. ϕ is a spatial homogeneity parameter. A small value implies randomness, and a large value implies uniformity. A negative value of ϕ implies dissimilarity between neighboring pixels, and is not of interest here. Our model is a hidden Markov model because the variables X are only known through the observed Y.

Let $N(X_i)$ be the neighborhood of X_i, e.g. 3×3 pixels. Let $U(N(X_i), k)$ be the number of neighborhood pixels with state k.

From $p(X)$ we have the conditional distribution:

$$p(X_i = j \mid N(X_i), \phi) = \frac{\exp(\phi U(N(X_i)), j)}{\sum_k \exp(\phi U(N(X_i)), j)} \tag{5}$$

Having looked at the latent space, we now return to the observed data. We assume the following conditional density model connecting the observed and hidden variables: $f(Y_i \mid X_i = j)$ is Gaussian with mean vector μ_j and variance and covariance matrix V_j. The Y_i are conditionally independent given the X_i or, alternatively expressed, dependence among the Y_i only occurs via dependence among the X_i. Call θ_k the set of parameters, (μ, σ^2) for state k. We have $f(Y \mid X) = \Pi_i f(Y_i \mid X_i) = \Pi_i f(Y_i \mid \theta_{X_i})$.

Our solution algorithm is as follows. It is based on Besag's[7] iterated conditional modes (ICM) algorithm, which reconstructs an image based on local properties modeled as a Markov random field. This iterative algorithm requires an initial estimate of X, $\hat{X}$, and proceeds to estimate the parameters of $p(Y_i \mid X_i)$, as well as ϕ and X. To initialize X, we note that in taking $p(Y_i \mid X_i)$ as Gaussian, then the marginal density of Y is a finite mixture of Gaussians. Therefore to initialize X we use a Gaussian mixture model fit to the marginal density of a selected band (the fourth or last one was used here). For this, we use the EM (expectation-maximization) iterative algorithm, which is a simple version of the full segmentation algorithm.

Segmentation Algorithm

Step 0: Initialize $\hat{X}$ using a marginal segmentation.

Step 1: Update $\hat{\theta} =$ argmax $f(Y \mid \hat{X})$ based on maximum likelihood estimates of $\theta_j = \{\mu_j, V_j\}$ for each class, j.

Step 2: Update ϕ using the maximum pseudo-likelihood: $\hat{\phi} = \text{argmin}_\phi(-\log \text{PL}(\hat{X} \mid \phi))$. The pseudo-likelihood is given by $\text{PL}(\hat{X} \mid \phi) = \Pi_i p(\hat{X}_i \mid N(\hat{X}_i, \phi))$.

Step 3: Update $\hat{X}$: for each pixel i, $\hat{X}_i = \text{argmax}_j f(Y_i \mid X_i = j) p(X_i = j \mid N(\hat{X}_i, \hat{\phi}))$.

Implementation details: In step 2, if $\hat{\phi}$ goes negative, then we reset it to zero. In all calculations, we exclude boundary pixels from consideration. Step 1 is one step of Besag's ICM algorithm.

Choosing the Number of Segments via Bayesian Model Selection

We now turn attention to model selection. This is developed not for the homogeneity parameter, ϕ, nor for the neighborhood, but rather for the number of segments, K (see Stanford and Raftery[6]).

The posterior distribution of X conditional on Y is: $f(X \mid Y) = f(Y \mid X)f(X)/f(Y) \propto f(Y \mid X)f(X)$. Since there is conditional independence between Y and X, we have that $f(Y \mid X) = \Pi_i f(Y_i \mid X_i)$ which, it has already been noted, is taken as Gaussian.

The density of x, $f(X)$, is related to all possible states, which is combinatorially explosive. Therefore the pseudo-likelihood, $\mathrm{PL}(X)$, is taken as a proxy for $f(X)$. The pseudo-likelihood, introduced in Besag,[8] restricts where the integrated likelihood is defined. We have

$$\mathrm{PL}(X, \phi) = \Pi_i p(X_k \mid N(X_i), \phi) = \Pi_i \frac{\exp(\phi U(N(X_i)), X_i)}{\sum_k \exp(\phi U(N(X_i)), k)} \tag{6}$$

The likelihood is made conditional on the neighborhood of pixel i. Previously we had

$$L(Y_i \mid X_i) = \sum_j f(Y_i \mid X_i = j)p(X_i = j)$$

for state or label j.

Instead, denoting X_{-i} the neighborhood of X_i not including pixel i, and with $\hat{X}$ denoting an estimate of X, we use:

$$L(Y_i \mid N(\hat{X}_{-i})) = \sum_j f(Y_i \mid X_i = j)p(X_i = j \mid N(\hat{X}_i)) \tag{7}$$

As already noted, the first part of the right hand side term requires evaluation of a Gaussian; and the second part uses the conditional distribution defined for $p(X)$ in equations 5 and 6.

Let us consider a model as M_k, a function of the number of classes, k, and defined by the estimates of means, variances and covariances, and other properties related to homogeneity parameter ϕ and neighborhood N. By Bayes theorem, the posterior probabilities are:

$$p(M_k \mid Y) = \frac{p(Y \mid M_k)p(M_k)}{\sum_{h=\mathrm{Kmin}}^{\mathrm{Kmax}} p(Y \mid M_h)p(M_h)}$$

Since each number of clusters is considered equi-likely, the prior $p(M_k)$ is constant. The model prior, $p(M_h)$, can be ignored as not affecting the final conclusions unduly. The term $p(Y \mid M_k)$ is the integrated likelihood of model M_k. The Schwarz criterion,[3, 4, 9] or Bayes information criterion, approximates the integrated likelihood as $p(Y \mid \hat{\theta}, M_k)$, for which we use equation 7.

Given our use of pseudo-likelihoods for all pixels, this criterion is termed the pseudo-likelihood information criterion, PLIC.[5, 6]

Our model selection algorithm in practice entails looking at values of $k = 1, 2, \ldots, 20$, and finding the first local maximum.

4. INITIAL DATA PROCESSING: PRINCIPAL COMPONENTS ANALYSIS

We used 5-band NOAA 14 image data from 2000/9/24, relating to the eastern Atlantic Ocean. Each image band was of dimensions 3313×2048. We used bands 1, 2, 4, and 5. Correlations between these bands were as follows:

$$\begin{pmatrix} 1.00 & 0.99 & -0.40 & -0.40 \\ 0.99 & 1.00 & -0.43 & -0.43 \\ -0.40 & -0.43 & 1.00 & 1.00 \\ -0.40 & -0.43 & 1.00 & 1.00 \end{pmatrix}$$

We see that bands 1 and 2, and bands 4 and 5, are very highly correlated.

A principal components analysis (PCA) carried out on this data yielded the following percentage variances explained by the principal axes: 70.5%, 29.25%, 0.25% and 0%. We see here that the 4-band data can be expressed in two new bands, and furthermore that one single synthetic band provided by the first component (an "eigen-band") explains most (i.e., 70.5%) of the variance.

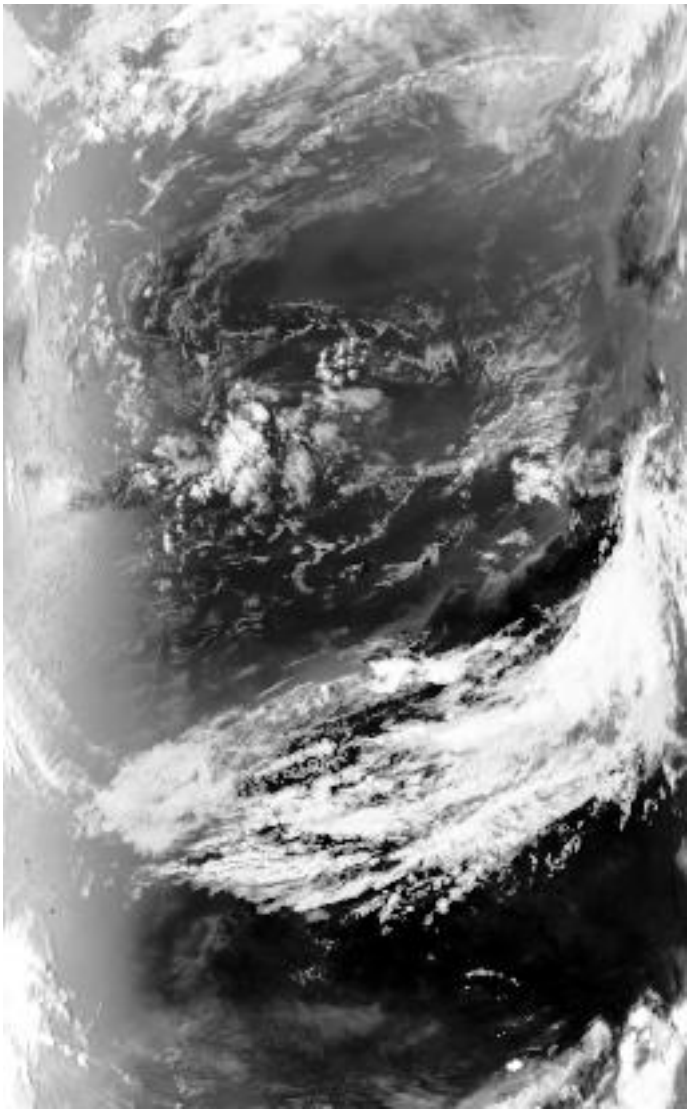

Figure 1. Principal component image of 4-band image data, shown with an 8-fold reduction in image dimensions, and histogram-equalized. The left extremity of the image is affected by sun glint.

Figure 1 shows the principal component image, on the basis of which an optimal graylevel clustering, and segmentation, will now be sought.

5. CLUSTERING AND SEGMENTATION

Two methods were applied to the principal component image. The results are shown in Table 1.

Clusters	$\hat{\phi}$	NegLogPL	BIC	PLIC
1	–	–	−67385806	−30696093
2	0.69	313171	−61625269	−4290711
3	0.69	583923	−60922631	−47581577
4	0.69	1065104	−60687086	−44727447
5	0.73	1425155	−60656468	−47109488
6	0.74	1485380	−60615300	−44369923
7	0.78	1752173	−60596591	−47602125
8	0.80	1849985	−60596352	−51635014
9	0.83	2160247	−60567044	−51973854
10	0.83	2445704	−60588785	−50832573
11	0.81	2668640	−59512566	−40948235
12	0.82	3004705	−60554098	−48542503
13	0.83	3177943	−59000604	−40627639

Table 1. Clustering or quantization, and Markov model-based spatial clustering, of the principal component image shown in Figure 1, for various number of clusters. BIC was determined for grayscale clustering or quantization. PLIC was determined for spatial clustering.

1. A mixture fit of Gaussians to the grayscale marginal distribution, leading to a grayscale clustering, or quantization. In this case, the BIC (Bayes Information Criterion) provided a test for the number of clusters. We begin with $K = 1$ cluster, and then increase K only as long as the BIC value increases. We are therefore sequentially comparing a model with K clusters to a model with $K + 1$ clusters, until the $K + 1$ cluster model fails to outperform the K cluster model, as judged by the BIC. In some cases, we find that BIC values reach an approximate plateau. In Table 1 we find the first maximum to be at $K = 9$. See Figure 2.

2. Segmentation is provided by a conditional likelihood of observed data conditioned on the segment labels which is Gaussian (implying, therefore, a Gaussian mixture of segments model); and a Markov prior which is given by a Potts or Ising spatial influence function. The Potts parameter ϕ, with estimated value $\hat{\phi}$, indicates the extent of spatial influence. The segmentation is initialized using a Gaussian mixture fit to the grayscale marginal distribution (i.e., method 1 which we used above). The generalization of the BIC goodness of fit criterion for the spatially-influenced segmentation is the PLIC, pseudo-likelihood information criterion. In Table 1, we exclude the case of $K = 1$ as uninteresting, and find that $K = 11$ segments gives the first minimum PLIC value. See Figure 3.

Figures 2 and 3, therefore, are the best results for, respectively, Gaussian marginal mixture modeling, and Markov spatial clustering. These figures are associated, respectively, with 9 and 11 classes.

To further assess these results, we took as ground truth the cloud map currently provided as a data product. This cloud map is shown in Figures 4 and 5.

Our results in Figures 2 and 3 can be compared visually with the ground truth in Figure 5. However, Figure 2 contains 9 class labels. Figure 3 contains 11 class labels. By design the ground truth in Figure 5 contains labels 0 and 1, only, in order to facilitate comparisons. Quite simply, therefore, we looked at all possible booleanizations of Figures 2 and 3. We looked at the Euclidean distance between Figure 5 and booleanized Figure 2; and we looked also at the Euclidean distance between Figure 5 and booleanized Figure 3.

The best fit between our clustering results and the ground truth image are shown in Figures 6 and 7.

From Table 2, the superiority of the spatially-based Markov model clustering is apparent. However we also note the lower false alarm rate, i.e. pixels found to be cloud in our results but not characterized as cloud in the ground truth image, in the case of the non-spatially based clustering.

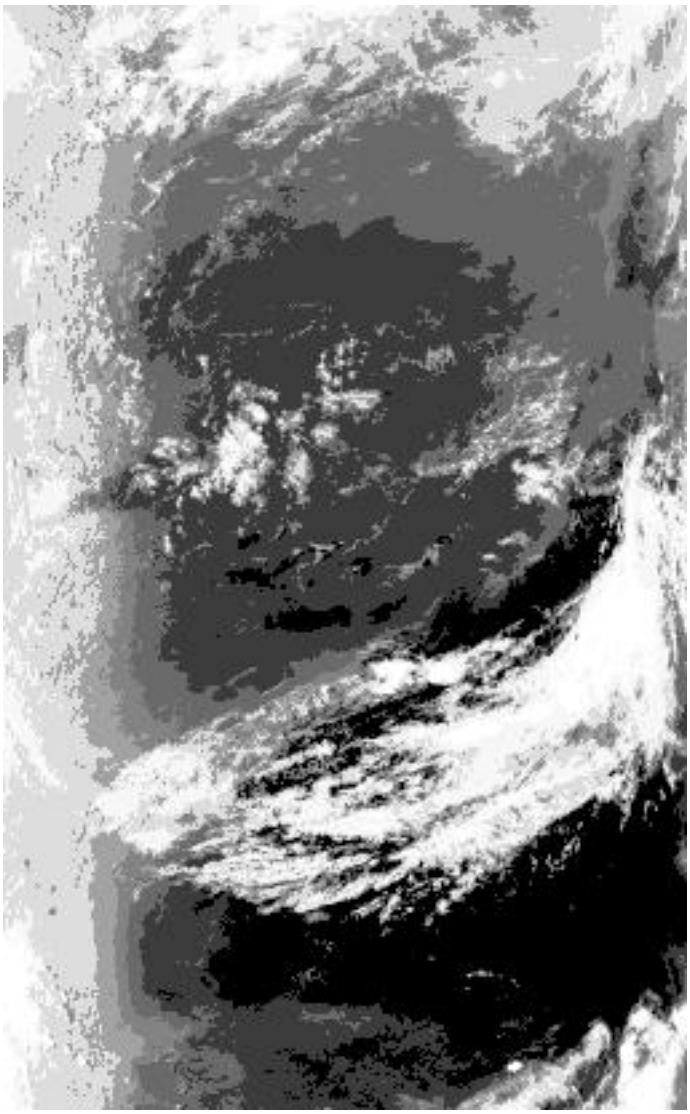

Figure 2. Gaussian marginal mixture model result, with $K = 9$ components. The image is shown contrast-stretched (histogram-equalized), and reduced 8-fold in image row and column dimensions.

	Quantized	Spatial
Cloud pixels recovered (% of ground truth)	86.58%	96.67%
Cloud pixels lost (% of ground truth)	13.42%	3.33%
False alarms (as % of result)	2.98%	15.91%

Table 2. Results of comparing clustering or quantization result in Figure 6, and the spatial clustering or segmentation result in Figure 7, with the ground truth image in Figure 5.

6. CONCLUSIONS

This article develops a cloud mask using statistical pattern recognition based on both marginal density clustering and spatial clustering. The results obtained from these approaches produce images which give detailed information about the original satellite bands. When comparing the results with these bands, it is easy to identify that cloudy regions have been well classified. To perform a better validation than simple visual inspec-

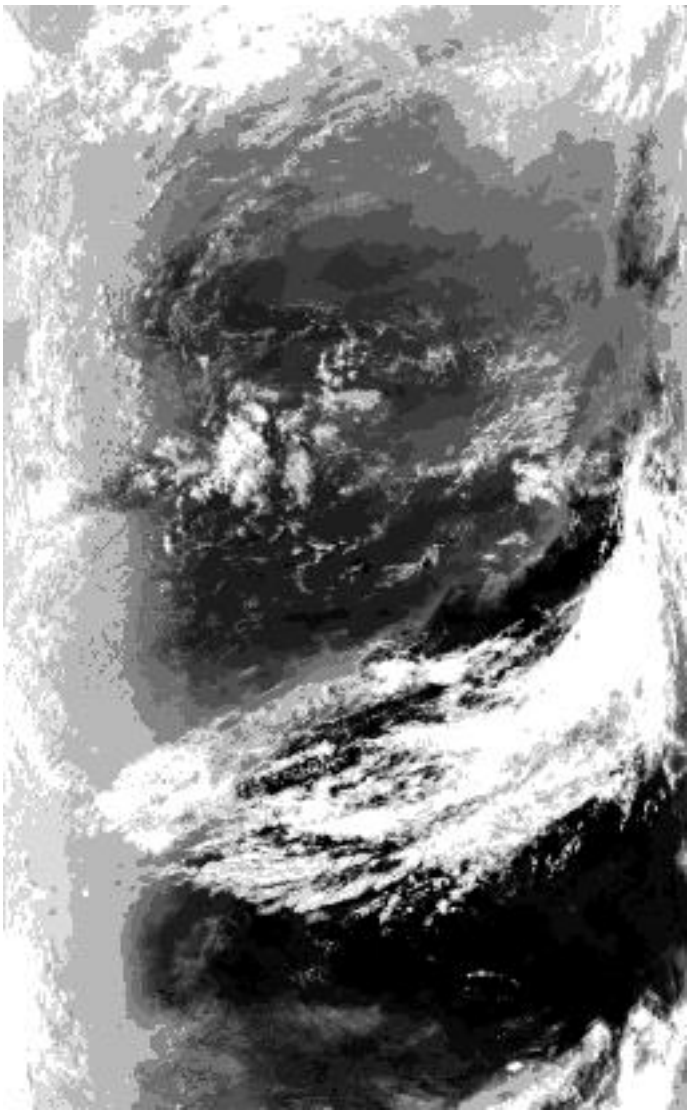

Figure 3. Markov segmentation model result, with $K = 11$ components. The image is shown contrast-stretched (histogram-equalized), and reduced 8-fold in image row and column dimensions

tion, a cloud mask obtained using thresholds in the University of Las Palmas de Gran Canaria has been taken as ground truth. Spatial clustering yields better results according to well-classified cloud pixels, but marginal density clustering produces a better false alarm rate. It is important to take into account that validation of satellite products is a very difficult issue. Therefore the ground truth cloud mask has some errors although it is good enough to make comparisons with.

A number of cloud detection algorithms are already in use, but all either need set thresholds or training sets for cloud detection and operate in localised geographic areas only. The clustering techniques introduced in this article seem to be very useful for cloud detection, especially because they provide the possibility of finding a cloud mask for every sensor and every geographic area.

Figure 4: Cloud map determined by standard algorithm based on thresholds.

REFERENCES

1. G. McLachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcek Dekker, 1988.
2. G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, 1997.
3. R. Kass and A. Raftery, "Bayes factors," *Journal of the American Statistical Association* **90**, pp. 773–795, 1995.
4. G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics* **6**, pp. 461–464, 1978.
5. D. Stanford, *Fast Automatic Unsupervised Image Segmentation and Curve Detection in Spatial Point Patterns*. PhD thesis, Department of Statistics, University of Washington, 1999.
6. D. Stanford and A. Raftery, "Determining the number of colors or gray levels in an image using approximate Bayes factors: the pseudolikelihood information criterion (PLIC)," tech. rep., Department of Statistics, University of Washington, 2001.

Figure 5: A thresholded version of the image shown in Fig. 4.

7. J. Besag, "Statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B* **48**, pp. 259–302, 1986.
8. J. Besag, "Statistical analysis of non-lattice data," *Statistician* **24**, pp. 179–195, 1975.
9. M. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association* **96**, pp. 746–774, 2001.

Figure 6. A thresholded version of the Gaussian mixture model result shown in Fig. 2. The latter image is thresholded at label 6 (among labels 1 to 9) which yields the best match with Fig. 5.

Figure 7. A thresholded version of the Markov segmentation model result shown in Fig. 3. The latter image is thresholded at label 9 (among labels 1 to 11) which yields the best match with Fig. 5.