Image Time-Series Data Mining Based on the Information-Bottleneck Principle

Lionel Gueguen and Mihai Datcu, *Senior Member, IEEE*

***Abstract*—Satellite image time series (SITS) consist of a time sequence of high-resolution spatial data. SITS may contain valuable information, but it may be deeply hidden. This paper addresses the problem of extracting relevant information from SITS based on the information-bottleneck principle. The method depends on suitable model selection, coupled with a rate–distortion analysis for determining the optimal number of clusters. We present how to use this method with the Gauss–Markov random fields and the autobinomial random fields model families in order to characterize the spatio-temporal structures contained in SITS. Experimental results on synthetic data and SITS from SPOT demonstrate the performance of the proposed methodology.**

***Index Terms*—Gibbs–Markov random field, information bottleneck, satellite image time series (SITS), soft clustering, unsupervised clustering.**

I. Introduction

NOWADAYS, huge quantities of satellite images are available thanks to the growing number of satellite sensors. A given scene can be observed repeatedly from space, resulting in satellite image time series (SITS). The high spatial resolution of the sensors give access to detailed spatial structures, which after series of revisits, are extended to spatio-temporal data structures. It follows that SITS are highly complex data sets that potentially contain valuable spatio-temporal information. For example in SITS, growth, maturation, or harvest of crops can be observed. Also, many applications for global monitoring and security need extraction of relevant information regarding the evolution of scene structures or objects. Specialized tools for information extraction in SITS have been devised in order to perform change detection, monitoring, or validation of physical models. However, these techniques usually are dedicated to specific applications. Consequently, in order to exploit the information contained in SITS, general analytical methods are required. Some methods for low-resolution images acquired at uniform sampled times have been studied in [1]. For high-resolution and nonuniform time-sampled SITS, new spatio-temporal analyzing algorithms are presented in [2] and [3]. They are based on a Bayesian hierarchical model of information content. The concept was first introduced in [4]–[6] for information mining in remote-sensing-image archives. The method is based on the synergy of two representations of the information: objective and subjective. The objective information extraction is a data-driven approach, while the subjective part is user driven. In fact, the subjective representation is obtained from the objective representation by machine learning under the constraints provided by a user. The advantage of such a concept is that it is free of the application specificity and adapts to the user's query.

This paper addresses the problem of objective representation of the information by unsupervised clustering and model selection. Based on the results of Bayesian inferences, the model selection is approached to better explain the SITS information content. Rate–distortion theory is used to analyze the information content of SITS as represented by clustering in a features space. Second, we present the models used for characterizing the spatio-temporal patterns and the relevant information. These models have been inspired from texture analysis and belong to the family of parametric Gibbs Markov random fields. Finally, we present an informational approach based on the information-bottleneck principle to compute an unsupervised clustering. The two methods of analysis have been unified in the framework of information bottleneck in order to compute a clustering of spatio-temporal patterns with the optimal number of clusters. This information-bottleneck principle has been used previously for clustering word or images [7], [8], and our approach generalizes this method by embedding a model selection and by determining an optimal number of clusters.

This paper is organized as follows. Section II introduces the information theoretical concept for spatio-temporal pattern detection and recognition. Section III presents the spatio-temporal patterns informational characterization. Section IV introduces the theory about the information-bottleneck principle. Section V presents the information-bottleneck approach for unsupervised clustering. Experiments and discussion are detailed in Section VI. Section VII concludes this paper.

II. Information Theoretical Concept for Pattern Detection and Recognition

A. Bayesian Approach

For efficient detection or recognition of spatio-temporal patterns, it is essential to characterize information in a low-dimensional space. To this end, features are extracted by fitting parametric models to data. This task can be viewed as a Bayesian hierarchical model in two stages [9], [10]. The second

Manuscript received December 20, 2005; revised November 18, 2006. This work was supported by the Competence Center in the field of Information Extraction and Image Understanding for Earth Observation funded by the French Space Agency (CNES), the German Aerospace Center (DLR), and École Nationale Supérieure des Télécommunications (ENST) Paris.

L. Gueguen is with the École Nationale Supérieure des Télécommunications (ENST) Paris, 75013 Paris, France, and the French Space Agency (CNES) 31401 Toulouse, France (e-mail: lionel.gueguen@enst.fr).

M. Datcu is with the German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany, and the École Nationale Supérieure des Télécommunications (ENST) Paris, 75013 Paris, France (e-mail: mihai.datcu@dlr.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2006.890557

level of inference is the model selection, and the first level of inference is the model fitting. Considering a stochastic process X which models the observations, $\mathcal{M}$ as a parametric model of X, and Θ as the parameters of the model, the Bayesian inferences are done using the Bayes rule. The model selection is done by maximizing the probability $p(\mathcal{M}|X)$ expressed in (1). Since the parametric model $\mathcal{M}$ is available, expressed as $p(X|\Theta,\mathcal{M})$, the model conditional likelihood usually named model evidence $p(X|\mathcal{M})$ is obtained by marginalization (2). The prior distribution of the parameters is required for calculation and is often considered to be uniform. The first Bayesian inference is parameters' estimation. It consists in choosing the parameters $\hat{\Theta}$ that maximize the *a posteriori* parameters probability (3). Finally, the signal is characterized by the estimated parameters

$$p(\mathcal{M}|X^n) = \frac{p(X^n|\mathcal{M})\,p(\mathcal{M})}{p(X^n)} \tag{1}$$

$$p(X^n|\mathcal{M}) = \int p(X^n|\Theta,\mathcal{M})p(\Theta|\mathcal{M})\,d\Theta \tag{2}$$

$$p(\Theta|X^n,\mathcal{M}) = \frac{p(X^n|\Theta,\mathcal{M})\,p(\Theta|\mathcal{M})}{p(X^n|\mathcal{M})}. \tag{3}$$

B. Rate–Distortion Approach

In the previous section, the choice of the model is done first. Then, the parameters are estimated, and an optimal clustering is calculated. Consequently, the clustering is totally dependent on model selection. As a result, there is no review or refinement of model selection after calculating the clustering. Our objective is to have a methodology to jointly choose the model and the clustering. On one hand, minimum-description-length-like criteria [11] state that the best model minimizes the entropy of the signal expressed with it and minimizes its own entropy. The entropy of the model is linked to its complexity, which is usually determined by the parametric dimension and the number of realizations. Considering that the signal is represented by the set of cluster centroids C in the features space, we want to minimize the entropy of the representation expressed as the mutual information between the signal and the centroids, $I(X,C) = H(C) - H(C|X)$. C is a random variable, and the centroid c is a realization of C

$$H(C) = -\sum_{c} p(c)\log p(c) \tag{4}$$

$$I(X,C) = \sum_{x,c} p(x,c)\log\frac{p(x,c)}{p(x)p(c)}. \tag{5}$$

Consequently, choosing a simpler model leads to minimizing $H(C)$ and $I(X,C)$, because the features space is less complex to represent. Moreover, from a compression point of view, a clustering can be seen as a vector quantization, where a distortion functional $d(X,C)$ (8) is minimized. The distortion functional measures the quality of the signal representation C. Combining the previous observations, the problem can be viewed as a rate–distortion problem (6). There is a tradeoff between the amount of information conserved (distortion) and the complexity of representation (rate). Considering the signal X, the features Θ, the clusters C, and the model $\mathcal{M}$, the problem can be viewed as a minimization of the following functional:

$$\min_{c,\,p(c|x,\mathcal{M}),\mathcal{M}} I(X,C) + \beta E_{X,C}\left[d(X,C)\right] \tag{6}$$

where

$$p(c|x,\mathcal{M}) = \int p(c|\Theta,\mathcal{M})p(\Theta|x,\mathcal{M})\,d\Theta \tag{7}$$

$$d(x,c) = d(\hat{\Theta}(x),c) \tag{8}$$

$$E_{X,C}\left[d(X,C)\right] = \sum_{x,c} p(x,c)d(x,c). \tag{9}$$

At the minimum of the previous functional, we define the distortion D_β and the rate R_β as

$$D_\beta = E_{X,C}\left[d(X,C)\right] \tag{10}$$

$$R_\beta = I(X,C) \tag{11}$$

where β is a tradeoff weighting between the distortion and the rate. For example, when $\beta \to 0$, simpler models are chosen to obtain only a few clusters. On the contrary, when $\beta \to \infty$, more complex models are chosen to fit data and more clusters are obtained. Therefore, from this rate–distortion approach, a clustering and a model can be calculated with β controlling the tradeoff between compression and distortion.

C. Determining the Natural Number of Clusters

In this section, we give an example of a rate–distortion approach for determining the optimal number of clusters. Such a method has been studied in [12]. It demonstrates theoretically and experimentally that the natural number of clusters can be determined from the rate–distortion curve. To explain the principle, we present the following example. Let a vectorial signal X consist of a mixture of K Gaussians (12). The Gaussian distributions each have distinct centroids μ_k and the same variance σ. An example of the realizations of X is shown in Fig. 1, for $K = 5$

$$p(X) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mu_k,\sigma). \tag{12}$$

A rate–distortion analysis is performed on data by a vector quantization. The k-means algorithm [13] is used to process the clustering with varying number of initial cluster centroids. Each realization x of X is associated with the nearest cluster centroid $c(x)$. Consequently, the conditional probability $p(c|x)$ is expressed by

$$p(c|x) = \begin{cases} 1, & \text{if } c = c(x) \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

In our special case, the k-means algorithm minimizes the rate–distortion criterion (6). Indeed, it calculates the cluster centroids c and the optimal assignments $p(c|x)$ and implicitly select the optimal model, which is dependent on the number

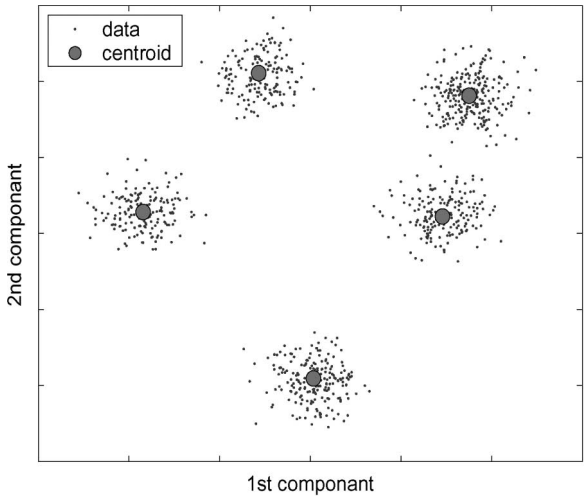


Fig. 1. Displayed are 1000 realizations of a stochastic process X generated from a mixture of five Gaussians. The five centroids μ_k of the Gaussians are represented by the big dots. The five Gaussian distributions have the same variance.

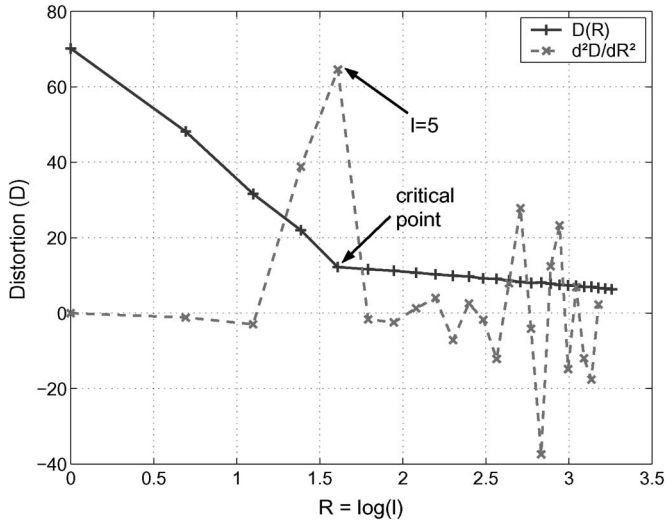


Fig. 2. Rate–distortion curve $D(R)$ and its second derivative $(\partial^2 D)/(\partial R^2)$. The rate–distortion curve has been obtained by clustering the signal generated with a mixture of five Gaussians. Two behaviors of the curve are noticeable: a strong decrease followed by a slow decrease. The second derivative highlights this change of behaviors, which corresponds to find the natural number of clusters. The second derivative has been computed with the differences which generate the effect of oscillations.

of clusters. If l is the number of clusters, the implicit selected model $\mathcal{M}$ is a mixture of Gaussian distributions defined by

$$p(X) = \frac{1}{l} \sum_{c} \mathcal{N}(c, \sigma_c) \tag{14}$$

where σ_c is the intracluster variance of the cluster represented by c. The mutual information $I(X, C)$ is deduced from the conditional probability $p(c|x)$. We approximate the mutual information by (18) considering that the clusters are equally distributed

$$I(X,C) = \sum_{c,x} p(c|x)\, p(x) \log \frac{p(c|x)}{p(c)} \tag{15}$$

$$= \sum_{c} \sum_{x, c(x)=c} -p(x) \log p(c) \tag{16}$$

$$= -\sum_{c} p(c) \log p(c) \tag{17}$$

$$\approx -\log\left(\frac{1}{l}\right). \tag{18}$$

We choose the Euclidean distance between the signal and the quantizers to define a distortion measure (19) which corresponds to the sum of the intracluster variances

$$E_{X,C}\left[d(X,C)\right] = \sum_{c,x} (x-c)^{\mathrm{T}} (x-c) p(c|x) p(x) \tag{19}$$

$$= \sum_{c} \sum_{x, c(x)=c} \left(x - c(x)\right)^{\mathrm{T}} \left(x - c(x)\right)\, p(x) \tag{20}$$

$$= \sum_{x} \left(x - c(x)\right)^{\mathrm{T}} \left(x - c(x)\right) p(x). \tag{21}$$

Consequently, a rate–distortion curve can be computed. In fact, the number of clusters l plays the role of the tradeoff parameter β. Therefore, the parametric rate–distortion function $D(R)$ is defined parametrically with l. An example of this curve is exhibited in Fig. 2, in which two distinct behaviors of the curve are noticeable. First, when $l \leq K$, D decreases rapidly with the rate. This behavior reveals that clusters fail to correctly represent the signal. Second, when $l \geq K$, D decreases very slowly with the rate which means that there are too many clusters for low distortion gains. Experimentally, the point of $D(R)$ obtained for $l = K$ is a critical point that corresponds to a maximum of the second derivative of $D(R)$. Indeed, due to the change in the decreasing behavior [12], at the critical point, a gap appears in the first derivative that corresponds to a maximum of the second derivative. Therefore, we can estimate the natural number of clusters from the parametric rate–distortion curve by maximizing the second derivative

$$\hat{K} = \arg\max_{l} \frac{\partial^2 D}{\partial R^2} \tag{22}$$

$$= \arg\max_{l} \frac{\partial^2 D}{\partial l^2} \frac{\partial^2 l}{\partial R^2}. \tag{23}$$

III. Spatio-Temporal Pattern Informational Characterization

In the following sections, we present two families of Gibbs–Markov random fields. SITS are high-complexity data; however they preserve spatial and temporal dependencies. Thus, the interest of Gibbs–Markov random fields is to discover and characterize these patterns. In addition, these parametric models are used to represent the relevant information, which is extracted in the information-bottleneck framework.

A. Gauss–Markov Random Fields (GMRF)

GMRF have interesting properties for characterizing textures in satellite images [10], [14]. GMRF are parametric models, which we use to model a 3-D random field. We consider that the random variable X is a random field defined on a rectangular

order	$dt=-1$	$dt=0$	$dt=1$	$r=(dx,dy,dt)$ and the index $j(r)$
1				r $(-1,0,0)$ $(0,-1,0)$ $(0,0,1)$ $j(r)$ 1 2 3
2				r $(-1,0,0)$ $(-1,-1,0)$ $(1,0,1)$ $j(r)$ 1 2 3 r $(0,-1,0)$ $(1,-1,0)$ $(0,-1,1)$ $j(r)$ 4 5 6 r $(0,0,1)$ $(-1,0,1)$ $(0,1,1)$ $j(r)$ 7 8 9
3				the length of the index is 13 and is too long to be described here

Fig. 3. Symmetric 3-D neighborhood. The pixel X_s is black. Pixels corresponding to X_{s+r} are white, and pixels corresponding to X_{s-r} are gray. dt is the time dimension, and (dx, dy) is the spatial dimension.

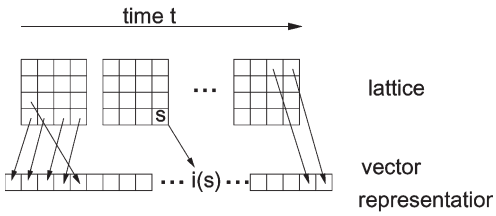


Fig. 4. Lattice Ω is mapped to a 1-D index $i(s)$. The squares represent the spatial grid evolving in time. s is a pixel location, while $i(s)$ is the corresponding location in the vector representation.

grid. Let X_s be the observations, s belonging to a 3-D lattice Ω, and N the half of a symmetric 3-D neighborhood (Fig. 3). Therefore, GMRF are defined as follows:

$$X_s = \sum_{r \in N} \theta_r (X_{s+r} + X_{s-r}) + e_s \tag{24}$$

where e_s is a white Gaussian noise of variance σ_e and θ_r is a scalar associated to each direction in the neighborhood. This Markov random field can be written in a probabilistic way by (25), and the corresponding Gibbs random field is expressed by the following expression [15], [16]:

$$p(X_s|\{X_{s+r}, X_{s-r}, \theta_r, r \in N\}) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp -\frac{1}{2}\frac{e_s^2}{\sigma_e^2} \tag{25}$$

$$p(X) = \prod_{s \in \Omega} p(X_s|\{X_{s+r}, X_{s-r}, \theta_r, r \in N\}). \tag{26}$$

In order to simplify the notation, we propose to introduce mapping functions from the lattice Ω and the neighborhood N to some 1-D indexes $i(s), j(r)$, respectively, as shown in Figs. 3 and 4. Then, the parameter set $\hat{\Theta}$ and the noise variance $\hat{\sigma}$ are estimated by least minimum squares, thus corresponding to a maximum-likelihood estimation when considering a white Gaussian error model. Equation (24) is expressed vectorially in (27) by introducing a matrix G expressed from X. The vectors X and E are formed by the values $X_{i(s)}, E_{i(s)}$, respectively. The vector Θ is composed of the values $\theta_{j(r)}$. The matrix G is defined by $G(i(s), j(r)) = X_{s+r} + X_{s-r}$. Hence, the estimated parameters are expressed by the following equations:

$$X = G\Theta + E \tag{27}$$

$$\hat{\Theta} = (GG^T)^{-1} G^{\mathrm{T}} X \tag{28}$$

$$\hat{\sigma}^2 = X^{\mathrm{T}} X - (G\hat{\Theta})^{\mathrm{T}} (G\hat{\Theta}). \tag{29}$$

With the previous equations, the problem is formulated as a linear system. From the estimated parameters, it is possible to approximate the model evidence (30), in the case of GMRF. A general formulation of the model evidence for linear systems is given in [17] and [18]. Considering N, Q being the respective dimension of X, Θ, the model conditional likelihood is given by

$$p(X|M) \approx \frac{\pi^{-N/2} \Gamma\left(\frac{Q}{2}\right) \Gamma\left(\frac{P-Q}{2}\right) |G^{\mathrm{T}} G|^{-1/2}}{4 R_\delta R_\sigma (\hat{\Theta}^{\mathrm{T}} \hat{\Theta})^{Q/2} \hat{\sigma}^{P-Q}} \tag{30}$$

where R_δ, R_σ are arbitrary constants. Then, by a two-stage Bayesian approach, it is possible to characterize spatio-temporal patterns. First, we choose the order of the model (Q) and the analyzing window size (N) by selecting the greatest model evidence. Then, the parameters are estimated. By introducing a distance in the features space, it is possible to efficiently compare spatio-temporal structures through their representatives. By doing this space transposition, the dimensionality of data to be treated is mainly reduced while the similarity between spatio-temporal structures is retained.

B. Autobinomial Gibbs Random Field

Autobinomial Gibbs random fields belong to the family of Gibbs stochastic processes. Like GMRF, they have interesting properties for characterizing textures [10], [19]. The field is defined as in the previous section. The maximum gray value in the image is $\mathcal{G}$ and $\binom{n}{k} = (n!)/(k!(n-k)!)$. The autobinomial

model is defined by introducing an energy function H and a partition function PF

$$H(X_s, N, \Theta) = -\log \binom{\mathcal{G}}{X_s} - X_s \eta \tag{31}$$

$$\eta = \theta_0 + \sum_{r \in N} \theta_r \frac{X_{s+r} + X_{s-r}}{\mathcal{G}} \tag{32}$$

$$\mathrm{PF}(N, \Theta) = \sum_{x_s} e^{-H(x_s, N, \Theta)}. \tag{33}$$

The probability distribution of the Gibbs random field is linked by (26) to a Markov random field described by the following equations:

$$p(X_s | \{X_{s+r}, X_{s-r}, \theta_r, r \in N\}) = \frac{1}{\mathrm{PF}(N, \Theta)} e^{-H(X_s, N, \Theta)} \tag{34}$$

$$p(X_s | \{X_{s+r}, X_{s-r}, \theta_r, r \in N\}) = \binom{\mathcal{G}}{X_s} q^{X_s} (1-q)^{G - X_s} \tag{35}$$

$$q = \frac{1}{1 + \mathrm{e}^{-\eta}}. \tag{36}$$

The parameters are computed with a conditional least squares estimator (CLSE) [20]. By making some approximations, the problem is reduced to a system of linear equations as in (27) with a different definition of X and G [10]. Therefore, the parameters are estimated as in (28) and (29). Since the system is linear, the model evidence can be approximated by (30) (see appendix for details of the calculation). An interesting property of autobinomial models is that they converge to GMRF when $\mathcal{G}$ tends to infinity. In consequence, the autobinomial model family is a richer class of parametric models that can approximate GMRF.

IV. Information-Bottleneck Principle

The marginal Bayesian inference and the rate–distortion approach are integrated in the information-bottleneck framework described in this section. Following, a theoretical introduction, we describe an algorithm of soft clustering derived from the principle. Finally, we show how to use this algorithm to cluster spatio-temporal events in SITS. In this section, uppercase letters are used to name random variables, and lowercase letters are used to indicate realizations of a random variable.

A. Problem Formulation

The information-bottleneck criterion emerged from rate–distortion theory. The problem is stated as follows [21]. We would like a relevant quantizer $\tilde{X}$ to compress X as much as possible under the constraint of a distortion measure between X and $\tilde{X}$. In contrast, we also want to capture as much information as possible about a third variable Y. For example, we would like to capture information from a set of models $\mathcal{M}$. In effect, we pass the information that X provides about Y through a bottleneck formed by the compact summary in $\tilde{X}$. An equivalent formulation is that we want to minimize the loss of mutual information caused by the compact representation of data, while the mutual information between $\tilde{X}$ and X is minimized too. There is a tradeoff between these two quantities that is controlled by the parameter β

$$\min_{p(\tilde{x}|x)} I(\tilde{X}, X) + \beta \left\{ I(X, Y) - I(\tilde{X}, Y) \right\}. \tag{37}$$

It is equivalent to the following criterion, because $I(X, Y)$ does not vary with $p(\tilde{x}|x)$:

$$\min_{p(\tilde{x}|x)} I(\tilde{X}, X) - \beta I(\tilde{X}, Y). \tag{38}$$

The assumption of the following Markov chain is made: $Y \leftrightarrow X \leftrightarrow \tilde{X}$. Banerjee demonstrated in [22] that information bottleneck can be viewed as a rate–distortion problem based on Bregman divergence [22]. He considered $Z = p(Y|X)$ and $\tilde{Z} = p(Y|\tilde{X})$ as sufficient statistics for X and $\tilde{X}$, respectively. Z takes its values over the set of the conditional distributions $\{p(Y|x)\}$, and $\tilde{Z}$ takes its values over the set of the conditional distributions $\{p(Y|\tilde{x})\} = \tilde{\mathcal{Z}}_s$. First, as Z and $\tilde{Z}$ are sufficient statistics, it is demonstrated that $p(\tilde{x}, x) = p(\tilde{z}, z)$ and $I(X, \tilde{X}) = I(Z, \tilde{Z})$. Second, considering the Markov chain described above, the following equations are obtained:

$$p(y, \tilde{x}|x) = p(y|x)\, p(\tilde{x}|x) \tag{39}$$

$$p(y, \tilde{x}) = \sum_x p(y|x)\, p(\tilde{x}|x) p(x) \tag{40}$$

$$d(x, \tilde{x}) = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})} \tag{41}$$

$$I(X, Y) - I(\tilde{X}, Y) = \sum_{\tilde{x}, x} d(x, \tilde{x}) p(\tilde{x}, x). \tag{42}$$

d is a Bregman divergence defined for Z and $\tilde{Z}$ that corresponds in this case to the Kullback–Leibler divergence $d(x, \tilde{x}) = d(z, \tilde{z})$. Therefore, the Bregman divergence is equal to the information loss

$$E_{Z,\tilde{Z}} [d(z, \tilde{z})] = E_{X,\tilde{X}} [d(x, \tilde{x})] \tag{43}$$

$$E_{Z,\tilde{Z}} [d(z, \tilde{z})] = I(X, Y) - I(\tilde{X}, Y). \tag{44}$$

Consequently, the problem equivalent to the information-bottleneck criterion is written as the following rate–distortion problem using the Kullback–Leibler divergence to define the distortion measure on $\tilde{\mathcal{Z}}_s$ and $\mathcal{Z}_s$ spaces:

$$\min_{\tilde{\mathcal{Z}}_s, p(\tilde{z}|z)} I(Z, \tilde{Z}) + \beta E_{Z,\tilde{Z}} [d(z, \tilde{z})]. \tag{45}$$

B. Optimal Solution

The optimal solutions for minimizing the information-bottleneck criterion have been expressed analytically in [21]. More general solutions have been found to solve the problem of (45). Cover and Thomas gave the solutions in [24]. They considered the Lagrangian $\mathcal{L}$ expressed as

$$\mathcal{L} = I(Z, \tilde{Z}) + \beta E_{Z,\tilde{Z}} [d(z, \tilde{z})] + \sum_{z, \tilde{z}} \lambda(z, \tilde{z}) p(\tilde{z}|z). \tag{46}$$

Deriving the Lagrangian $\mathcal{L}$ and finding the roots leads to find a local optimum of the problem

$$\frac{\partial \mathcal{L}}{\partial p(\tilde{z}|z)} = 0 \tag{47}$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{z}} = 0. \tag{48}$$

First, for a fixed set $\tilde{\mathcal{Z}}_s$, the solution to (47) is given by

$$p(\tilde{z}|z) = \frac{p(\tilde{z})}{N(z,\beta)} e^{-\beta d(z,\tilde{z})} \tag{49}$$

$$N(z,\beta) = \sum_{\tilde{z}} p(\tilde{z}) e^{-\beta d(z,\tilde{z})}. \tag{50}$$

Second, for fixed probabilistic assignments $p(\tilde{z}|z)$, the solution to (48) is given by

$$\begin{aligned} \tilde{z} &= E_{Z|\tilde{z}}[Z] \\ &= \sum_{z} z p(z|\tilde{z}). \end{aligned} \tag{51}$$

It should be noted that (49) and (51) are interdependent, thus, making the solutions difficult to compute.

C. Algorithm

Using the two properties (49), (51), Banerjee proposed in [22] and [23] an iterative algorithm to compute $\tilde{\mathcal{Z}}_s$ and $p(\tilde{z}|z)$. This algorithm is used to solve the problem and to reach a local optimum of the functional. This section presents the proposed algorithm. Moreover, this method makes it possible to compute the divergence D_β and the rate R_β. The algorithm takes in input $\{p(z_i)\}_{i=1}^n$, the set $\mathcal{Z} = \{z_i\}_{i=1}^n$, the tradeoff parameter β, and the number of $\tilde{z}$ denoted k. It returns the optimal set $\tilde{\mathcal{Z}}_s = \{\tilde{z}_h\}_{h=1}^k$ and the optimal conditional distribution $\{p(\tilde{z}_h|z_i)\}_{\substack{1\le h\le k \\ 1\le i\le n}}$. In the following description, we denote by V^t a variable V at the step t of the algorithm. The following procedure is iterated until convergence after initializing $\tilde{z}_h$ and $p(\tilde{z}_h)$. It is demonstrated that the method converges to a local minimum

$$(a1) \quad \forall i \quad N^t(z_i,\beta) = \sum_h p^t(\tilde{z}_h^{t-1}) e^{-\beta d(z_i,\tilde{z}_h^t)}$$

$$(a2) \quad \forall i,h \quad p^{t+1}\left(\tilde{z}_h^t|z\right) = \frac{p^t(\tilde{z}_h^{t-1})}{N^t(z_i,\beta)} e^{-\beta d(z_i,\tilde{z}_h^t)}$$

$$(a3) \quad \forall h \quad p^{t+1}\left(\tilde{z}_h^t\right) = \sum_i p^{t+1}\left(\tilde{z}_h^t|z_i\right) p(z_i)$$

$$(a4) \quad \forall h \quad \tilde{z}_h^{t+1} = \sum_i p^{t+1}\left(z_i|\tilde{z}_h^t\right) z_i.$$

The rate and the distortion are computed using (52) and (53). Then, by varying the parameter β, the following rate–distortion curve is computed. In addition, to obtain consistent calculus when increasing β, the following algorithm should be derived in a simulated annealing process [25] in order to avoid local minima:

$$D_\beta = \sum_{i,h} p(z_i) p(\tilde{z}_h|z_i) d(z_i,\tilde{z}_h) \tag{52}$$

$$R_\beta = \sum_{i,h} p(z_i) p(\tilde{z}_h|z_i) \log \frac{p(\tilde{z}_h|z_i)}{p(\tilde{z}_h)}. \tag{53}$$

V. Information-Bottleneck Approach for Clustering

The information-bottleneck principle has been used for clustering in a variety of contexts. For example, a word-clustering method is presented in [7], and an image-clustering method is presented in [8]. These methods take into account the relevant information contained in a feature space of a predetermined parametric model. Our method presents a general framework of these ideas, where the relevant information is contained in a model space. This generalization highlights the embedded-model selection and enables to compare models of different dimensions.

A. Information Bottleneck Used for Data Characterization

We consider that X contains the observations, and $\tilde{X}$ is the summary of X. The SITS is partitioned in parallelepipeds of fixed size, which are the realizations of the random field X. Let the model $\mathcal{M} = Y$ be the random variable that contains the relevant information. $\mathcal{M}$ takes its values in the set of models composed of GMRF and autobinomial models of three first orders. These models represent the relevant information that we want to extract from SITS. Given these models, we focus on the spatio-temporal texture information extraction, and the information will be represented as a clustering of the realizations of X. Therefore, the information bottleneck gives a formalism to express the tradeoff between compression (short summary) and the relevant information contained in the summary. Thus, this principle indicates how much information can be extracted from the data by a predetermined set of models. Consequently, the problem to be solved is

$$\min_{p(\tilde{x}|x)} I(X,\tilde{X}) - \beta I(\tilde{X},\mathcal{M}). \tag{54}$$

Considering $z = p(\mathcal{M}|x)$ and $\tilde{z} = p(\mathcal{M}|\tilde{x})$, this may be rewritten as in (45). To solve this problem, we need to evaluate $p(z)$ and z. The other variables are evaluated with the Banerjee algorithm presented in the previous section.

B. Calculus of Model Evidence

Calculating model evidence is not an easy task and requires some precautions. Indeed, only the logarithm of the evidence is calculable, and some preprocessing is required. First, in the case of a general linear model, model evidence (30) is known proportionally to a constant $R_\delta R_\sigma$. However, this constant may vary with the models. Therefore, we normalize the model evidence by assuming that the conditional probabilities integrate to unity for each realization m of the model $\mathcal{M}$

$$\forall \quad m, \qquad \sum_x p(x|m) = 1. \tag{55}$$

Then, we evaluate $z = p(\mathcal{M}|x)$ by using the Bayes rule. We assume that $p(\mathcal{M})$ can be calculated and that the conditional

probability $p(\mathcal{M}|x)$ integrates to one for each realizations x of the random variable X

$$\sum_m p(m|x) = 1 \tag{56}$$

$$p(\mathcal{M}|x) = \frac{p(x|\mathcal{M})\, p(\mathcal{M})}{\sum_m p(x|m)\, p(m)}. \tag{57}$$

To compute $p(\mathcal{M})$, the first possibility to consider is that $\mathcal{M}$ follows a uniform distribution. In other words, no prior knowledge on models is introduced. A second possibility to evaluate $p(\mathcal{M})$ is done by counting occurrences of Bayesian choice. After normalization (55), the best model m_x is chosen by maximum likelihood for each x. For each realization m of $\mathcal{M}$, we count the number of times it has been chosen. Then, $p(\mathcal{M})$ is derived from occurrence counts by normalizing with the number of realizations of X, denoted by N_x

$$\forall \quad x, \qquad m_x = \arg\max_m p(x|m) \tag{58}$$

$$\forall \quad m, \qquad p(m) = \frac{1}{N_x} \sum_x \delta(m, m_x) \tag{59}$$

$$\delta(m_1, m_2) = \begin{cases} 1, & \text{if } m_1 = m_2 \\ 0, & \text{otherwise.} \end{cases} \tag{60}$$

However, computing $p(x|\mathcal{M})$ is not feasible because of the required precision. Only the log-evidence $\log p(x|\mathcal{M})$ can be calculated. We shift the log-evidence in order to be able to calculate quantities proportional to the evidence (30) before normalizing by considering that conditional probabilities integrate to one (55). For example, the shift is done to set the maximum log-evidence to zero. Finally, since we have computed z, we estimate $p(z)$ using a histogram based on a Parzen window defined as follows:

$$\mathcal{K}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^{\mathrm{T}} z}{2}\right) \tag{61}$$

$$p(z_0) = \frac{1}{N_z \Delta} \sum_z \mathcal{K}\left(\frac{z_0 - z}{\Delta}\right) \tag{62}$$

where a Gaussian kernel $\mathcal{K}$ is used to compute probabilities and N_z is the number of realizations of Z. The parameter Δ is chosen to be small compared to the values of Z in order to have a good estimate. Finally, we compute the probability $p(z)$ for each z by (62).

C. Unsupervised Clustering

Unsupervised clustering using Bregman divergence has been studied in [23]. Our approach makes use of this general framework. The novelty resides in the choice of the variable that contains the relevant information. By making this assumption, our method calculates the optimal clustering while taking into account the relevance of the models. If we consider $\tilde{z}$ being the cluster centroids, the Banerjee algorithm makes it possible to calculate $p(\tilde{z}|z)$ that is a soft clustering that quantifies the probability that z belongs to the clusters $\tilde{z}$. Moreover, we know that $p(\tilde{z}|z) = p(\tilde{x}|x)$. Therefore, the clustering is based on the features $\tilde{X}$ that are not accessible. This clustering method makes use of all models in the set, unlike the techniques that are constrained to the best model. All models contribute to extract distinguishable types of information from data. Finally, to obtain a hard clustering from $p(\tilde{z}|z)$, a centroid $\tilde{z}^*$ is linked to each z by maximum *a posteriori*

$$\tilde{z}^* = \arg\max_{\tilde{z}} p(\tilde{z}|z)\,. \tag{63}$$

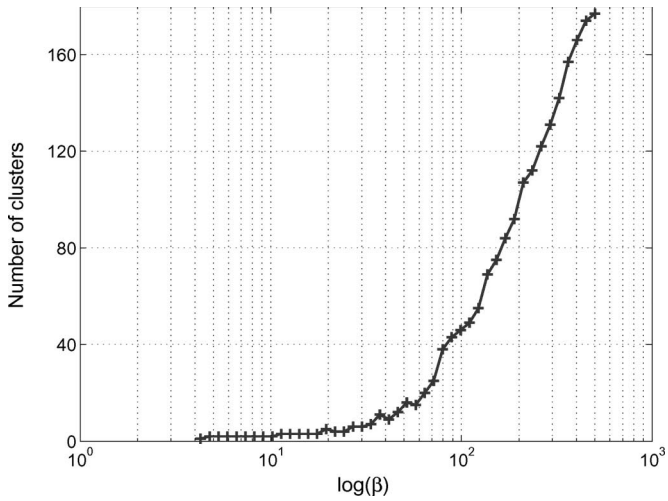


Fig. 5. Variation of the number of centroids with beta. The number of centroids varies almost exponentially with $\log \beta$, which means that the number of centroids varies linearly with β. Therefore, the parameter β determines the number of centroids obtained by the clustering method.

D. Choice of the Optimal Number of Clusters

In the algorithm of the information bottleneck, the number k of $\tilde{z}$ is preset. However, the real number of distinguishable $\tilde{z}$ obtained after optimization is constrained by β. Therefore, the initial number k is chosen to be equal to the number of realizations z. Then, β influences the effective number of clusters found (Fig. 5). As these two quantities are linked, we give a criterion for the optimal choice of β. This criterion is based on the rate–distortion curve $D(R)$, which is a parametric function of β. The optimal $\hat{\beta}$ maximizes the second derivative of $D(R)$ (64), as previously discussed in the Section II-C. This is the critical point on the curve that corresponds to the natural number of clusters. Moreover, local maxima of $(\partial^2 D_\beta)/(\partial R_\beta^2)$ are also critical points in the sense that they highlight subcluster structures contained within clusters. Hence, a natural hierarchical clustering could be derived by calculating clusterings at each local maxima in increasing order. This point is described in [12]

$$\begin{aligned} \hat{\beta} &= \arg\sup_\beta \frac{\partial^2 D_\beta}{\partial R_\beta^2} \\ &= \arg\sup_\beta \frac{\partial^2 D(\beta)}{\partial \beta^2} \left(\frac{\partial^2 R(\beta)}{\partial \beta^2}\right)^{-1}. \end{aligned} \tag{64}$$

VI. Experiments and Discussion

A. Experiments on Synthetic Markov Random Fields

For method evaluation, in this section, we compare the k-means and the AutoClass [26] clustering to the

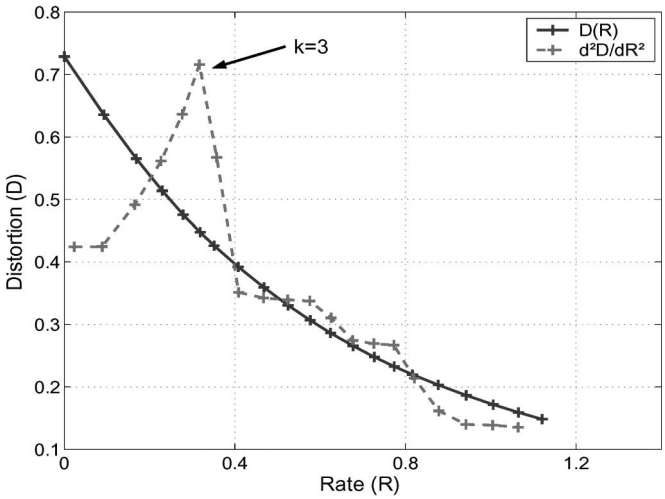


Fig. 6. Rate–distortion curve $D(R)$ and the second derivative of $D(R)$ obtained for the information-bottleneck clustering done on causal autoregressive signals. The curve is discrete, and the second derivative is computed with the differences. The first local maximum determines the optimal number of clusters, which is equal to three. This number is equal to the number of model orders. The second local maximum reveals the existence of subclusters.

information-bottleneck-based clustering on synthetic causal autoregressive signals. Unlike k-means, the AutoClass algorithm has been designed to find the optimal number of clusters. We use three different models to generate the signals. An example of a second-order process is given in (65) with the parameters θ_1, θ_2, and a white Gaussian noise e_n. The first method consists in choosing the best model by comparing model evidences, then processing the k-means algorithm in the parameters space. The second method consists in applying the method presented in Section V and the AutoClass algorithm on the model evidences. In order to generate the signal, we computed the parameters of each model following uniform distributions. Then, from a generated white Gaussian noise, we compute several signals for each model characterized by its order. Since the parametric distribution is known, we compute the model evidence by (2)

$$x_n = \theta_1 x_{n-1} + \theta_2 x_{n-2} + e_n. \tag{65}$$

By the first method, the selected model order is the first one. Then, we estimate parameter for each signal by least minimum squares (28). Finally, we process the k-means algorithm in the feature space taking $k = 3$. By the second method, we apply the method described in Section V-D to calculate the optimal number of clusters. The resulting rate–distortion curve is shown in Fig. 6, and at the critical point, the optimal number of clusters is equal to three. In a sense, the first maxima of the second derivative of $D(R)$ gives the number of models that generate the signals. Then, the following maxima of $(\partial^2 D)/(\partial R^2)$ reveal the existence of subclusters. In Table I, we present the confusion matrix between the results of clustering and the real class of signals, which correspond to the orders of model chosen to generate the signals. We observe that our method performs better than the k-means and AutoClass clustering, indeed these last methods do not distinguish the signals generated by different models. Consequently, for this study case, our method is better suited to extract information from data generated by several parametric models, like GMRF and autobinomial Gibbs random fields. To conclude, this experiment underlines the fact that the information-bottleneck-based method seems to exploit the whole set of parametric models to extract the relevant information.

TABLE I
CONFUSION MATRIX FOR THE k-MEANS, THE AUTOCLASS, AND THE INFORMATION BOTTLENECK CLUSTERINGS. THE RESULTS SHOW THAT THE k-MEANS AND THE AUTOCLASS CLUSTERING DO NOT EXPLOIT THE INFORMATION CONTAINED IN THE MODELS BY GIVING A CONFUSED MATRIX. AUTOCLASS FINDS THREE CLUSTERS AS THE INFORMATION BOTTLENECK, WHILE GIVING A CONFUSED MATRIX

	IB clustering			k-means			AutoClass		
	clusters			clusters			clusters		
class	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
1	**83**	16	1	**38**	41	21	83	10	7
2	27	**52**	21	30	**47**	23	89	7	4
3	7	36	**57**	34	41	**25**	63	31	6

B. Experiments on Sits

For this paper, we have worked on the ADAM[1] data set provided by the Centre National d'Etudes Spatiales (French Aerospace Center). The images, constituting the SITS, have been acquired by three instruments: SPOT 1, 2, and 4 and have a resolution of 20 m. This SITS comprises 38 images of size 3000×2000 and each image contains three spectral bands. The SITS is not uniformly time-sampled, meaning that the elapsed time between two images ranges from one day to one month. The SITS has been intercalibrated to take into account the use of different instruments. In addition, the images have been coregistered with a subpixel precision. As a remark, the information visible in data is the evolution of spatial structures. Therefore, we want to characterize spatio-temporal patterns of this series. For our experiments, we only used the first spectral band and two subsequences of sizes $100 \times 100 \times 7$ and $200 \times 200 \times 7$. We considered the SITS as a realization of a stochastic process. We partitioned the series in time-overlapping parallelepipeds of size $10 \times 10 \times 5$ and we considered each parallelepiped as an independent realization of a stochastic process X. Note that the parallelepipeds do not overlap spatially. Therefore, we have $10 \times 10 \times 7 = 700$ realizations of the process for the first sequence and $20 \times 20 \times 7 = 2800$ realizations for the second series. The set of models chosen is composed of autobinomial Gibbs random fields (three first orders) and Gauss–Markov random fields (three first orders). Then, we applied the method to compute the rate–distortion curve using the algorithm of Section IV-C with varying β, and we made it vary exponentially. By finding the first local maximum of the second derivative of the rate–distortion curve (Fig. 7), we selected the optimal β linked to the optimal number of clusters. From the soft clustering obtained using the Information Bottleneck principle, we derive a hard clustering using (63). Thus, parallelepipeds are characterized spatially and temporally by their belonging to a cluster. The clustering results are displayed for the two sequences in Figs. 8 and 9.

[1]The ADAM data set is in free access at http://medias.obs-mip.fr/adam/. The set is composed of 57 images. Images SPOT: copyright CNES, 2000–2003.

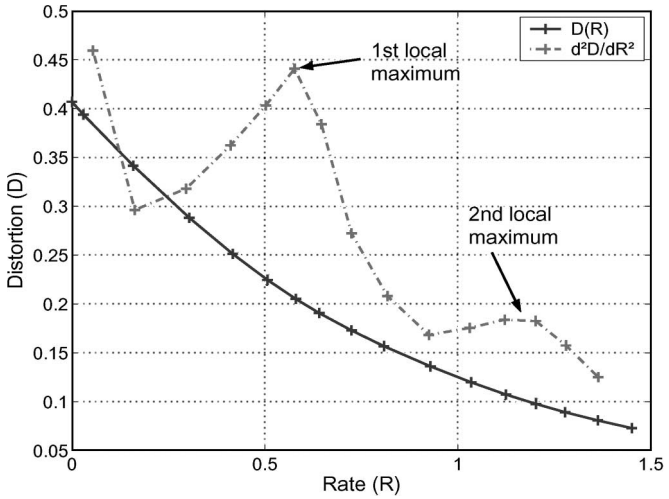


Fig. 7. Rate–distortion curve $D(R)$ and the second derivative $(\partial^2 D)/(\partial R^2)$ obtained. This curve has been obtained in the rate–distortion analysis of the stochastic process modeling the second subsequence shown in Fig. 9. The two local maxima underline the existence of a subclusters structure.

C. Results and Discussion

As shown in Figs. 8 and 9, the algorithm succeeds in characterizing spatio-temporal structures in three or four classes. First, the stable frontiers have been detected where a frontier is a border between two different regions. Second, stable linear structures also constitute a class, which could be roads or talus delineating two similar areas. Third, a class contains the punctual changes. It reveals the existence of small object dynamics. Finally, slow temporal variation areas are gathered. In conclusion, we are able to extract relevant information contained in SITS regarding spatio-temporal dependencies. This paper shows in this case that four classes are meaningful in the SITS using Gibbs–Markov random fields models. In addition, the advantages of our clustering method can be summarized in three remarks. First, the calculus of clustering and the selection of model are done jointly, unlike hierarchical Bayesian methods. Indeed, the selection of the model is embedded in the method by using the model evidence. Second, from the rate–distortion analysis, the optimal number of clusters is selected automatically. Using this criterion, we try to find the natural number of clusters, and at the same time, we try to extract the significant part of the information contained in SITS. Finally, more than a clustering technique, our method gives a way to quantify and to qualify the information contained in data. On the one hand, the information is quantified by the compression or the clustering. On the other hand, the information is qualified by the third variable containing relevant information. Indeed, we assumed that the relevant information is contained in the spatio-temporal dependencies. However, spectral or geometrical information was not taken into account in this paper. Our method can be generalized by considering spectral and geometrical models. Thus, the problem can be viewed as a multi-information-bottleneck problem [27].

Nevertheless, one of the drawbacks of the method is the computation of the model evidence. The computation of log-evidence produces large nonpositive numbers, and the exponential of these numbers is not always representable on computers. However, this tricky step has been solved by the normalization described in Section V. Another drawback of the method is the computational cost. Indeed, 9 and 36 h, respectively, are required to compute the results shown in Figs. 8 and 9. The computational cost of the algorithm is proportional to the number of realizations, which is of 700 in the first case and of 2800 in the second. The high computational cost is due to reiterating the algorithm described in Section IV-C for varying β in order to compute the rate–distortion curve. To speed up the algorithm, one can compute the rate–distorsion curve with less samples of β, thus loosing precision in the estimation of the optimal number of clusters. Another way to speed up the algorithm is by relaxing the criterion of convergence of the information-bottleneck algorithm, thus loosing precision of the clustering.

VII. Conclusion

We have presented a novel method to cluster spatio-temporal parameters of a random field for modeling a SITS. This informational method enables to characterize spatio-temporal structures based on the GMRF and autobinomial Gibbs random fields models. As these models highlight spatio-temporal dependencies in SITS, the information-bottleneck method extracts this type of information. Then, by coupling the information-bottleneck principle to a rate–distortion analysis, the method leads to find the natural number of classes contained in SITS by determining the critical number of clusters. Finally, our experiments show that the method is suited to cluster Gibbs–Markov random fields. Thus, highly informative structures have been extracted from SITS. Future work will be done in the multi-information-bottleneck framework to take into account spectral and geometrical information. Then, from the characterization results, other work will be done in order to provide a short length index for spatio-temporal structure retrieval in large SITS databases.

Appendix I
LSE of the GMRF

Equation (24) is expressed vectorially by (27), in which the matrix G is introduced. Here, we give the formulation of G. For each pixel s and each parameter index r, G is expressed as

$$[G]_{s,r} = X_{s+r} + X_{s-r}. \tag{66}$$

For simple notation, we make the confusion between the index of parameter and r that represents a displacement in the neighborhood N on the field.

Appendix II
CLSE of the Autobinomial Model

The mean and variance of X_s are expressed as

$$E[X_s] = \frac{\mathcal{G}}{1+e^{-\eta}} \tag{67}$$

$$E\left[(X_s - E[X_s])^2\right] = \frac{\mathcal{G}e^{-\eta}}{(1+e^{-\eta})^2}. \tag{68}$$

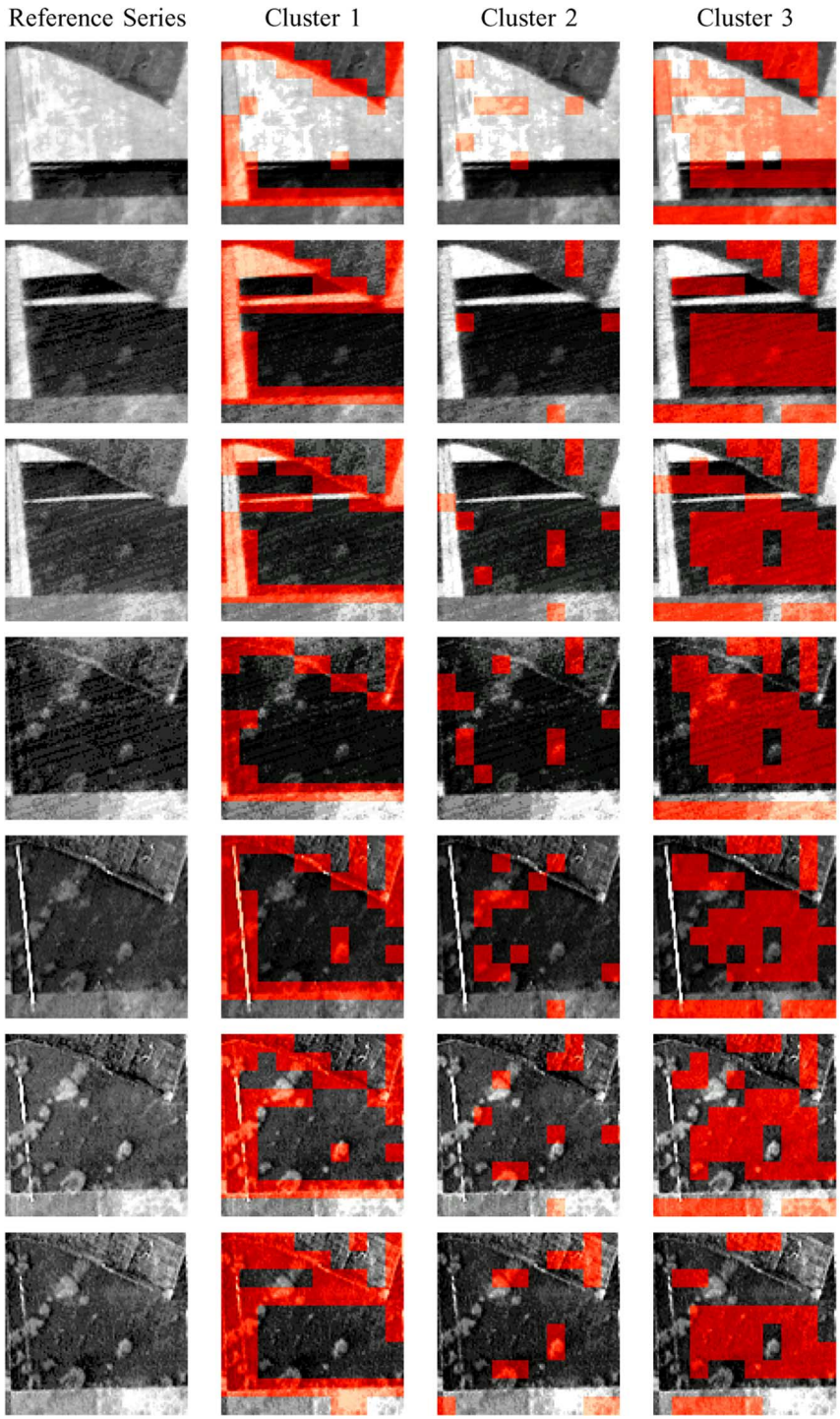


Fig. 8. Sequence represents the harvest of a field. The clustering results are presented on series of size 100 × 100 × 7, and the optimal number of clusters is three. In the first cluster, the stable frontiers and the linear structures have been gathered. In the second cluster, we observe punctual changes. This cluster represents the object dynamics. We can see that humidity patches have been detected. Finally, the third cluster contains slow background changes, which represent stable vegetation. In addition, this cluster has detected the linear artifacts caused by the instruments.

The CLSE is defined as

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{s} (x_s - E[X_s])^2 \tag{69}$$

$$= \arg\min_{\Theta} \sum_{s} \left(x_s - \frac{\mathcal{G}}{1 + e^{-\eta}} \right)^2. \tag{70}$$

We can perform a Taylor expansion and obtain

$$\eta \approx -\log\left(\frac{\mathcal{G}}{x_s} - 1\right) + e_s \tag{71}$$

where e_s is a Gaussian noise of zero mean and small variance. By replacing η in (71) by its exact formulation (32), we obtain

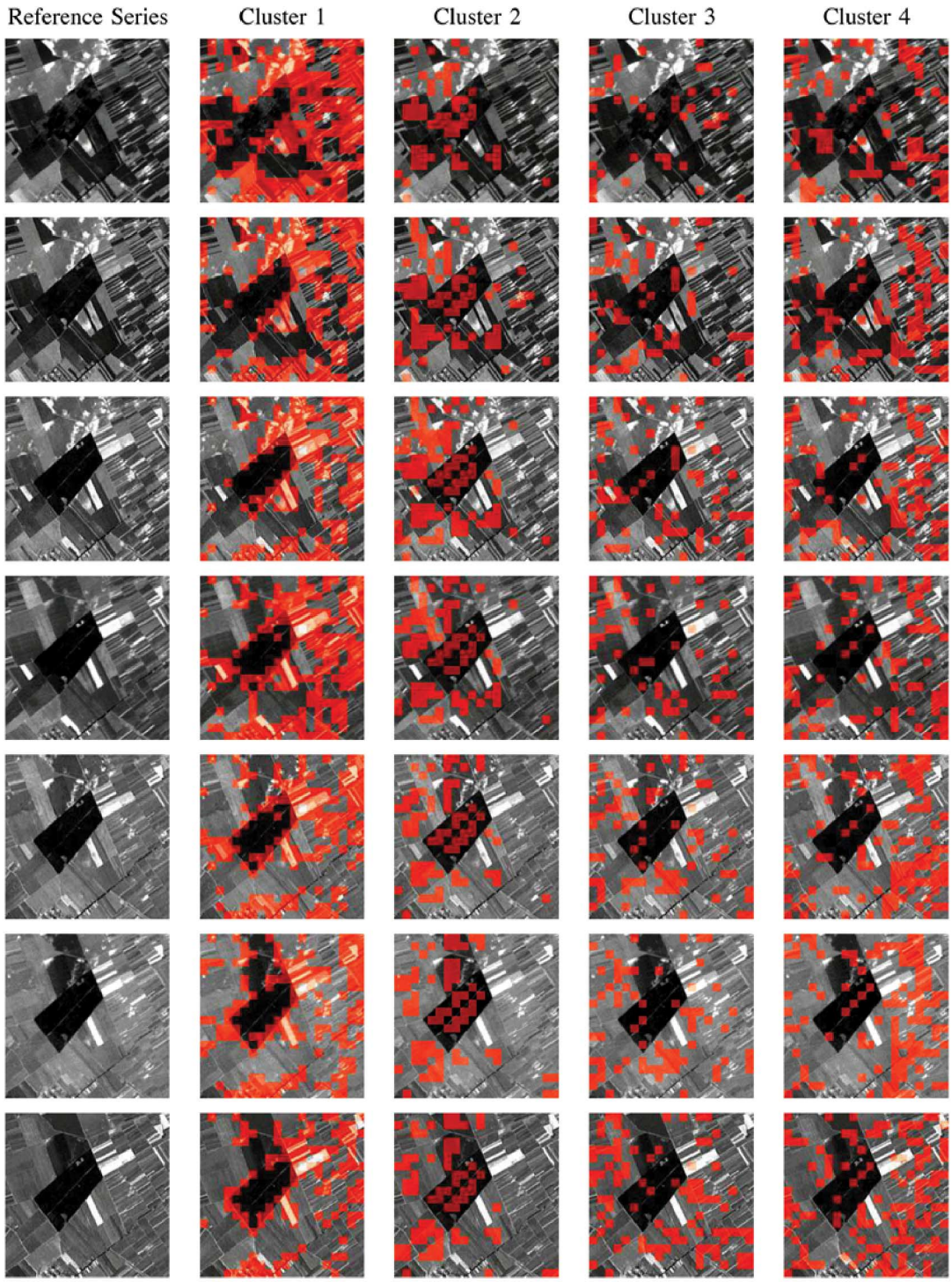


Fig. 9. Clustering results are presented on series of size 200 × 200 × 7, and the optimal number of clusters is four. In the first cluster, we observe the stable frontiers between two different types of vegetation. In the second cluster, the slow temporal variation regions have been gathered. The third cluster represents the punctual and small-object dynamics. Finally, the fourth cluster contains the stable linear structures, such as roads or talus between two similar types of vegetation. In addition, thin fields belong to this cluster, as shown on the right part of images.

the following linear system:

$$G\Theta = d + E \tag{72}$$

with the vector of unknown parameters Θ, the vector d of transformed pixel values obtained from the right terms of (71), a Gaussian noise vector E, and the matrix G of neighboring pixel values.

$$[d]_s = -\log\left(\frac{\mathcal{G}}{x_s} - 1\right) \tag{73}$$

$$[G]_{s,r} = \begin{cases} 1, & \text{if } r = 0 \\ \frac{X_{s+r}+X_{s-r}}{\mathcal{G}}, & \text{if } r \in N. \end{cases} \tag{74}$$

Then, the minimum mean squared error estimate of Θ is

$$\hat{\Theta} = (G^{\mathrm{T}}G)^{-1}G^{\mathrm{T}}d. \tag{75}$$

Finally, $p(X|\mathcal{M})$ is expressed with the (30).

ACKNOWLEDGMENT

The authors would like to thank the role of K. Raney (Johns Hopkins University) for contributing to the clarity in the presented concepts, and of K. Seidel (ETH Zurich) for the discussions about information mining. They would also like to thank A. Giros (CNES) for initiating the frame of work on SITS and for numerous stimulative discussions.

REFERENCES

[1] C. M. Antunes and A. L. Oliveira, "Temporal data mining: An overview," in *Proc. Workshop Temporal Data Mining*, 2001, pp. 1–13.

[2] P. Heas, P. Marthon, M. Datcu, and A. Giros, "Image time-series mining," in *Proc. IGARSS*, Anchorage, AK, Sep. 2004, vol. 4, pp. 2420–2423.

[3] P. Heas and M. Datcu, "Modelling trajectory of dynamic cluster in image-time-series for spatio-temporal reasoning," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1635–1647, Jul. 2005.

[4] M. Datcu and K. Seidel, "Image information mining: Exploration of image content in large archives," in *Proc. IEEE Aerosp. Conf.*, Mar. 2000, vol. 3, pp. 253–264. 18–25.

[5] M. Datcu, K. Seidel, S. D'Elia, and P. G. Marchetti, "Knowledge-driven information mining in remote-sensing image archives," *ESA Bull.*, no. 110, pp. 26–33, May 2002.

[6] M. Datcu, H. Daschiel *et al.* "Information mining in remote sensing image archives: System description," *IEEE Trans. Geosci. Remote Sens.,* vol. 41, no.12, pp. 2923–2936, Dec. 2003.

[7] F. C. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in *Proc. 30th Annu. Meeting Assoc. Comput. Linguistics*, 1993, pp. 190–193.

[8] J. Goldberger, H. Greenspan, and S. Gordon, "Unsupervised image clustering using the information bottleneck method," in *Proc. 24th DAGM Symp., Pattern Recog.*, Zurich, Switzerland, Sep. 2002, pp. 158–165.

[9] M. Datcu, K. Seidel, and M. Walessa, "Spatial information retrieval from remote-sensing images. I. Information theoretical perspectives," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 5, pp. 1431–1445, Sep. 1998.

[10] M. Schroder, H. Rehrauer, K. Seidel, and M. Datcu, "Spatial information retrieval from remote-sensing images. II. Gibbs–Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 5, pp. 1446–1455, Sep. 1998.

[11] J. J. Rissanen, "A universal data compression system," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 5, pp. 656–664, Sep. 1983.

[12] C. Sugar and G. James, "Finding the number of clusters in a dataset: An information theoretic approach," *J. Amer. Stat. Assoc.*, vol. 98, no. 463, pp. 750–763, 1998.

[13] J. Hartigan and M. Wong, "A k-means clustering algorithm," *Appl. Stat.*, vol. 1, no. 28, pp. 100–108, 1979.

[14] R. Chellappa and R. I. Kashyap, "Texture synthesis using 2-D noncausal autoregressive models," *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-33, no. 1, pp. 194–204, Feb. 1985.

[15] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[16] F. Spitzer, "Markov random field and Gibbs ensembles," *Amer. Math. Mon.*, vol. 78, no. 2, pp. 142–154, Feb. 1971.

[17] J. J. K. O Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York: Springer-Verlag, 1996, ch. 2.

[18] G. L. Bretthorst, "Bayesian analysis. II. Signal detection and model selection," *J. Magn. Reson.*, vol. 88, no. 3, pp. 552–570, Jul. 1990.

[19] G. R. Cross and A. K. Jain, "Markov random field texture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 1, pp. 25–39, Jan. 1983.

[20] S. R. Lele and J. K. Ord, "Conditional least squares estimation for spatial processes: Some asymptotics results," Dept. Statist., Pennsylvania State Univ., University Park, PA, Tech. Rep. 65, 1986.

[21] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proc 37th Annu. Allerton Conf. Commun., Control and Comput.*, 1999, pp. 368–377.

[22] A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu, "An information theoretic analysis of maximum likelihood mixture estimation for exponential families," in *Proc. ACM 21st Int. Conf. Mach. Learn.*, Jul. 2004, vol. 8, p. 8.

[23] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 234–245.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

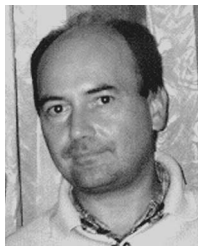
[25] K. Rose, "Deterministic annealing for clustering, compression, classification, regression and related optimization problems," *Proc. IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov. 1998.

[26] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, "Autoclass: A Bayesian classification system," in *Proc. 5th Int. Conf. Mach. Learn.*, Jun. 1988, pp. 54–64.

[27] N. Friedman, O. Mozenzon, N. Slonim, and N. Tishby, "Multivariate information bottleneck," in *Proc. UAI*, 2001, pp. 152–161.

Lionel Gueguen received the engineering degree in telecommunications from the Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, and the M.S. degree in signal and image processing from the Université de Rennes 1, Rennes, France, both in 2004.

He is currently working toward the Ph.D. degree at the École Nationale Supérieure des Télécommunications (ENST), Paris, France. He has been funded by the French Aerospace Center (CNES), Toulouse, France, and the German Aerospace Center, Wessling, Germany, since 2005. His research interests include satellite image time-series mining and compression.



Mihai Datcu (SM'04) received the M.S. and Ph.D. degrees in electronics and telecommunications from the University "Politechnica" Bucharest UPB, Romania, and the "Habilitation à diriger des recherches" from Université Louis Pasteur, Strasbourg, France, in 1978, 1986, and 1999, respectively.

He holds a Professorship in electronics and telecommunications with the University "Politechnica" Bucharest UPB, Romania, since 1981. Since 1993, he has been a Scientist with the German Aerospace Center (DLR), Oberpfaffenhofen, Wessling, Germany. He is developing algorithms for model-based information retrieval from high-complexity signals and methods for scene understanding from SAR and interferometric SAR data. He is engaged in research related to information theoretical aspects and semantic representations in advanced communication systems. Currently, he is Senior Scientist and Image Analysis Research Group Leader with the Remote Sensing Technology Institute of DLR, Oberpfaffenhofen, the Coordinator of the CNES/DLR/ENST Competence Centre on Information Extraction and Image Understanding for Earth Observation, and Professor with the École Nationale Supérieure des Télécommunications Paris. His interests are in Bayesian inference, information and complexity theory, stochastic processes, model-based scene understanding image semantic coding, image information mining for applications in information retrieval and understanding of high-resolution SAR, and optical observations. He has held Visiting Professor appointments from 1991 to 1992 with the Department of Mathematics of the University of Oviedo, Spain, and from 2000 to 2002, with the Université Louis Pasteur and the International Space University, both in Strasbourg, France. From 1992 to 2002, he had a longer Invited Professor assignment with the Swiss Federal Institute of Technology ETH, Zürich. In 1994, he was a Guest Scientist with the Swiss Center for Scientific Computing, Manno, and in 2003, he was a Visiting Professor with the University of Siegen, Germany. He is involved in advanced research programs for information extraction, data mining, and knowledge discovery and data understanding with the ESA, CNES, NASA, and in a variety of European projects.

Prof. Datcu is member of the European Image Information Mining Coordination Group.