Ieee Journal Of Selected Topics In Applied Earth Observations And Remote Sensing
January 2015, Volume 8 Issue 1 Pages 204-213
http://dx.doi.org/10.1109/JSTARS.2014.2341219
http://archimer.ifremer.fr/doc/00255/36643/

Hybrid hidden Markov model for marine environment monitoring

Rousseeuw Kevin [1, 2], Poison Caillault Emilie [1], Lefebvre Alain [2], Hamad Denis [1]

[1] ULCO/LISIC, BP 719, FR-62228 Calais, France
[2] IFREMER, Centre Manche Mer du Nord, BP 699, FR-62321 Boulogne-sur-Mer, France

Abstract :

Phytoplankton is an important indicator of water quality assessment. To understand phytoplankton dynamics, many fixed buoys and ferry boxes were implemented, resulting in the generation of substantial data signals. Collected data are used as inputs of an effective monitoring system. The system, based on unsupervised hidden Markov model (HMM), is designed not only to detect phytoplancton blooms but also to understand their dynamics. HMM parameters are usually estimated by an iterative expectation-maximization (EM) approach. We propose to estimate HMM parameters by using spectral clustering algorithm. The monitoring system is assessed based on database signals from MAREL-Carnot station, Boulogne-sur-Mer, France. Experimental results show that the proposed system is efficient to detect environmental states such as phytoplankton productive and nonproductive periods without a priori knowledge. Furthermore, discovered states are consistent with biological interpretation.

Keywords : Hybrid Hidden Markov Model, marine water monitoring, Phytoplankton blooms, spectral clustering

1. Introduction

In the framework of coastal and river water quality assessment and management, phytoplankton plays an important role as an indicator of short- and long-term changes in water quality. Indeed phytoplankton cells are capable of integrating natural and human induced disturbances by changing their physiology. The Marine Strategy Framework Directives (MSFD) underlined the importance to prevent and early detect phytoplankton blooms (harmful, and non-toxic as well), and to understand their physical and outbreak nutrient conditions [1], [2].

Advances in monitoring systems arise from the evolution of computer technology, the availability of effective low-cost sensors, and the deployment of remote sensing generating multidimensional signals. Mathematical models and powerful tools are therefore needed to effectively monitor complex systems with multivariate time series. Recently, machine learning approaches are used to detect harmful algae blooms thanks to available information on cell taxonomy. Such systems are trained from global observation by remote sensing (Support Vector Machine [3], Probabilistic Neural Networks [4]) or local observations like flow cytometry datasets (Radial Basis Function Neural Network [5]).

To monitor phytoplankton dynamics, many marine instrumented stations, fixed buoys and ferry boxes, were implemented with High Frequency (HF) multi-sensor systems. Often, collected data are incomplete due to problems of sensor readings, communication failures and the lack of environmental information (taxa). Accordingly, unsupervised or semi-supervised machine learning approaches are suitable for phytoplankton dynamics monitoring.

This paper focuses on how to build a marine monitoring system based on HF multisensor signal collected from MARELCarnot station (IFREMER, Boulogne-sur-Mer, France) in an unsupervised context. This marine station measures physicochemical and biological parameters every 20 minutes. A lack of information stops to set up a training database at high frequency. Indeed, no information are directly acquired by MAREL-Carnot station about phytoplankton taxonomic composition and local activities (e.g., dredging, opening dams). And, the resolution of complementary regional monitoring programmes in the area is too low (the objectives are differents).

Hidden Markov Model, noted HMM, is a well-adapted stochastic signal model to represent time series dynamics. The success of HMM in speech and handwriting recognition [6] leads to their application in marine monitoring. HMM approaches are based on static parameters defined by states and symbols, and dynamic parameters related to state transition and observation symbol probabilities. For instance, in speech recognition, a word is a sequence of phonemes (states) structured by transition probabilities where each phoneme is considered to be a spectral fingerprint (symbols) with some occurrence probabilities.

HMM building needs to estimate not only the number of states but also the characteristics of each of them. Commonly, HMM parameters are learned with labeled database or fixed with a priori information. Here, we address phytoplankton bloom

This work was supported in part by IFREMER, in part by the Artois-Picardie Water Agency with a Ph.D. Grant, and in part by the DYMAPHY Interreg IV A 2 Mers Seas and Zee en program. The IFREMERs Department REM and its unit Research and Technological Development have operational control for servicing the MAREL system.

K. Rousseeuw is both with the LISIC Lab (LISIC : Laboratoire dInformatique Signal et Image de la Cte dOpale address: ULCO/LISIC, BP 719, FR-62228 Calais cedex) and with French Research Institute for Exploitation of the Sea (IFREMER, address: IFREMER Centre MancheMer du Nord, BP 699, FR-62321 Boulogne-sur-Mer, France) (e-mail : kevin.rousseeuw@gmail.com).

E. Poisson Caillault and D. Hamad are with LISIC lab (e-mail: emilie.caillault@lisic.univ-littoral.fr; denis.hamad@lisic.univ-littoral.fr).

A. Lefebvre is with IFREMER - Centre MancheMer du Nord (e-mail: Alain.Lefebvre@ifremer.fr). Digital Object Identifier 10.1109/JSTARS.2014.2341219

forecasting issue using a hybrid HMM. The specific objective of this work is to design a system able to model phytoplankton dynamics from large database and no prior knowledge. For this purpose, a fully unsupervised HMM is built using spectral clustering algorithm to generate HMM symbols and states.

The paper is organized as follows. Section II describes the monitoring system and the proposed hybrid HMM with three parts. Part II-A discusses about usual unsupervised techniques to build HMM, and spectral clustering approach is argued to estimate HMM static parameters (states and symbols). Part II-B details HMM symbol generation by a self tuning fast K-means proposed algorithm. Part II-C defines HMM state generation by spectral clustering algorithm. Section III describes protocol of experimentations and collected data used from the IFREMER MAREL-Carnot station that registers water characteristics at HF resolution. First a fixed 2-state HMM is built in order to assess symbol and state generations thanks to an artificial labeling. Thus, our algorithms are compared with other machine learning techniques. Then, in section IV experiment results of a fully unsupervised N-state HMM are presented, and are related to examine biological expectations.

II. Monitoring system based on hybrid HMM

Fig. 1 presents the proposed monitoring system architecture. Data collected at high frequency resolution from 2005 to 2008 are first pretreated. Then, the clustering step is applied in order to find environmental states. The final step relies on temporal information between these states to develop a phytoplakton dynamics model. The built model is used to predict a new or forthcoming phytoplankton bloom, or specific states (classification/alert box in Fig. 1).

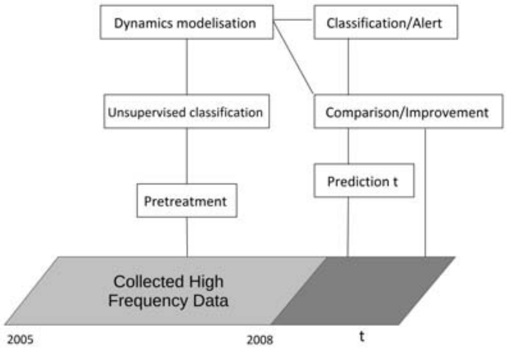


Fig. 1. General scheme of the monitoring system.

A. Hidden Markov Model

According to normal course of phytoplankton succession highlighted by Margalef [7] and Reynolds *et al.* [8] works, we assume that phytoplankton biomass is constrained by a high level of dependence among successive observations. Besides, it may be viewed as the result of a probabilistic walk along the environmental states. So let us see how to design one ergodic Hidden Markov Model to characterize the dynamics of phytoplankton blooms from physico-chemical and biological parameters in an unsupervised context.

A HMM noted $\lambda = \lambda(\mathbf{S}, \mathbf{V}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ is defined with 2 static sets $(\mathbf{S}, \mathbf{V})$ and 3 computed sets of probabilities $(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ that we recalled as follows [6]:

1) $\mathbf{S} = \{s_1, s_2, \dots, s_N\}$ is the set of states with N the number of distinct states. For instance: non-productive period, pre-bloom, bloom, post-bloom and other rare events, like dam opening, factory or agricultural discharge. The number of states is generally set by expert people in relation with the applications, or automatically determined by penalized maximum likelihood criterion.
2) $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_M\}$ is the set of symbols and M is the number of distinct symbols. In the simpliest cases, observation symbols correspond to the system outputs. Environmental state is not characterized by a unique representative per state, and two system outputs can belong to several states. Thus, it is necessary to build a codebook of symbols by vector quantization [9]–[11]. So data space will be represented by this codebook $\mathbf{V}$ of M symbols.
3) $\boldsymbol{\pi} = \{\pi_i\}$ of size N, defines the initial probability distribution, $\pi_i = P(s(t = 1) = s_i)\}$. There is no information about the state that will be predominant during data acquisition. *A priori* initial states are equiprobable.
4) $\mathbf{A} = \{a_{ij}\}$ of size $N \times N$, defines the transition matrix with $a_{ij} = P(s(t) = s_i | s(t-1) = s_j)$ the conditional probability. Therefore, the number of times that we move from a state s_i to state s_j is estimated, then $\mathbf{A}$ is normalized in row.
5) $\mathbf{B} = \{b_{ik}\}$ of size $N \times M$, defines the emission probability with $b_{ik} = P(\mathbf{v}(t) = \mathbf{v}_k | s(t) = s_i)$.

From a finite observation sequence without any labeled states, HMM symbols, transition and emission matrices [12] are adapted iteratively. Expectation-Maximisation approach (EM) is used with entropy criterion with Minimum Description Length

(MDL) constraint as Penalized Maximum Likelihood criterion [13]. Whatever the used criterion is (Bayesian Information Criterion and its derived), EM performance depends on initialization step, and can be time-consuming for large complex database. To avoid HMM iterative parameter estimate and the initialization step, we choose to use a spectral clustering approach to generate HMM state and symbol parameters from spatial information in one-pass algorithm.

Spectral clustering (SC) [14], [15] is a multi-cut method based on the eigen-decomposition of the Gram affinity matrix from the original dataset. Eigenvectors represent a new feature space where data are simply clustered by a K-means algorithm. It succeeds in clustering convex and non-convex distributed data. SC algorithm has been addressed for several applications: image segmentation, speech recognition, information retrieval, and so on [16]. Recently, algorithms have been developped to avoid their tuning requirements: to build the affinity function and to find the number of clusters. These steps are automatically completed using techniques, especially from [17], [18]. And, works [19], [20] allow to treat applications with a high volume of data.

The hybrid HMM building is schematically shown in Fig. 2. A step of vector quantization allows to extract HMM symbols. From these symbols, SC algorithm extracts HMM states. The HMM emission and transition probability matrices are then computed from the observed sequence. The transition matrix $\mathbf{A}$ is determined with the number of occurences moving from one state to another. $\mathbf{B}$ matrix corresponds to the number of times that observation $\mathbf{o}_t$ is both in a state s_i and in a symbol $\mathbf{v}_k$.

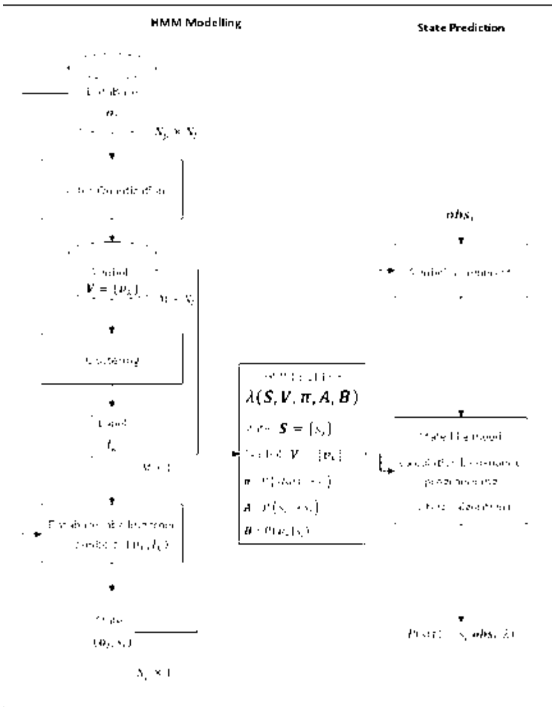


Fig. 2. Hybrid HMM building scheme. Dotted line separates HMM modelling on left side from state prediction of a new observation $\mathbf{obs}_t$ on right side.

When new data $\mathbf{o}_t$ is collected, it is associated with its nearest symbol. Viterbi algorithm [21], [22] is then applied to estimate its environmental state.

B. Symbol generation

MAREL-Carnot database consists of 26,280 × 19 parameters per year as from November 2004. To discover underlying states in this large database, instance selection is required. K-means algorithm is a well-adapted vector quantization method, and is popular for data clustering too [23]. The main idea is to build vector prototypes from a set of observations denoted $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{N_p}\}$ of N_p data points preserving HF information. The fast K-means algorithm [24] is modified to obtain a self tuning K-means based on Hartigan-Wong algorithm [25], and on Elbow criterion: the number of clusters K is incremented until a fixed percentage of explained variance or a $Kmax$ number of retained prototypes (i.e. symbols) is met. The principle of this proposed algorithm, named Self Tuning Fast K-Means (STFKM), is described in Fig 3. The center initialization for large databases (number of points $Np > 20,000$) is very important here to speed up the process convergence. $Kmax$ is the maximum number of reduced points specified by the user or in the default case, the number of measures in the time series. $varExplained$ is the explained variance desired by the user, by default this number is set to 95 percent.

```
procedure STFKM($\mathbf{O}, Kmax, varExplained$)
    if $varExplained$ not defined then
        $varExplained$=0.95
    end if
    if $Kmax$ not defined then $Kmax$=nrow($\mathbf{O}$)
    end if
    Variable: $k$=1, $vE$=0;
    while $k < Kmax$ or $vE < varExplained$ do
        $k = k + 1$;
        Step 1: Initialization of $k$ centers
        ..Cut Data in n subsamples of 20,000 points
        ..Compute K-means (K=$k$) on each subsample
        ..Select the $k$ clusters centers from the best partition according $MS(within)/MS(between)$
        Step 2: Decide the class memberships of the $N_p$ points by assigning them to the nearest center.
        Step 3: Re-estimate the $k$ cluster centers, by assuming the new memberships found
        Step 4: If none of $N_p$ points changed membership in the last iteration, Otherwise goto 2.
        Step 5: $vE = MS(between)/MS(total)$
    end while
    return $k$ obtained centers
end procedure
```

Fig. 3. Outline of Self Tuning Fast K-Means algorithm noted STFKM. The variance, MS(.) for Mean Square defined the variance between or within groups, or the total.

C. State generation by spectral clustering

After STFKM procedure on the pretreated data, M symbols summarize the entire database. From these M symbols, N states are detected by unsupervised clustering. Each MAREL-Carnot physico-chemical parameter follows a stochatic, non-linear and non-stationary process (except sea-level), see section III. They have not-gaussian distributions, and environmental state characterisation is unknown. So, SC method is the best way to avoid some assumptions about data shape. SC is capable of classifying data which are connected, but which are not necessarily compact, or clustered within convex boundaries. Indeed, the key idea of SC is to transform the input data space into a new feature space where K-means clustering could be applied. The most typical method by Ng, Jordan *et al.* [14] is recalled in Fig. 4.

```
procedure SpectralClustering($\mathbf{O}$, $K$)
    Variable: $\mathbf{W}, \mathbf{D}, \mathbf{Lap}, \mathbf{X}, \mathbf{Y}, \mathbf{l}$
    Compute a Gram affinity matrix $\mathbf{W}_{M\times M}$ from $\mathbf{O}$
    $\mathbf{D} = diag(\mathbf{W})$;                                        ▷ Degree matrix
    $\mathbf{Lap} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$;                                        ▷ Laplacian matrix
    Select the $K$-largest eigenvectors $\mathbf{x}$ of $\mathbf{Lap}$;
    Form the matrix $\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_k] \in \mathbb{R}^M$ by stacking the eigenvectors in column
    Form the $\mathbf{Y}$ matrix from the row-normalisation of $\mathbf{X}$ thus $y_{ij} = w_{ij}/(\sum_j x_{ij}^2)^{1/2}$
    $\mathbf{l}$ =K-means($\mathbf{Y}$,$K$)                                        ▷ each row of $\mathbf{Y}$ is a point
    Assign original Point $\mathbf{o}_i$ to the cluster $l_i$
    return label vector : $\mathbf{l}$
end procedure
```

Fig. 4. Spectral clustering algorithm

The number of clusters K in input of the SC algorithm and the way to build the Gram Affinity matrix $\mathbf{W}$ have both significant effects on the classification result. Gaussian kernel function is the most widely used function for constructing $\mathbf{W} = \{w_{ij}\}$ defined as:

$$w_{ij} = exp(-\frac{\|\mathbf{o}_i - \mathbf{o}_j\|^2}{2\sigma^2}) \tag{1}$$

The scaling parameter σ helps to sparse the matrix and tends to obtain an ideal case with a robust eigen-decomposition (i.e. in

TABLE I
LIST OF MAREL-CARNOT SIGNALS: ACRONYM, NAME AND MEASUREMENT FREQUENCY.

Acronym	Name	Frequency	RD	NC
E_TA	Air temperature	20 minutes	✓	
C_O21	Corrected dissolved oxygen	20 minutes	✓	✓
CSAL1	Salinity	20 minutes	✓	✓
CSAT1	Oxygen saturation percentage	20 minutes	✓	
E_CO1	Conductivity	20 minutes	✓	
E_LU1	P.A.R. Photosynthetically Available Radiation	20 minutes	✓	✓
E_O21	Non-corrected dissolved oxygen	20 minutes	✓	
E_PH1	pH	20 minutes		
E_TU1	Turbidity	20 minutes	✓	✓
E_VDM	Direction wind	20 minutes		
E_VVR	Gust wind speed	20 minutes	✓	✓
E_VVM	Average wind speed	20 minutes	✓	
ECHL1	Fluorescence	20 minutes		
EMAHH	Sea-level (measured)	20 minutes		
ETCO1	Water temperature	20 minutes	✓	✓
XMAHH	Sea-level (calculated)	20 minutes	✓	✓
C_PO1	Phosphate concentration	12 hours	✓	✓
C_NI1	Nitrate concentration	12 hours	✓	✓
C_SI1	Silicate concentration	12 hours	✓	✓

the ideal case, the first K eigenvalues are equal to one). However, a bad choice of σ brings an incorrect classification. Zelnik and Perona (ZP) [17], or Kong et al. [18] proposed a local scale parameter σ_i for each data $\mathbf{o}_i$ based on its neighborhood, instead of selecting a uniform scaling parameter σ. The ZP affinity matrix $\mathbf{W}$ is chosen with a z-neighborhood ($\mathbf{o}_{nz}$ the z^{th} neigborhood of the point $\mathbf{o}_i$):

$$w_{ij} = exp(-\frac{\|\mathbf{o}_i - \mathbf{o}_j\|^2}{2\sigma_i\sigma_j}) \text{ with } \sigma_i = \|\mathbf{o}_i - \mathbf{o}_{nz}\| \tag{2}$$

Many authors proposed to overcome the choice of the number of clusters K by analysing either eigenvalues magnitude (equal or nearest to one) or eigengap, or eigenvectors (see [18], [26], [27]). To select the number of states N for HMM topology, the eigengap method is used: it is the simpliest one to implement, and it has the lowest complexity.

From the spectral clustering of the M symbols $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_M\}$ issued from the STFKM step, we assign the observation data $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{N_p}\}$ thanks to the label $s_i = l_k$ of their cluster center $\mathbf{v}_k$.

III. DATA AND FIXED 2-STATE HMM VALIDATION

One HMM is built according to the scheme described in the previous section from MAREL-Carnot multivariate marine signals in order to model the phytoplankton dynamics in the French Channel coast around Boulogne-sur-Mer. Data and their curves are available on the website (http://www.ifremer.fr/difMarelCarnot/) with authorisation request. These data are first presented, then the experiment validation protocol follows.

Without ground truth on the environmental states and in order to assess our system, we decide to create an automatic data labeling based on the monitoring sampling strategy for the EU Water Directive Framework (EU-WFD). Thus, data from March to October are labeled s_1, corresponding to the productive period (in terms of biomass production capacity), and the others s_2 for the non-productive period. Furthermore, this labeling will allow to compare our sytem with other machine learning algorithms.

A. Data presentation

MAREL-Carnot station registers 19 signals: 16 water characteristics every 20 minutes, and 3 nutrient levels every 12 hours. These signals are detailed in Table I. Collected data signals come from different sensors. They require pretreatments to respect sensor range, and sensor time alignment. In case of sensor failures, its measurements are not retained (all pH values and measured sea-level values are removed). The sensor ranges are adapted to Boulogne-sur-Mer ecosystem. Time alignment is obtained by a moving average technique on a small sampling rate (20 minutes). Nutrient parameters are duplicated to obtain the same time resolution. From this step, signal database for 2005-2008 contains 105,192 points in $\mathbb{R}^{19}$. Only half of the data

(48,157 points) have no missing values (due to sensor default). To reduce missing value, a moving average over one week (temporal scale of a bloom) is applied, leading to a completed data of N_p =84,614 points.

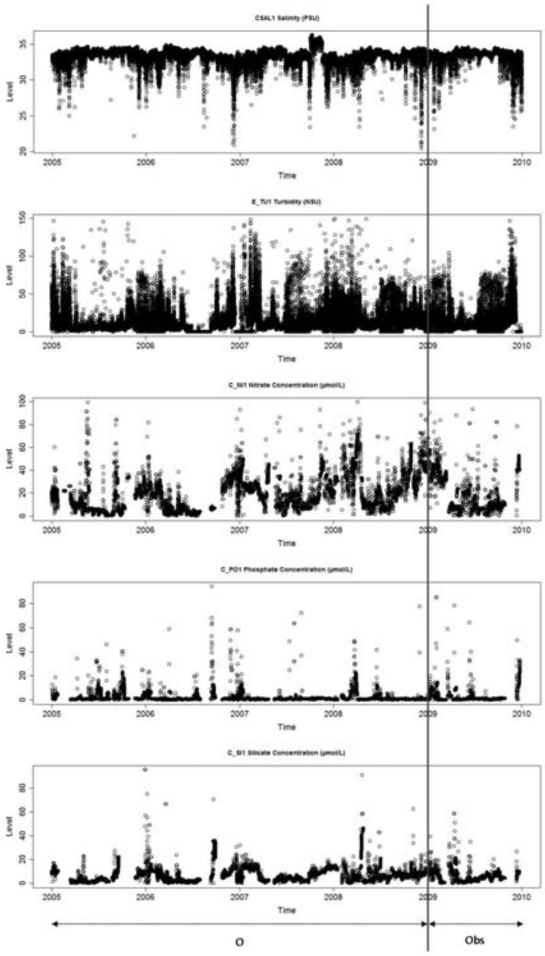


Fig. 5. Five completed MAREL-Carnot signals from 2005 to 2009 by mobile average: C_SAL1, E_TU1, C_NI1, C_PO1, C_SI1. Vertical line separates data into 2 subsets. Subset **O** corresponds to 2005-2008 period which is used for HMM building. Subset **Obs** relates to the year 2009 which is used for generalisation and accordingly does not participate in HMM building.

Fig. 5 illustrates 5 signals after this data completion: Salinity, Turbidity and nutrients (Nitrate, Phosphate and Silicate); residual missing measurements of nitrate concentrations, for instance, can be visualized. After a correlation analysis, N_f =10 physico-chemical parameters, not correlated, are retained and detailed in NC column of the Table I. Note that Fluorescence signal is not taken into account, but is used to validate clustering results since we have no ground truth: it is a presence indicator of phytoplankton cells.

When an observation data contains at least one missing value (among in $\mathbb{R}^{10}$), this point does not participate to the generation of symbols and states. Centering and standard deviation scalings are achieved on each parameter to avoid parameter range influence.

Data from 2005 to 2008 are considered to build HMM parameters. To test the time modeling, data from 2009 will be tested with the same pretreatment protocol.

B. Vector quantization validation

The number of symbols $\mathbf{V}$ required to characterize a state is first analyzed. Selection of the M symbols from the data is performed by 100 random drawings. A 1-Nearest Neighbor algorithm (1-NN) is used to evaluate the approximation of a mixture of components per state. 1-NN is first applied on each parameter, and then on the multidimensional matrix (from 2005 to 2008). K-means algorithm is also applied to build $K = M$ representatives per state.

Two scores are considered: Rate Recognition (RR) and the monthly Overlap defined by the following equation:

$$Overlap = \frac{\sum_i [|s_1(i)| + |s_2(i)| - max(|s_1(i)|, |s_2(i)|)]}{|s_1 \cup s_2|} \qquad (3)$$

TABLE II
1-NN RR and Overlap scores (mean and standard deviation in percent) for M representatives per state according EU-WFD labeling : 2005-2008 database

M values	Random selection		K-means	
	RR	Overlap	RR	Overlap
1	68.1 (8.9)	18.4 (6.8)	82.7 (0.1)	11.3 (0.1)
10	79.4 (4.0)	18.5 (3.7)	89.8 (0.8)	10.0 (0.7)
100	87.6 (0.9)	12.2 (0.9)	95.1 (0.2)	4.9 (0.2)
1,000	94.7 (0.2)	5.2 (0.2)	**98.4 (0.1)**	**1.6 (0.1)**

TABLE III
STFKM influence measured with a SVM approach: RR and Overlap scores (in percent)

Sampling	Training		Test	
	RR	Overlap	RR	Overlap
No sampling-SVM	97.9	2.1	92.9	7.0
Random 1,000-SVM	93.3	6.7	92.2	6.7
STFKM-SVM	95.4	4.6	92.6	7.4

$|.|$ is the cardinal operator and $|s_1(i)|$ defines the number of points labeled s_1 during the i^{th} month. Phytoplankton productive and non-productive periods are expected to have no overlap according to EU-WFD.

For the monodimensional analysis, water temperature ETCO1 is the most discriminative parameter, with a recognition rate from 75.1% (± 3.5) for one representative per state, to 77.8% (± 0.4) for 1,000 representatives. For the multidimensional analysis, Table II summarizes the mean and standard deviation of the two scores, RR and Overlap, for different M-values. Approximating data distribution with one unique random representative gives poor recognition rate (68.1%) and often an important Overlap (18.4%) of the two desired environmental states. To decrease this Overlap around 10%, more than 100 random representatives are required.

K-means is a geometric approach adapted for linearly separable data sets. This algorithm requires to know the desired number of symbols (centers) M. Here with 10 symbols per state, the Overlap is around 10%. To reduce this Overlap around 5%, 100 representatives per state are required.

The proposed STFKM automatically searches the number of symbols that describes the data structure, and it is fast running. The impact of the STFKM selection is tested with a learning machine: a Support Vector Machine (SVM). A SVM model (radial basis kernel) was trained on 2005-2008 data and was tested on 2009 data with 10 cross-validation. Three experiments for the SVM training were led: on all training data, on 1,000 randomly selected representatives per state (M =2,000) of this data, and on symbols issued from STFKM vector quantization.

Table III summarizes SVM capacities of training and generalisation (test) for these 3 studies with RR and Overlap scores. In generalisation, STFKM algorithm allows to keep a similar recognition rate (92.6%) and Overlap score (7.4%) to no sampling-SVM. STFKM-SVM gives better training capacity than a random selection of M symbols. In this supervised context, we can conclude that the obtained vector quantization by STFKM is a relevant data reduction.

The stability of STFKM algorithm is then assessed according to the Rand Index (RI) of 10 achieved symbol generations. RI score [28] is a measure of similarity between two data clusterings $\mathbf{P}_1$ and $\mathbf{P}_2$ of a given set of n elements $\mathbf{E} = \{e_1, \ldots, e_n\}$. Note that the number of clusters in each partition can be different. RI is computed according to the following equation:

$$RI(\mathbf{P}_1, \mathbf{P}_2) = \frac{(a+b)}{\binom{n}{2}} \tag{4}$$

where a (resp. b) is defined by the number of pairs of elements in $\mathbf{E}$ that are in the same set in $\mathbf{P}_1$ (resp. in different sets) and in the same set in $\mathbf{P}_2$ (resp. in different sets) .

In a fully unsupervised context, STFKM approach allows to respect the high frequency information without losing data structure. Table IV shows that RI scores of the 10 obtained partitions and the number of retained symbols are quite similar. The RI score near to one shows that STFKM algorithm gives a robust vector quantization.

TABLE IV
Rand Index and M-value boxplot values for 10 symbol generations

Boxplot	Min	Q1	Median	Mean	Q3	Max
RI	0.99	0.99	0.99	0.99	0.99	0.99
M	2744	2749	2754	2759	2763	2790

TABLE V
RR AND OVERLAP SCORES (IN PERCENT) OF UNSUPERVISED CLASSIFICATION

Sampling	Building Database		Test Database	
	RR	Overlap	RR	Overlap
S-EM (C_SAL1)	70.1	14.7	77.0	11.9
STFKM-EM (EII)	83.6	13.4	91.2	4.3
STFKM-HC	66.9	0.6	66.9	0.5
STFKM-SC	79.1	11.9	84.1	5.7

TABLE VI
SIMILARITY SCORES FOR SIGNAL RECONSTRUCTION BASED ON STATE CLASSIFICATION FROM HMM MODEL.

Year	2005	2006	2007	2008	Mean	2009
sim	0.87	0.84	0.83	0.86	0.85	0.79

C. State generation validation

Then, the spectral clustering algorithm is compared to usual unsupervised approaches : EM, Hierarchical Clustering (HC) for a set number of states $N = 2$ with the same STFKM symbols. For experiments, a 7-neighborhood is considered for the scaling parameter of the similarity matrixin SC. We used EM and HC algorithms implemented in R-Gui (library Mclust and stats: http://www.r-project.org/). Only the best or runnable options are retained: Expectation-Maximisation (EII model), HC with complete linkage. EM is also performed on each of the 10 parameters, only results of the most discriminative parameter (salinity CSAL1) are presented. 1-NN algorithm is used to label data from the built model. The 2005-2008 period is named building database, and the year 2009 test database.

Table V presents RR and Overlap scores to analyse jointly. In spite of its very low Overlap score around 0.5%, hierarchical clustering (cutting obtained tree to 2 clusters) does not separate productive and non-productive periods. State 1 represents 84,118 of 84,614 points, 99% of the building database and state 2 concerns few points in August, September and November for the building database and February, April and June for the test database without any biological or sensor interpretation. EM approach offers the best RR results for the two databases. STFKM approach gives lower RR than EM one, but its Overlap for the largest database (building database) is reduced. STFKM-SC approach is a balanced one for EU-WFD labeling, and will be relevant for a number of states greater than 2. Indeed, we expect that our hybrid HMM system can detect more than 2 states like phytoplankton spring bloom or automnal bloom, rare events.

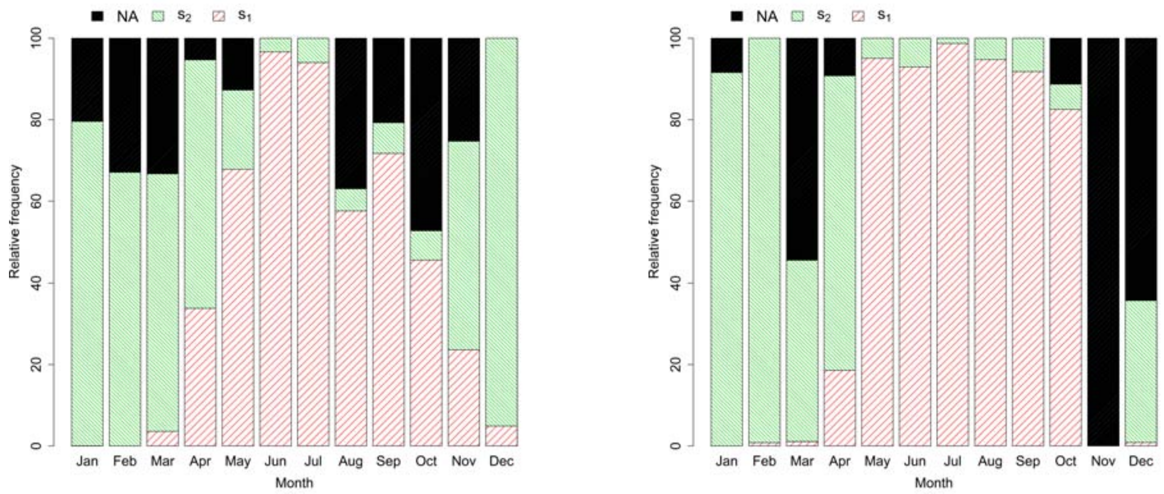
D. Time modeling validation, fixed 2-state HMM

We evaluate, through experiments, the reliability of our hybrid model: the entire procedure for building one HMM from clustering is repeated 10 times. The 10 partitions have a mean RI score of 0.95, so the whole STFKM-SC step (symbol and state generation) is robust. We keep the symbol and state partition with the smallest normalised multi-cut, MNCut [14] to build HMM. For experiments, the number of states is set: $N = 2$ and other parameters are automatically tuned: explained variance is fixed to 95 percent. According to the MNCut criterion, HMM built has $M = 2794$ symbols.

According to the EU-WFD labeling, 79.3 percent of building database is well recognized with an Overlap of 11.7%, and 82.1 percent of 2009 test database is well recognized with an Overlap of 6.7%. Fig. 6 illustrates the distribution of labeled states by HMM prediction for the building database and the 2009 test database. In 2005-2008, state s_1 in red color ties in with the period from March to December with a dominant April-November period whereas state s_2 in green color is dominant in the November-April period. Over the period 2009 state s_2 ties in with the period from April to October, whereas state s_2 in green color is dominant in the December-April period. Many data in March and August-November have no estimated state (noted NA in black color in Fig. 6) due to one least missing value in $\mathbb{R}^{10}$; that means the system forecasts confusions in transition periods. But states s_1 and s_2 match well with the two main environmental EU-WFD states: s_1 and s_2 characterize the phytoplankton dynamics, the productive and the non-productive periods.

To validate the relevancy of the built model from 2005-2008 and its symbol quantization in a fully unsupervised context, MAREL-Carnot signals $\widehat{\mathbf{Obs}}$ on the year 2009 are reconstructed and compared with the original data $\mathbf{Obs}$. For an observation $\mathbf{obs}_t$, the system estimates its state s_i with the higher likelihood. Then the most present symbols $\mathbf{v}_k$ in this state are retained, see Fig. 7. A similarity sim score is defined by the following equation 5:

$$sim(\mathbf{Obs}, \widehat{\mathbf{Obs}}) = \frac{1}{|\mathbf{Obs}|} \sum_{t=1}^{|\mathbf{Obs}|} \frac{1}{\left\| \mathbf{obs}_t - \widehat{\mathbf{obs}_t} \right\| + 1} \tag{5}$$



(a) Period 2005-2008, HMM building database.

(b) 2009, test database.

Fig. 6. State distribution per month for a typical seasonal cycle: state s_1 is represented in red color and s_2 in green color. The black color, named NA, concerns measures whose state is not estimated (one least missing values).

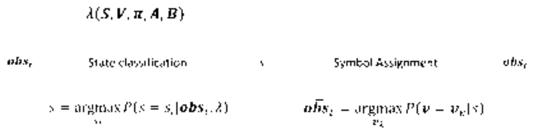

$\lambda(S, V, \pi, A, B)$
obs_t
State classification
s
Symbol Assignment
$\hat{obs}_t$
$s = \underset{s_i}{\operatorname{argmax}} P(s = s_i | obs_t, \lambda)$
$\hat{obs}_t = \underset{v_k}{\operatorname{argmax}} P(v = v_k | s)$

Fig. 7. Time series reconstruction from HMM classification. Signal $\widehat{\mathbf{Obs}}$ is reconstructed from original observation $\mathbf{Obs}$ thanks to the M symbols.

From 2005 to 2008, each year participates to HMM building, and reconstructed signals are compared to the original ones so as to assess the modeling power. Table VI shows the similarity scores, Eq.(5), between original and reconstructed signals for each year over the period 2005-2008 and their related mean. The last column corresponds to the year 2009. This similarity score is bounded by 0 and 1: 1-value implies that reconstructed signals are exactly the same as the original ones. The similarity scores and their mean are greater than 0.83 from 2005 to 2008. So, we conclude that reconstruted signals are very close to the original data, and that the vector quantization algorithm for HMM states is efficient. Therefore, the proposed system has an interesting generalisation power, for the year 2009, which does not participate in HMM building. Indeed, the time series are built with a high similarity score: 0.79 according to the choice of the most probable symbol of the state. Fig. 8 illustrates original and reconstructed signal of one parameter, dissolved oxygen concentration C_O21 for 2009.

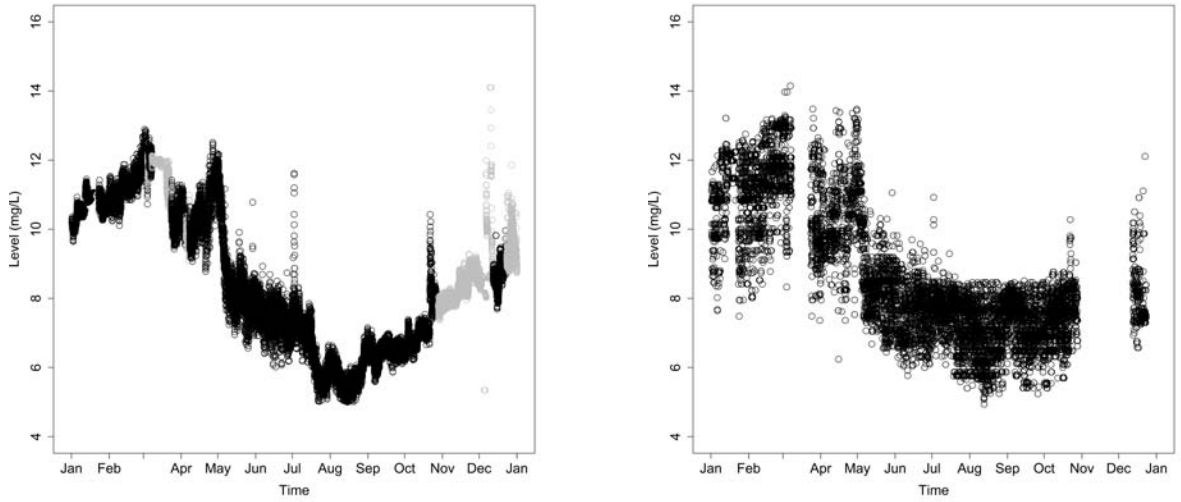
IV. N-STATE HMM FOR PHYTOPLANKTON DYNAMICS

Considering that HMM is now validated on a fixed 2-state biological dynamics, the next step is to increase the number of states to refine and to try to better understand the bloom determinism and its dynamics. A 7-state HMM with M =2,884 symbols is built from 2005-2008 database, $N = 7$ according to the eigengap technique. Time modeling validation is achieved according to the same protocol as the one of the 2-state HMM. Reconstructed 2009 signals have a similarity score above 0.8 with the original data.

Fig. 9 is the color-state projection of the estimated states over the period 2005-2008 on Fluorescence signal; black color denotes unlabelled observations due to missing value. The state distribution and sequencing are illustrated in Fig. 10 with the same color standard.

To interpret this model and to relate a ecological meaning, we analyze the state sequencing and the characterization of each state by a correlation analysis (Table VII).

State s_6 (yellow) clearly highlights high salinity values, ranging from 33.9 to 36.2 with a mean value of 35.4 (Fig. 5). These values are more representative of offshore waters. And then we can conclude in this coastal zone that mainly state s_6 corresponds to salinity anomalies (sensor failures). Nevertheless, s_6 may sometimes be explained when west winds persist and consequently bring more offshore waters to the coast. State s_3 (blue) is representative of the winter non-productive period, with high nutrient concentrations and low temperature (Table VII).



(a) C_O21 original signal.

(b) $\widehat{C\_O21}$ estimated signal.

Fig. 8. 2009 - Dissolved oxygen concentration signals: original signal (a) and estimated signal (b) by HMM. In gray color, time measurements are not estimated by HMM due to at least one missing parameter at this time.

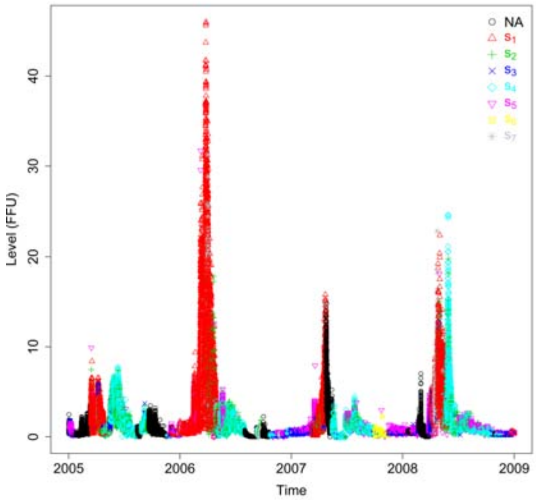
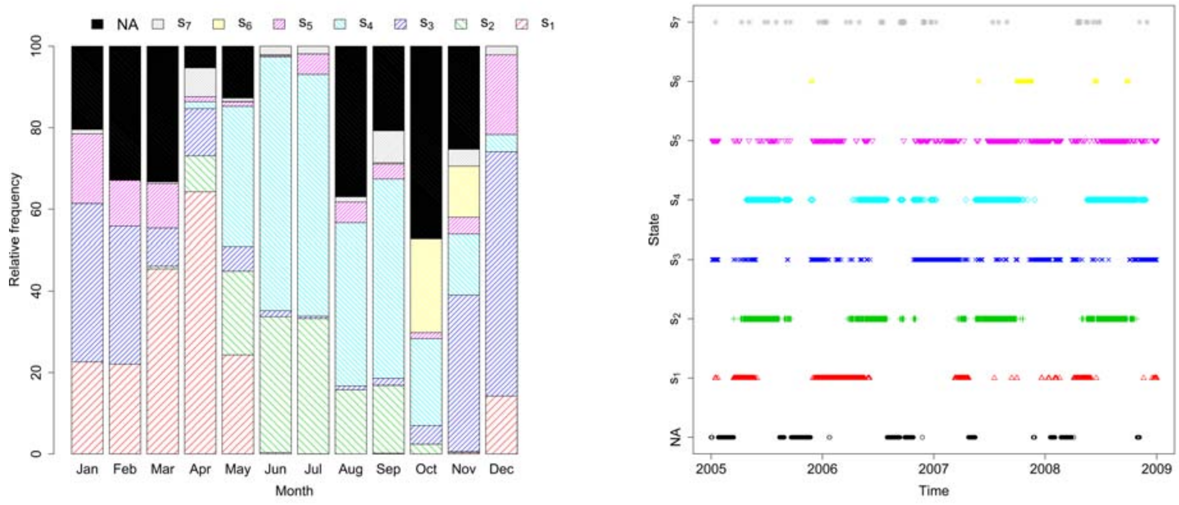


Fig. 9. Clustering results: Color-state sequencing projection on the fluorescence signal for 2005-2008. Black color, named NA concerns measures whose state is not estimated (one least missing value).

The initiation of the main phytoplankton bloom, and then the growing phytoplankton stage (between February and May: inter-annual variability of the bloom) are characterized by s_1 (red). High oxygen concentrations, explained by a high phytoplankton production (photosynthesis) are observed during this stage (Table VII). During state s_1, phytoplankton mainly uses the winter nutrients stock, and consequently this state corresponds to the new production period [29]. States s_2 and s_4 follow state s_1, and are identified as the regenerated production period when phytoplankton production is based on regenerated nutrients (transformation of the organic matter from the previous bloom - state s_1 - into new available nutrients).

States s_5 (pink) and s_7 (grey) correspond to rare or short events, respectively, with high turbidity during storm events, and high phosphate and silicate concentrations (C_PO1 and C_SI1 in Table VII). More investigations are required to better understand the main processes involved during these periods.

Fig. 11 illustrates the predicted states for the year 2009 with their sequencing. The 7-state HMM succeeds in predicting phytoplankton biomass dynamics. The state sequencing matches our assumption with a pre-bloom winter period (state s_3 mainly) followed by the main phytoplankton bloom based on external nutrient inputs (state s_1), and the regenerated bloom (states s_2 and s_4).



(a) State distribution per month for a typical seasonal cycle over the period 2005-2008.

(b) Measure sequencing in each state over the period 2005-2008.

Fig. 10. Clustering results: NA concerns measures whose state is not estimated (one least missing value).

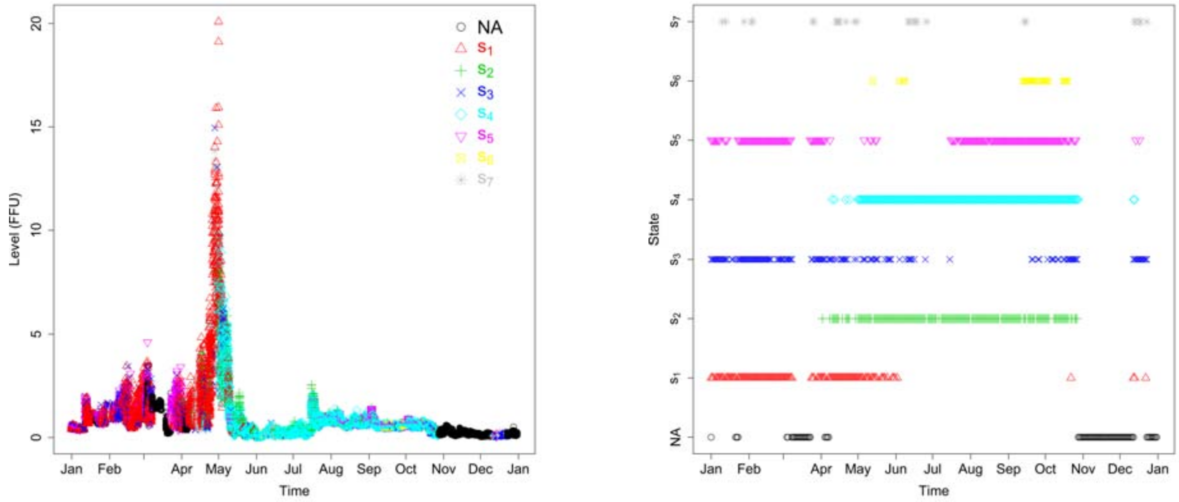
TABLE VII
CORRELATIONS BETWEEN PARAMETERS AND STATES (IN BOLD: THE HIGHEST CORRELATION COEFFICIENTS).

State	s_1	s_2	s_3	s_4	s_5	s_6	s_7
Color	red	green	blue	cyan	pink	yellow	grey
C_NI1	-0.14	-0.24	**0.54**	-0.24	0.12	0.02	-0.03
C_O21	**0.64**	-0.16	-0.03	**-0.38**	0.05	-0.17	0.01
C_PO1	-0.08	-0.03	-0.04	-0.07	-0.02	-0.03	**0.57**
C_SI1	-0.11	-0.16	0.20	-0.25	0.15	0.04	**0.47**
CSAL1	0.12	0.13	**-0.36**	0.10	-0.25	**0.42**	0.00
E_LU1	-0.08	**0.73**	-0.21	-0.24	-0.06	-0.06	0.00
E_TU1	-0.11	-0.14	-0.01	-0.22	**0.76**	-0.03	-0.05
E_VVR	-0.25	-0.03	0.16	-0.05	**0.31**	-0.06	-0.03
ETCO1	**-0.50**	**0.32**	**-0.37**	**0.56**	-0.14	0.07	0.05
XMAHH	-0.00	0.03	0.02	-0.03	0.01	0.01	0.00

V. CONCLUSIONS AND FUTURE WORKS

Two N-state HMM was built in order to forecast phytoplankton blooms near the French Channel coast from MAREL-Carnot signals (IFREMER, Boulogne-sur-Mer) without any biological knowledge. HMM building requires at least two parameters: a number of states, a number of symbols that characterize states. These parameters are commonly estimated iteratively by Expectation-Maximisation. We propose a one-pass process to estimate HMM symbols and states in a fully unsupervised context. A proposed Self Tuning Fast K-Means STFKM algorithm extracts symbols from observation data. From this vector quantization, a spectral clustering approach, with no tuning too, generates HMM states that allows to treat non-convex data. A signal reconstruction approach is proposed to assess HMM prediction.

Result analyses from the MAREL-Carnot buoy data first demonstrate interests and the stability of each used algorithm (state and symbol generation) throughout the monitoring chain. The high resolution information is preserved. Built 2-state HMM permits to detect the main productive and non-productive periods, as used for the purposes of the EU Water Framework Directive to assess good environmental status. A 7-state HMM was proposed to refine knowledge about phytoplankton bloom dynamics in a temperate ecosystem, temporarily dominated by a harmful algae (*Phaeocsytis globosa*). The obtained state sequencing coincides with dynamics described using measurements from low resolution system near the MAREL-Carnot (Rephy/SRN data [30]). The proposed HMM system succeeds in characterizing phytoplankton dynamics from new incoming data (in near real-time approach). Using the main statistical characteristics of the parameters underlying the definition of a given state, the system will allow to further increase knowledge about the main controlling or forcing parameters (*i.e.*, nutrient pressure, light availability, turbidity), the environmental status (*e.g.*, phytoplankton biomass), and the direct and/or indirect effects of such blooms (*e.g.*, oxygen concentration).



(a) Color-state projection on Fluorescence signal.

(b) State sequencing.

Fig. 11. Predicted state results in 2009 by HMM: Predict state results in 2009 by HMM: (a) Color-state projection on Fluorescence signal and (b) state sequencing. NA (black color) corresponds to measures with no estimated state.

The main limiting step in the monitoring chain is the removing samples with missing values. Indeed, the latter affects the state estimation and characterization (symbol process). Some phytoplankton blooms were not taken into account for HMM building.

Several environmental monitoring and research programmes could benefit from the proposed method to avoid the critical expert labeling step when modelling. It could help to process large multivariate time series as generated by high resolution (in time and/or space) platforms, more and more frequently implemented for the integrated observation of pelagic ecosystems and biogeochemical cycles in the oceans. Moreover, the possibility of identifying environmental states (characterized by a combination of several parameters) is a clear opportunity to better understand what a good environmental status is, as defined and used for the needs of the WFD, the MSFD or other regional sea convention (as OSPAR).

References

[1] *Directive 2000/60/EC of the European Parliament and of the Council. Establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327/1.*, 2000.

[2] *Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive)*, 2008.

[3] B. Gokaraju, S. Durbha, R. King, and N. Younan, "A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the gulf of mexico," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 4, no. 3, pp. 710–720, Sept 2011.

[4] ——, "Ensemble methodology using multistage learning for improved detection of harmful algal blooms," *Geoscience and Remote Sensing Letters, IEEE*, vol. 9, no. 5, pp. 827–831, Sept 2012.

[5] G. Pereira and N. Ebecken, "Combining in situ flow cytometry and artificial neural networks for aquatic systems monitoring." *Expert Systems with Applications: An International Journal*, vol. 38, no. 8, pp. 9626–9632, 2011.

[6] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] R. Margalef, "Life-forms of phytoplankton as survival alternatives in an unstable environment," *Oceanologica acta*, vol. 1, pp. 493–509, 1978.

[8] C. S. Reynolds, V. Huszar, C. Kruk, L. Naselli-Flores, and S. Melo, "Towards a functional classification of the freshwater phytoplankton," *Journal of Plankton Research*, vol. 24, no. 5, pp. 417–428, May 2002.

[9] J. M. Koo, H. Lee, and C. Un, "An improved vq codebook design algorithm for hmm," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, 1992, pp. 357–360 vol.1.

[10] A. H.-R. Ko, R. Sabourin, and A. de Souza Britto Jr., "A new hmm training and testing scheme." in *ICPR*. IEEE, 2008, pp. 1–4.

[11] M. Debyeche, J. P. Haton, and A. Houacine, "Improved vector quantization approach for discrete hmm speech recognition system." *Int. Arab J. Inf. Technol.*, vol. 4, no. 4, pp. 338–344, 2007.

[12] X. Liao, P. Runkle, and L. Carin, "Identification of ground targets from sequential high-range-resolution radar signatures," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 4, 2002.

[13] M. A. T. Figueiredo, S. Member, and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 381–396, 2002.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 849–856.

[15] U. von Luxburg, "A tutorial on spectral clustering," *CoRR*, vol. abs/0711.0189, 2007.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[17] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 1601–1608.

[18] W. Kong, S. Hu, J. Zhang, and G. Dai, "Robust and smart spectral culstering from normalized cut," *Neural Computing and Applications*, vol. 23, no. 5, pp. 1503–1512, October 2013.

[19] D. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in *15th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, Paris, France, 2009, pp. 907–916.

[20] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation." in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, W. Burgard and D. Roth, Eds. AAAI Press, 2011, pp. 313–318.

[21] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, 1967.

[22] G. J. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[23] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[24] M. Shindler, A. Wong, and A. Meyerson, "Fast and accurate k-means for large datasets." in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Granada, Spain: Curran Associates, Inc., 2011, pp. 2375–2383.

[25] J. Hatrigan and M. Wong, "A k-means clustering algorithm," *Journal of Royal Statistical Society. Series C (Applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[26] G. Sanguinetti, J. Laidler, and N. D. Lawrence, "Automatic determination of the number of clusters using spectral algorithms.in," in *IEEE Machine Learning for Signal Processing. 28-30 Sept 2005*, 2005, pp. 28–33.

[27] T. Xiang and S. Gong, "Spectral clustering with eigenvector selection," *Pattern Recogn.*, vol. 41, no. 3, pp. 1012–1029, Mar. 2008.

[28] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[29] V. Gentilhomme and F. Lizon, "Seasonal cycle of nitrogen and phytoplankton biomass in a well-mixed coastal system (eastern english channel)," *Hydrobiologia*, vol. 361, pp. 191–199, 1997.

[30] A. Lefebvre, N. Guiselin, F. Barbet, and F. L. Artigas, "Long-term hydrological and phytoplankton monitoring (1992-2007) of three potentially eutrophic systems in the eastern English Channel and the Southern Bight of the North Sea," *ICES Journal of Marine Science*, vol. 68, no. 10, pp. 2029–2043, Sep. 2011.