Analysis of Satellite Image Time Series Based on Information Bottleneck

Lionel Gueguen[*,†], Camille Le Men[*,†] and Mihai Datcu[**,*]

[*]*GET-Télécom Paris - 46 rue Barrault, 75013 Paris, France*
[†]*CNES - 18 avenue Edouard Belin, 31401 Toulouse, France*
[**]*German Aerospace Center DLR - Oberpfaffenhofen, D-82234 Wessling, Germany*

Abstract. Derived from Information Theory, the Information Bottleneck principle enables to quantify and qualify the information contained in a signal. This paper presents an algorithm based on the Information Bottleneck principle to analyze Satellite Image Time Series (SITS). The method is composed of a parameter estimation and a model selection. This method has been previously applied to textural and radiometric parametric models and we propose here to extend it to take into account the geometry information. Two approaches are presented. In the first approach, each image of the SITS is segmented and the obtained regions are described by textural models. The Information Bottleneck method is further used to characterize the image segments of the SITS a spatio-temporal way. In the second method, the geometrical information is extracted from a temporal adjacency graph of the spatial regions, and the radiometric and textural information is then extracted through the Information Bottleneck method. This approach leads to a temporal characterization of the spatial regions of the SITS.

Keywords: Unsupervised clustering, Satellite Image Time Series, Information Bottleneck, segmentation.
PACS: 89.20.Ff

INTRODUCTION

Nowadays, huge quantities of satellite images are available due to the growing number of satellite sensors. Moreover, a scene can be observed very often, thus enabling to create Satellite Image Time-Series (SITS). These SITS contain highly detailed spatial information and some information about the scene dynamic. They are therefore highly complex data containing numerous and various spatio-temporal information. For example in a SITS, growth, maturation or harvest of cultures 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 made such as change detection, monitoring or validation of physical models. However, these techniques are dedicated to specific restricted applications. Consequently in order to exploit the information contained in SITS, more general analyzing methods are required. Some methods for low resolution images and uniform sampled have been studied in [1]. For high resolution and non-uniform time-sampled SITS, new spatio-temporal analyzing algorithms are presented in [2, 3]. They are based on a Bayesian hierarchical model of information content. The concept was first introduced in [4, 5] 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 adapted to the user's query. Alternatively, this paper addresses the problem of representing

objectively the information by unsupervised clustering and model selection. In order to cluster spatio-temporal events of SITS, we present a new method which clusters stochastic processes without any supervision since the optimal number of clusters is computed. This method is based on the Information Bottleneck (IB) principle which is an extension of the Rate-Distortion analysis. The paper is organized as follows. First, we present the theoretical concept for modelling stochastic process. Then, we present the models which describe the spatio-temporal informational content of SITS. Finally, we give two experiments of the method on a SITS.

MODELLING STOCHASTIC PROCESS : THE CONCEPT

In the following sections, we give a theoretical substantiation for using the IB principle to model a stochastic process. First, we remind the Minimax problem of Redundancy-Coding introduced by Davisson [6] and exploited by Rissanen [7] to derive the Minimum Description Length criterion. Then, we point out the link between the IB principle and the Minimax problem. Finally, we highlight the pertinence of the IB principle for modelling stochastic processes.

Redundancy coding

We are interested in finding a universal generative model of a stochastic process X[1]. As there are an infinity of models, we restrict the research to a finite set of generative model $\{m_i\}$ which define the conditional probabilities $\{p(X|m_i)\}$. Thus, we look for a distribution $q(X)$ whose ideal code length is the shortest in mean for the worst model generating the stochastic process. Then, we define the redundancy as the additional amount of information required to encode the realizations of X using a distribution $q(X)$ instead of $p(X)$. The redundancy $R(p,q)$ is given by the Kullback-Leibler divergence between the two distributions:

$$R(p,q) = D_{KL}(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

Considering a set of generative models, the Minimax problem of Redundancy-Coding consists in finding $q(X)$ minimizing the largest redundancy which is the one associated with the worst model:

$$R^+ = \min_{q(X)} \max_{\mathscr{M} \in \{m_i\}} R(p(X|\mathscr{M}), q(X)) \tag{2}$$

Rissanen embeds it within a wider problem by considering the mean redundancy instead of the worst case redundancy [8]. He introduces a probability distribution on the set of generative models, $w(\mathscr{M})$ and the redundancy is:

$$R(w,q) = \sum_{\mathscr{N} \in \{m_i\}} w(\mathscr{N}) R(p(X|\mathscr{N}), q(X)) = I(X, \mathscr{M}) \tag{3}$$

It follows that the Minimax problem is reformulated as:

$$R^+ = \min_{q(X)} \max_{w(\mathscr{M})} R(w(\mathscr{M}), q(X)) \tag{4}$$

Solving this problem, we find first that $q(X)$ is a mixture of the generative models for any distribution $w(\mathscr{M})$. The following formula is also obtained using the Bayes rule:$q^*(x) =$

[1] Upper case letters are used to name random variables and lower case letters are used to name the realizations.

$\sum_{\mathscr{N}\in\{m_i\}} w(\mathscr{N})p(x|\mathscr{N})$. After minimizing over the distribution $q(X)$, we want to maximize the mutual information $I(X,M)$ over the distribution $w(\mathscr{M})$. The maximum obtained is the capacity of the channel $\mathscr{M} \to X$. In conclusion, solving the Minimax Redundancy-Coding problem corresponds to find a universal model for a stochastic process that could be generated by any models of the set.

The Information Bottleneck principle

In this section, we make the link between the IB principle and the Minimax problem of Redundancy-Coding. Tishby introduced the IB principle in [9]. We want to compress a signal while preserving the relevant information contained in another random variable. Let X be the signal, $\tilde{X}$ be the compressor and Y be the variable containing the relevant information, thus the IB principle is expressed as the following minimization.

$$\min_{p(\tilde{x}|x),p(y|\tilde{x})} I(\tilde{X},X) - \beta I(\tilde{X},Y) \tag{5}$$

Where β is a trade off parameter between the compression and the relevant information extracted. To make the link with redundancy coding, we consider that the relevant information is contained in a random model $\mathscr{M}$ taking its values in a set of models $\{m_i\}$. Then, the previous principle is expressed as:

$$\min_{p(\tilde{x}|x),p(\mathscr{N}|\tilde{x})} I(\tilde{X},X) - \beta I(\tilde{X},\mathscr{M}) \tag{6}$$

In this context, we have access to the real distributions $p(X)$ and $p(\mathscr{M}|X)$. Hence, while we minimize the whole criterion, the mutual information $I(\tilde{X},\mathscr{M})$ is maximized over the distribution $p(\mathscr{N} \mid \tilde{x})$. We suppose that $p(\tilde{X})$ is fixed during the maximization. Moreover, we suppose that $p(\tilde{X}) = \sum_{\mathscr{N}\in\{m_i\}} w(\mathscr{N})p(\tilde{X}|\mathscr{N})$, using the Bayes rule. In consequence, it is equivalent to maximize over the distribution $w(\mathscr{N}) = \sum_{\tilde{x}} p(\mathscr{N} \mid \tilde{x})p(\tilde{x})$. Relating to the Minimax Redundancy-Coding problem, we try to find the distribution of models which causes the worst redundancy with $p(\tilde{X})$. However, $p(\tilde{X})$ is not fixed and is related to the minimization of $I(\tilde{X},X)$ over the distribution $p(\tilde{x} \mid x)$. In the Minimax problem, we look for a distribution $q(X)$ whose ideal code length is the shortest in mean for the worst model generating the data. In the IB case, on one hand, we try to find a distribution $p(\tilde{X})$ whose ideal code length is the shortest by minimizing $I(\tilde{X},X)$. On the other hand, we try to find the distribution $w(\mathscr{N})$ which gives the largest redundancy with the distribution $p(\tilde{X})$. In a sense, we try to extract from the probability $p(X)$ a new distribution $p(\tilde{X})$ which would be a universal model for any stochastic process generated by any of the models. However, unlike the Minimax problem, there is a trade off in the IB principle. For example, as β tends to zero, $I(\tilde{X},\mathscr{M})$ is not maximized, resulting into a non-universal model for the set. On the contrary, when β tends to ∞, $p(\tilde{X})$ approaches a universal model. In fact the previous principle enables to find the subpart of information of X which is well modeled in mean by the set of generative models in the sense of redundancy coding.

Classification of stochastic process

In this section, we highlight the use of the Information Bottleneck principle for classifying stochastic processes. From the Minimax Redundancy-Coding problem, Rissanen derived the Minimum Description Length (MDL) principle [7] in order to optimally model a stochastic process. Generally, the MDL principle enables to select a model among a family of parametric models in order to characterize stochastic processes by their corresponding estimated parameters. Then, a classification of the processes is possible through a clustering of the estimated

parameters [10]. Conversely, in our approach we try to find a subpart of information from the stochastic process which is optimally modelled by a set of models while the processes are clustered at the same time. In this last case, we extract the meaningful information related to the models and used for clustering. Consequently, our method jointly clusters and selects models. Actually, the model selection is embedded in the clustering computation. However, the choice of the trade-off parameter β remains critical. This aspect will be discussed in a following section. If we consider a stochastic process X, we can always compute a non-parametric estimation of its distribution. Then, using the IB principle we try to extract the relevant information related to the set of models thus finding the conditional probability $p(\tilde{x} \mid x)$. We interpret this probability like a soft clustering. In a sense, we cluster the realizations of X using the information related to the set of models used.

Optimal trade-off parameter

As previously mentioned, the selection of the trade-off parameter β is critical. We give an heuristic criterion based on a Rate-Distortion analysis. Tishby describes the self-consistent equations of the minimization problem of the equation (5) in [9]. As the equations are intricate, an Expectation-Maximization like algorithm has been derived for the IB problem in [9, 11]. In addition, we can define a Rate-Distortion function $(R(\beta), D(\beta))$ as explained by Banerjee in [11]. In fact, β influences the effective number of clusters found. 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 curvature of $D(R)$ (7). Consequently, more than selecting the trade-off parameter, we give a criterion to estimate the optimal number of classes during the clustering process. The reason for selecting such a criterion is described in [12]. In fact, we try to find the point where the decrease in distortion does not provide significantly more information.

$$\hat{\beta} = \arg\sup_{\beta} \frac{|R'(\beta)D''(\beta) - R''(\beta)D'(\beta)|}{(R'(\beta)^2 + D'(\beta)^2)^{\frac{3}{2}}} \tag{7}$$

INFORMATIONAL CHARACTERIZATION OF SPATIO-TEMPORAL PATTERNS

Gauss-Markov Random Field

Gauss-Markov Random Field (GMRF) have shown interesting properties for characterizing textures in satellite images [10, 13]. In the following sections, we present a family of GMRF for the analysis of SITS since these latter preserve spatial and temporal dependencies. Using GMRF, these dependencies will be characterized. These parametric models are then used to represent the relevant information extracted in the IB framework. In the first method (described in) GMRF are generalized to a 3-dimensional random field: We consider that the random variable X is a random field defined on a grid. Let X_s be the observations, s belonging to a 3-dimensional lattice Ω and N the half of a symmetric 3-dimensional neighborhood (Figure 1(a)). So GMRF are defined as follows:

$$X_s = \sum_{r \in N} \theta_r (X_{s+r} + X_{s-r}) + e_s \tag{8}$$

where e_s is a white Gaussian noise of variance σ_e and θ_r is a scalar parameter associated to each direction in the neighborhood. The equation (8) is expressed vectorially as follows ordering X

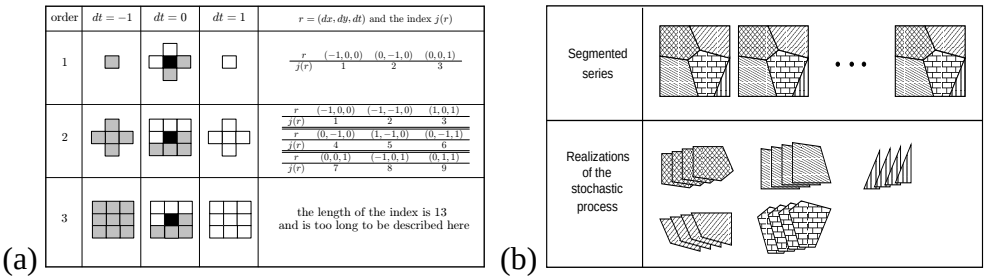

FIGURE 1. (a)Symmetric 3-d neighborhood. The pixel X_s is black. Pixels corresponding to X_{s+r} are white and pixels corresponding to X_{s-r} are grey. dt is the time dimension and (dx,dy) is the spatial dimension.(b)Definition of the stochastic process derived from the global spatial segmentation. At each time, the regions stay the same. So we define a realization of the stochastic process as the concatenation of the same region in time following images.

in a matrix G: $X = G\Theta + E$. From this equation, $\hat{\Theta}$ is estimated with the Least Minimum Square method. From the estimated parameters, it is possible to approximate the model evidence (9). A general formulation of the model evidence for linear systems is given in [14, 15]. Considering N, Q being the respective dimension of X, Θ, the model conditional likelihood is given by:

$$p(X|\mathcal{M}) \approx \frac{\pi^{-N/2}\Gamma(\frac{Q}{2})\Gamma(\frac{P-Q}{2})|G^TG|^{-1/2}}{4R_\delta R_\sigma(\hat{\Theta}^T\hat{\Theta})^{Q/2}(X^TX-(G\hat{\Theta})^T(G\hat{\Theta}))^{\frac{P-Q}{2}}} \tag{9}$$

Where R_δ, R_σ are some constants not calculable. As discussed in the section , we need to define a set of models in order to apply the IB principle for stochastic process classification. We propose to use a set of GMRF models of different orders. The Figure 1(a) showing a 3-dimensional GMRF of order 1, 2 and 3 illustrates the fact that the order of the model is linked to the number of parameters required.

CLUSTERING OF SPATIO-TEMPORAL STRUCTURES IN SATELLITE IMAGE TIME SERIES

We present two experiments for clustering spatio-temporal structures in SITS. Both experiments use the IB based clustering method. First, we use a spatial reference segmentation (RS) to define the stochastic process to be modeled. This assumption leads to a spatio-temporal characterization of the regions contained in the SITS. Secondly, we consider temporal chains of spatial regions which are characterized by textural and radiometric parameters. Then, we consider the corresponding parameters chains as a stochastic process. Thus, we obtain a temporal modelling of the regions evolution.

Spatio-temporal modelling

First of all, we assume that a reference spatial segmentation (RS) of the SITS exists. This assumption is motivated by the stability of the spatial regions over the time. The RS of the series is obtained considering the SITS as a vectorial image, the components of the vector being the different time samples. The Euclidean norm of its vectorial gradient is then computed and used to initialize a segmentation algorithm. The RS is used as a mask on successive images thus defining the realizations of a stochastic process as shown in the Figure 1(b). However the spatial regions have different size, therefore we need to define a normalized stochastic

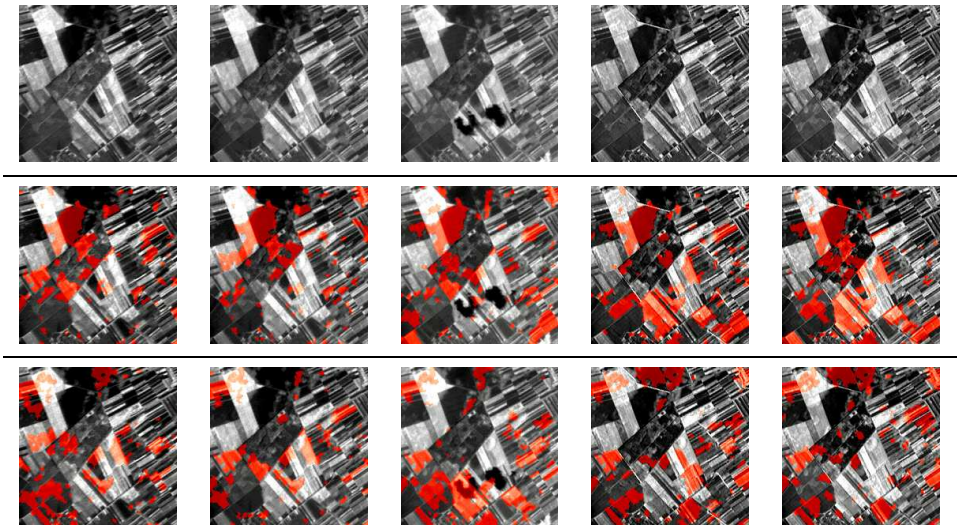

FIGURE 2. One the top the reference series is displayed.Below, two clusters are displayed in red in the spatio-temporal domain. The first cluster highlights flat regions which do not vary in time. Finally, the second cluster shows an intermediate evolution where only intensities vary in time which corresponds to humidity of the ground.

process. For example, if we consider a GMRF X defined on a 3-dimensional field of size n where pixels are indexed form 1 to n, the probability of X is approximated by the pseudo likelihood $\prod_{i=1}^{n} p(X_i \mid \{X_{i+r}, X_{i-r}, \theta_r, r \in N\})$. We consider a normalized stochastic process X_s which follows the distribution $p(X_s) = p(X)^{\frac{1}{n}}$. We normalize in the same way the evidence of the model, which means that we compute $p(X \mid \mathscr{M})^{\frac{1}{n}}$ where n is the size of the lattice Ω. Consequently, the resulting evidence does not depend on the size of the regions considered. We made the experiment on series of 20 images 200×200. The RS having approximatively 1000 regions, we obtain 20×1000 realizations of the stochastic process composed of 5 regions following in time (there is an overlapping of the realizations in time). In order to compute the Rate-Distortion curve, we start by applying the IB algorithm with a very small β and with a random initialization. The result is a unique cluster and does not depend on the initialization. Then, we apply several times the algorithm with exponentially growing trade-off parameters. This is a simulated annealing like scheme as described for Expectation-Maximization algorithms in [16]. We obtain 15 clusters after the Rate-Distortion analysis. We show in the Figure 2 the results of two clusters drawn in the spatio-temporal domain.

Temporal modelling

The hypothesis of a common segmentation for all images is actually not completely verified. Therefore, we propose to take into account the structural changes before the temporal features characterization. Given the nature of the scene observed, we make the assumption that the objects do not move but that their structure and geometry may change. Consequently, the segments of two images can be linked according to an intersection criterion. These structural changes are represented in a spatio-temporal graph of the spatial regions. The nodes of the graph correspond to the segments and a node is linked to nodes of the previous and next images if their intersection is not empty. In order to be able to link two segments, the two following

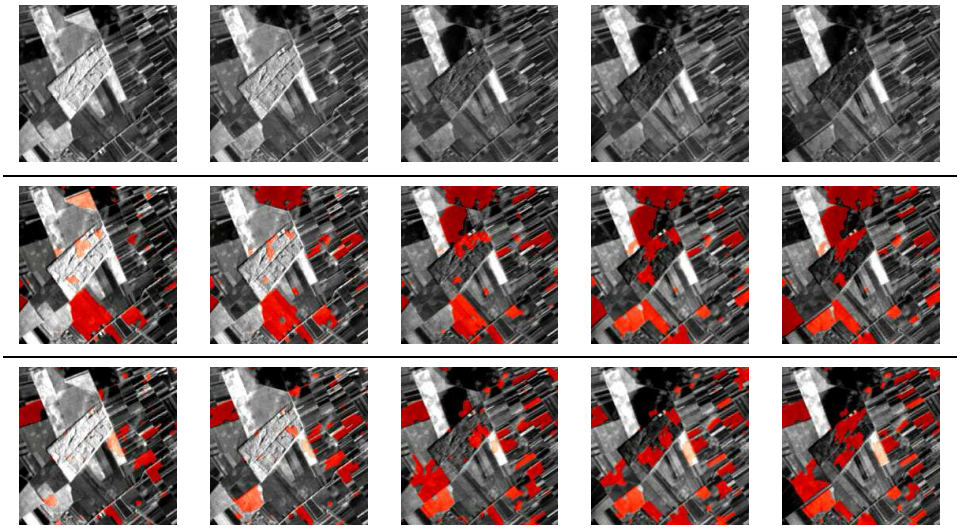

FIGURE 3. One the top the reference series is displayed.Below, two clusters are displayed. The first cluster highlights strong temporal changes such as harvest. The second cluster shows some stable regions in time with few spectral variations.

segmentations have to be comparable, ie most structurally similar. In order to obtain those series of segmentations, we use the structural criteria of segmentation comparison described in [17]. We define a dissimilarity measure inspired from the coding theory: a distance between two partitions $\mathscr{P}_1$ and $\mathscr{P}_2$ is defined as the sum of the lengths of the two messages:*describing* $\mathscr{P}_1$ *knowing* $\mathscr{P}_2$ and *describing* $\mathscr{P}_2$ *knowing* $\mathscr{P}_1$. We look for a maximum of these messages lengths (l^*) which is the conditional entropy. We then obtain: $\mathscr{D}(\mathscr{P}_2,\mathscr{P}_1) = l^*(\mathscr{P}_2|\mathscr{P}_1) + l^*(\mathscr{P}_1|\mathscr{P}_2) = -\sum_{k=1}^{\overline{\mathscr{P}_1}} \sum_{l=1}^{\overline{\mathscr{P}_2}} \frac{\overline{\mathscr{Q}_l \cap \mathscr{R}_k}}{N} \log_2 \frac{\overline{\mathscr{Q}_l \cap \mathscr{R}_k}}{\overline{\mathscr{R}_k}\,\overline{\mathscr{Q}_l}}$ where $\overline{E}$ represents the number of elements in E, $\mathscr{R}_k$ (respectively $\mathscr{Q}_l$) is a region of $\mathscr{P}_1$ (respectively $\mathscr{P}_2$), and N is the number of pixels in an image. This distance enables to select the closest segmentation to a RS among a set of scale varying segmentations. In order to obtain series of segmentations of the SITS, we use the iterative method described in [18]. The result is a series of segmentations which are the most similar (in the sense defined above) to a RS under the constraint that each segmentation is also the most similar to the segmentations of the previous and the following images. We then use the IB method in order to characterize the evolution of the region's mean intensities in time. We consider that the chains of spectral means of regions form the realizations of the stochastic process. In addition, we use Markov models of order one, two and three for representing the chains. Consequently, we characterize the evolution of the spectral means in time, thus obtaining a temporal analysis of the spatial regions of the temporal adjacency graph. We made the experiment on series of size $200 \times 200 \times 14$ which contain approximately 1000 chains of spectral means parameters. We then apply the previously described IB algorithm and find 26 clusters at the trade-off parameter optimum. The Figure 3 shows the results of two clusters drawn in the signal domain.

CONCLUSION

We have presented a general method to compute a soft clustering of stochastic processes. This method is unsupervised and enables to find the natural number of clusters in a set of the realizations of a stochastic process. In addition, we have shown that the method embeds a model selection during the clustering process as done in the Bayesian framework. We also presented

how to apply the method on SITS using Gauss-Markov Random Fields and segmentation methods. In the first experiment, we clustered spatio-temporal structures and in the second, we clustered the evolution of the structures. From the results, we can conclude that the method achieves to retrieve meaningful information from SITS in both cases and it is a suitable tool for creating indexes of SITS. However, the method is complex and requires a lot of computation power.

REFERENCES

1. C.M. Antunes and A.L. Oliveira, "Temporal Data Mining: an Overview," Workshop on temporal data mining, IST, Lisbon Technical University, 2001.
2. P. Heas, P. Marthon, M. Datcu, and A. Giros, "Image time-series mining," in *IGARSS'04*, Anchorage, USA, Sept. 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 Transactions on Geoscience and Remote Sensing*, vol. 43, no. 7, pp. 1635–1647, 2005.
4. M. Datcu and K. Seidel, "Image Information Mining: Exploration of Image Content in Large Archives," in *IEEE Aerospace Conference Proceedings*, March 2000, vol. 3 of *18-25*, pp. 253–264.
5. M. Datcu, H. Daschiel, and al., "Information Mining in Remote Sensing Image Archives: System Description," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2923–2936, Dec. 2003.
6. L. D. Davisson, "Universal Noiseless Coding," *IEEE Transactions on Information Theory*, vol. IT-19, no. 6, pp. 783–795, Nov. 1973.
7. J.J. Rissanen, "A Universal Data Compression System," *IEEE Transactions on Information Theory*, vol. IT-29, no. 5, pp. 656–664, Sept. 1983.
8. J.J. Rissanen, "Lectures on Statictical Modeling Theory," Tech. Rep., Technical Universities of Tampere and Helsinki, Helsinki, Finland, 2002.
9. N. Tishby, F. Pereira, and W. Bialek, "The Infomation Bottleneck Method," in *Proc 37th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.
10. M. Schroder, H. Rehrauer, K. Seidel, and M. Datcu, "Spatial Information Retrieval from Remote-Sensing Images. ii. Gibbs-Markov Random Fileds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 5, pp. 1446–1455, Sept. 1998.
11. A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu, "An information theoretic analysis of maximum likelihood mixture estimation for exponential families," in *ACM Twenty-first international conference on Machine learning*, Alberta, Canada, July 2004, vol. 8, ACM Press.
12. C. Sugar and G. James, "Finding the Number of Clusters in a DataSet: An Information Theoretic Approach," *Journal of the American Statistical Association*, pp. 750–763, 1998.
13. R. Chellappa and R.I. Kashyap, "Texture Synthesis Using 2-D Noncausal Autoregressive Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 1, pp. 194–204, Feb 1985.
14. J.J.K. O Ruanaidh and W.J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, chapter 2, Springer, 1996.
15. G. L. Bretthorst, "Bayesian Analysis. II. Signal Detection and Model Selection," *Journal of Magnetic Resonance*, vol. 88, pp. 552–570, 1990.
16. K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression and Related Optimization Problems," *Proceedings of IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov. 1998.
17. A.Giros, "Comparison of partitions of two images for satellite image time series segmentation," *IGARSS'06 proceedings, Denver (USA)*, august 2006.
18. Y.Lemeur, "Segmentation multitemporelle d'une séquence d'images spot,*MsC thesis*," september 2004.