Hybrid Hidden Markov Model for Marine Environment Monitoring
Kévin Rousseuw, Emilie Poisson Caillault, Alain Lefebvre, Denis Hamad

To cite this version:
Kévin Rousseuw, Emilie Poisson Caillault, Alain Lefebvre, Denis Hamad. Hybrid Hidden Markov Model for Marine Environment Monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IEEE, 2015, 8 (1), pp.204-213. 10.1109/JSTARS.2014.2341219. hal-01442974

HAL Id: hal-01442974
https://hal.archives-ouvertes.fr/hal-01442974
Submitted on 21 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Hybrid Hidden Markov Model for Marine Environment Monitoring

Kévin Rousseeuw, Ph.D. Student Member, IEEE, Émilie Poisson-Caillault, Alain Lefebvre, and Denis Hamad

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 phytoplankton blooms but also to understand their dynamics. HMM parameters are usually estimated by an iterative Expectation-Maximisation approach. We propose to estimate HMM parameters by using spectral clustering algorithm. The monitoring system is assessed on database signals from MAREL-Carnot station (Boulogne-sur-Mer, France). Experiment results show that the proposed system is efficient to detect environmental states such as phytoplankton productive and non productive periods without a priori knowledge. Furthermore, discovered states are consistent with biological interpretation.

Index Terms
Hybrid Hidden Markov Model, marine water monitoring, Phytoplankton blooms, spectral clustering.

I. 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 MAREL-Carnot station (IFREMER, Boulogne-sur-Mer, France) in an unsupervised context. This marine station measures physico-chemical 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 different).

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
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 phytoplankton dynamics model. The built model is used to predict a new or forthcoming phytoplankton bloom, or specific states (classification/alert box in Fig. 1).

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 $λ$ = $(S, V, π, A, B)$ is defined with 2 static sets $(S, V)$ and 3 computed sets of probabilities $(π, A, B)$ that we recalled as follows [6]:

1) $S = \{s_1, s_2, \ldots, s_N\}$ 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) $V = \{v_1, v_2, \ldots, v_M\}$ is the set of symbols and $M$ is the number of distinct symbols. In the simplest 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 $V$ of $M$ symbols.

3) $π = \{π_i\}$ of size $N$, defines the initial probability distribution, $π_i = P(s(t=1) = s_i)$. There is no information about the state that will be predominant during data acquisition. A priori initial states are equiprobable.

4) $A = \{a_{ij}\}$ 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 $A$ is normalized in row.

5) $B = \{b_{ik}\}$ of size $N \times M$, defines the emission probability with $b_{ik} = P(v(t) = v_k|s(t) = s_i)$.

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 developed 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 $A$ is determined with the number of occurrences moving from one state to another. $B$ matrix corresponds to the number of times that observation $o_t$ is both in a state $s_i$ and in a symbol $v_k$.

![Fig. 2. Hybrid HMM building scheme. Dotted line separates HMM modelling on left side from state prediction of a new observation $\text{obs}_t$ on right side.](image)

When new data $o_t$ 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 $26,280 \times 19$ 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 $O = \{o_1, o_2, \ldots, o_{N_p}\}$ 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 $K_{max}$ 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 $N_p > 20,000$) is very important here to speed up the process convergence. $K_{max}$ is the maximum number of reduced points specified by the user or in the default case, the number of measures in the time series. $\text{var}^{\text{Explained}}$ is the explained variance desired by the user, by default this number is set to 95 percent.
1: procedure STFKM(O,Kmax,varExplained)
2: if varExplained not defined then
3: varExplained=0.95
4: end if
5: if Kmax not defined then
6: Kmax=nrow(O)
7: end if
8: Variable: k=1, vE=0;
9: while k<Kmax or vE < varExplained do
10: k = k + 1;
11: Step 1: Initialization of k centers
12: ..Cut Data in n subsamples of 20,000 points
13: ..Compute K-means (K=k) on each subsample
14: ..Select the k clusters centers from the best partition according
15: MS(within)/MS(between)
16: Step 2: Decide the class memberships of the Np points by assigning them to the nearest center.
17: Step 3: Re-estimate the k cluster centers, by assuming the new memberships found
18: Step 4: If none of Np points changed membership in the last iteration, Otherwise goto 2.
19: Step 5: vE = MS(between)/MS(total)
20: end while
21: return k obtained centers
22: end procedure

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 stochastic, 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.

1: procedure SPECTRALCLUSTERING(O, K)
2: Variable: W, D, Lap, X, Y, l
3: Compute a Gram affinity matrix W M×M from O
4: D = diag(W); \hfill \triangleright \text{Degree matrix}
5: Lap = D^{−1/2}WD^{−1/2}; \hfill \triangleright \text{Laplacian matrix}
6: Select the K-largest eigenvectors x of Lap;
7: Form the matrix X = [x_{1}, x_{2}, \ldots x_{k}] ∈ R^{M} by stacking the eigenvectors in column
8: Form the Y matrix from the row-normalisation of X thus y_{ij} = w_{ij}/( \sum_{j} w_{ij} )^{1/2}
9: l = K-means(Y, K); \hfill \triangleright \text{each row of Y is a point}
10: Assign original Point o_{i} to the cluster l_{i}
11: return label vector : l
12: 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 W have both significant effects on the classification result. Gaussian kernel function is the most widely used function for constructing W = \{w_{ij}\} defined as:

\[ w_{ij} = \exp\left(-\frac{||o_{i} - o_{j}||^{2}}{2\sigma^{2}}\right) \]  

The scaling parameter \sigma 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.**

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

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

\[
w_{ij} = \exp\left(-\frac{||o_i - o_j||^2}{2\sigma_i \sigma_j}\right) \quad \text{with} \quad \sigma_i = ||o_i - o_{nz}||
\]

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 simplest one to implement, and it has the lowest complexity.

From the spectral clustering of the \( M \) symbols \( V = \{v_1, v_2, \ldots, v_M\} \) issued from the STFKM step, we assign the observation data \( O = \{o_1, o_2, \ldots, o_N\} \) thanks to the label \( s_i = l_k \) of their cluster center \( v_k \).

### 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 system 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...
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_NH1, 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 \( O_{obs} \) 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 \( V \) 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:

\[
Overlap = \frac{\sum_l [\| s_1(l) \| + \| s_2(l) \| - \max(\| s_1(l) \|, \| s_2(l) \|)]}{| s_1 \cup s_2 |}
\]
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

<table>
<thead>
<tr>
<th>M values</th>
<th>Random selection</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>Overlap</td>
</tr>
<tr>
<td>1</td>
<td>68.1 (8.9)</td>
<td>18.4 (6.8)</td>
</tr>
<tr>
<td>10</td>
<td>79.4 (4.0)</td>
<td>18.5 (3.7)</td>
</tr>
<tr>
<td>100</td>
<td>87.6 (0.9)</td>
<td>12.2 (0.9)</td>
</tr>
<tr>
<td>1,000</td>
<td>94.7 (0.2)</td>
<td>5.2 (0.2)</td>
</tr>
</tbody>
</table>

TABLE III
STFKM INFLUENCE MEASURED WITH A SVM APPROACH: RR AND OVERLAP SCORES (IN PERCENT)

<table>
<thead>
<tr>
<th>Sampling</th>
<th>Training RR</th>
<th>Training Overlap</th>
<th>Test RR</th>
<th>Test Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>No sampling-SVM</td>
<td>97.9</td>
<td>2.1</td>
<td>92.9</td>
<td>7.0</td>
</tr>
<tr>
<td>Random 1,000-SVM</td>
<td>93.3</td>
<td>6.7</td>
<td>92.2</td>
<td>6.7</td>
</tr>
<tr>
<td>STFKM-SVM</td>
<td>95.4</td>
<td>4.6</td>
<td>92.6</td>
<td>7.4</td>
</tr>
</tbody>
</table>

| | 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 \(P_1\) and \(P_2\) of a given set of \(n\) elements \(E = \{e_1, \ldots, e_n\}\). Note that the number of clusters in each partition can be different. RI is computed according to the following equation:

\[
RI(P_1, P_2) = \frac{(a + b)}{\binom{n}{2}}
\]

where \(a\) (resp. \(b\)) is defined by the number of pairs of elements in \(E\) that are in the same set in \(P_1\) (resp. in different sets) and in the same set in \(P_2\) (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

<table>
<thead>
<tr>
<th>Boxplot</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>M</td>
<td>2744</td>
<td>2749</td>
<td>2754</td>
<td>2759</td>
<td>2763</td>
<td>2790</td>
</tr>
</tbody>
</table>
TABLE V
RR AND OVERLAP SCORES (IN PERCENT) OF UNSUPERVISED CLASSIFICATION

<table>
<thead>
<tr>
<th>Sampling</th>
<th>Building Database</th>
<th>Test Database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR</td>
<td>Overlap</td>
</tr>
<tr>
<td>S-EM (C_SAL1)</td>
<td>70.1</td>
<td>14.7</td>
</tr>
<tr>
<td>STFKM-EM (EII)</td>
<td>83.6</td>
<td>13.4</td>
</tr>
<tr>
<td>STFKM-HC</td>
<td>66.9</td>
<td>0.6</td>
</tr>
<tr>
<td>STFKM-SC</td>
<td>79.1</td>
<td>11.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Year</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>Mean</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim</td>
<td>0.87</td>
<td>0.84</td>
<td>0.83</td>
<td>0.86</td>
<td>0.85</td>
<td>0.79</td>
</tr>
</tbody>
</table>

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 matrix in 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 autumnal 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 $R_{10}$; 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 $\hat{\text{Obs}}$ on the year 2009 are reconstructed and compared with the original data $\text{Obs}$. For an observation $\text{obs}_t$, the system estimates its state $s_i$ with the higher likelihood. Then the most present symbols $v_k$ in this state are retained, see Fig. 7. A similarity $\text{sim}$ score is defined by the following equation 5:

$$sim(\text{Obs}, \hat{\text{Obs}}) = \frac{1}{|\text{Obs}|} \sum_{t=1}^{|\text{Obs}|} \frac{1}{\|\text{obs}_t - \hat{\text{obs}}_t\| + 1}$$

(5)
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).

Fig. 7. Time series reconstruction from HMM classification. Signal \( \hat{\text{Obs}} \) is reconstructed from original observation \( \text{Obs} \) 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 reconstructed 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 \( C_{O2} \) 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).
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).**

<table>
<thead>
<tr>
<th>State</th>
<th>( s_1 )</th>
<th>( s_2 )</th>
<th>( s_3 )</th>
<th>( s_4 )</th>
<th>( s_5 )</th>
<th>( s_6 )</th>
<th>( s_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>red</td>
<td>green</td>
<td>blue</td>
<td>cyan</td>
<td>pink</td>
<td>yellow</td>
<td>grey</td>
</tr>
<tr>
<td>( C_{NI1} )</td>
<td>-0.14</td>
<td>-0.24</td>
<td><strong>0.54</strong></td>
<td>-0.24</td>
<td>0.12</td>
<td>0.02</td>
<td>-0.03</td>
</tr>
<tr>
<td>( C_{OZ1} )</td>
<td><strong>0.64</strong></td>
<td>-0.16</td>
<td>-0.03</td>
<td>-0.38</td>
<td>0.05</td>
<td>-0.17</td>
<td>0.01</td>
</tr>
<tr>
<td>( C_{PO1} )</td>
<td>-0.08</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.03</td>
<td><strong>0.57</strong></td>
</tr>
<tr>
<td>( C_{SI1} )</td>
<td>-0.11</td>
<td>-0.16</td>
<td>0.20</td>
<td>-0.25</td>
<td>0.15</td>
<td>0.04</td>
<td><strong>0.47</strong></td>
</tr>
<tr>
<td>( CSAL1 )</td>
<td>0.12</td>
<td>0.13</td>
<td><strong>-0.36</strong></td>
<td>0.10</td>
<td>-0.25</td>
<td><strong>0.42</strong></td>
<td>0.00</td>
</tr>
<tr>
<td>( E_{LU1} )</td>
<td>-0.08</td>
<td><strong>0.73</strong></td>
<td>-0.21</td>
<td>-0.24</td>
<td>-0.06</td>
<td>-0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>( E_{TU1} )</td>
<td>-0.11</td>
<td>-0.14</td>
<td>-0.01</td>
<td>-0.22</td>
<td><strong>0.76</strong></td>
<td>-0.03</td>
<td>-0.05</td>
</tr>
<tr>
<td>( E_{VVR} )</td>
<td>-0.25</td>
<td>-0.03</td>
<td>0.16</td>
<td>-0.05</td>
<td><strong>0.31</strong></td>
<td>-0.06</td>
<td>-0.03</td>
</tr>
<tr>
<td>( ETCO1 )</td>
<td><strong>-0.50</strong></td>
<td><strong>0.32</strong></td>
<td><strong>-0.37</strong></td>
<td><strong>0.56</strong></td>
<td>-0.14</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>( XMAHH )</td>
<td>-0.00</td>
<td>0.03</td>
<td>0.02</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

### 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 (*Phaeocystis 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).
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


