INTERREG IVA 2 Mers Seas Zeeën Crossborder Cooperation Programme 2007-2013
Part-financed by the European Regional Development Fund (**ERDF**)

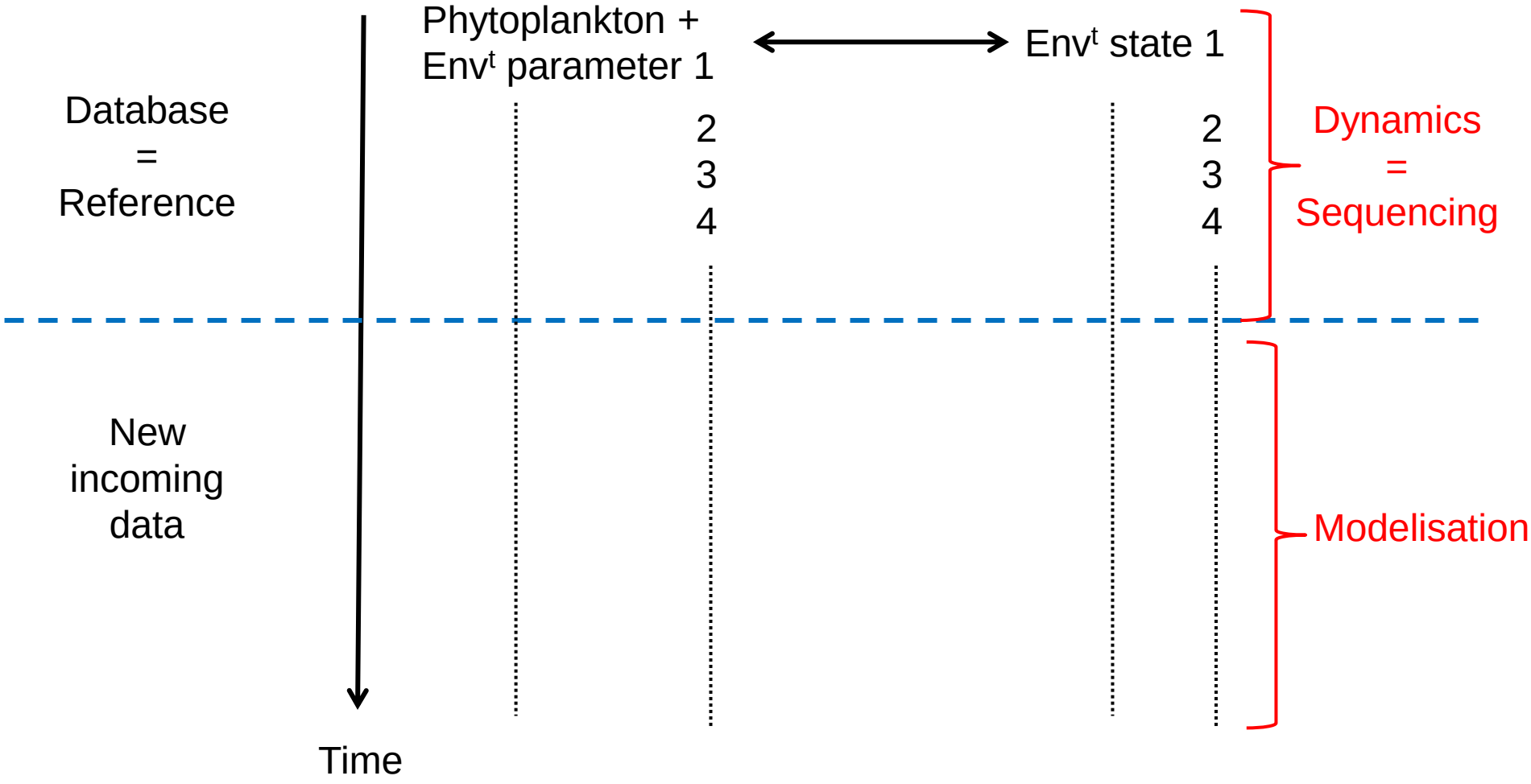
Unsupervised Hidden Markov Model building for high frequency data

Kévin Rousseeuw[1,2], Alain Lefebvre[1], Emilie Poisson Caillault[2], Denis Hamad[2]

[1] IFREMER Centre Manche - Mer du Nord, BP 699, FR-62321 BOULOGNE-SUR-MER Cedex
Mail: Firstname.name@ifremer.fr
[2] LISIC - Université Lille Nord de France - ULCO, BP 719, FR-62228 CALAIS
Mail: Firstname.name@lisic.univ-littoral.fr

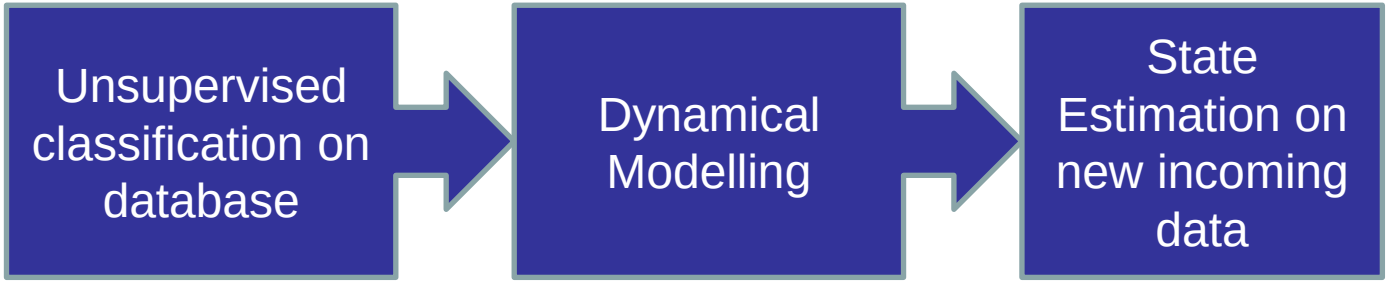
Main objective

To understand the dynamics and determinism of phytoplankton blooms



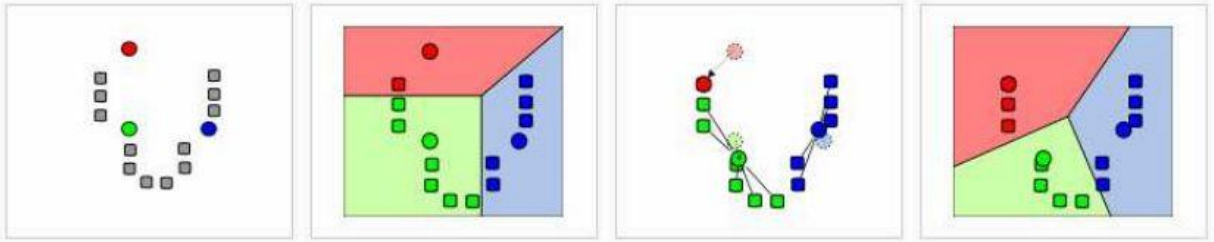
Methodology

- Detection of different environmental states
  - Unsupervised classification (any *a priori* knowledge on the data)
  - The number of states is automatically computed or set by an expert
- Comprehension of environmental states dynamics
  - Dynamical modelling



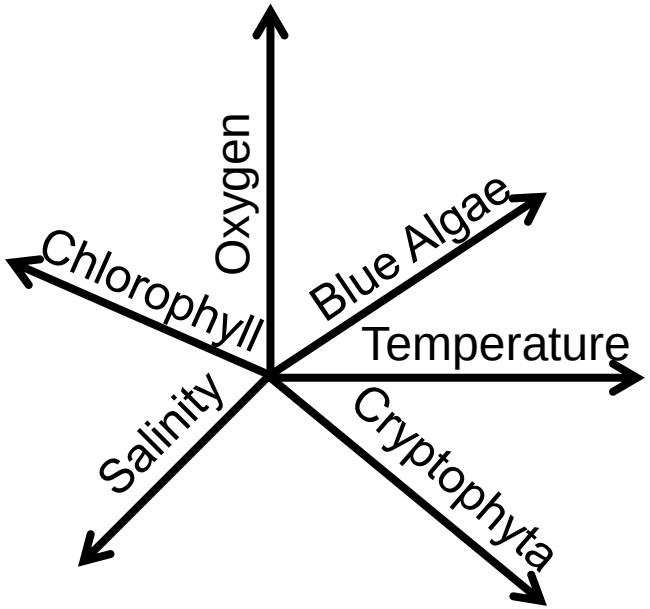
Environmental states search

- Unsupervised Classification
  - K-means method (Hartigan-Wong 1979) or Kernel K-means



1 2 3 4

High frequency (1') and
Multi parameter (9)



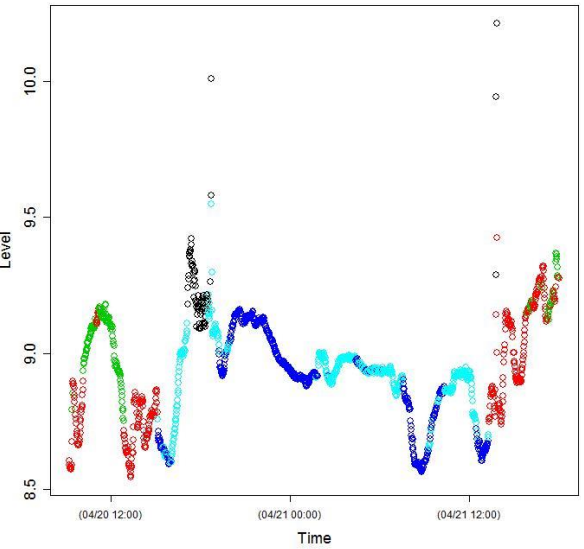
Step 1: To initialize K centers
Step 2: To associate data to the nearest center
Step 3: To calculate the new center for each group
Step 4: To associate data to the nearest center, if cluster centers do not move so break, or else return step 3

Temporality not

taken into account

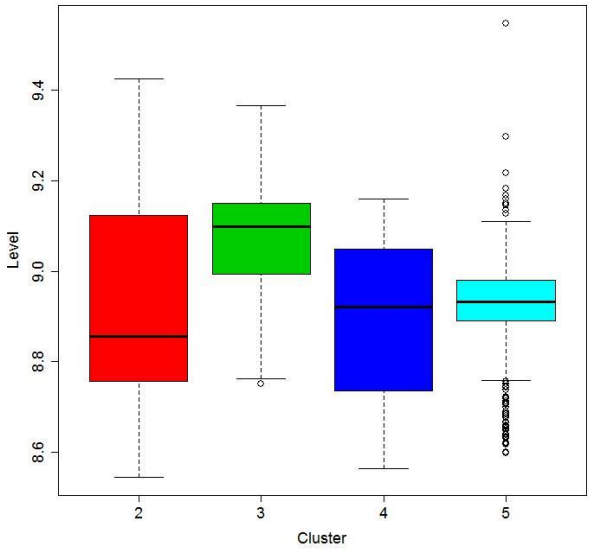
States characterization and hierarchical organization of the key parameters by cluster/state (1/2)

Temperature

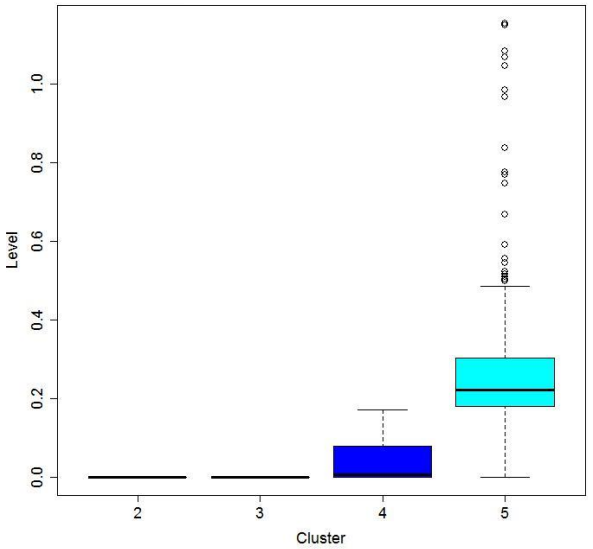


State color on the temperature signal

Temperature

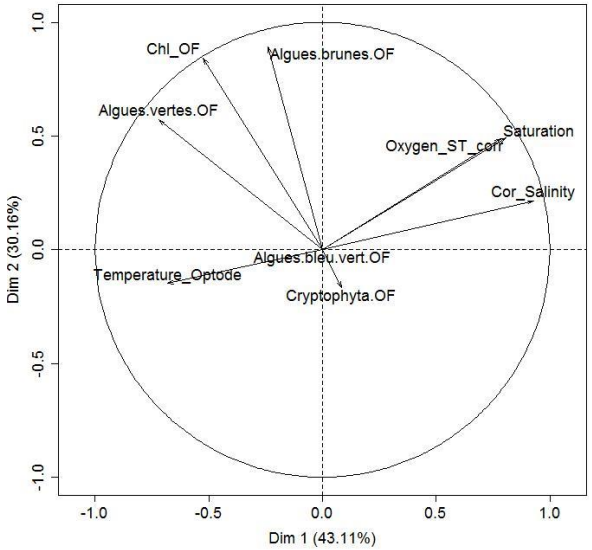


Blue-green Algae

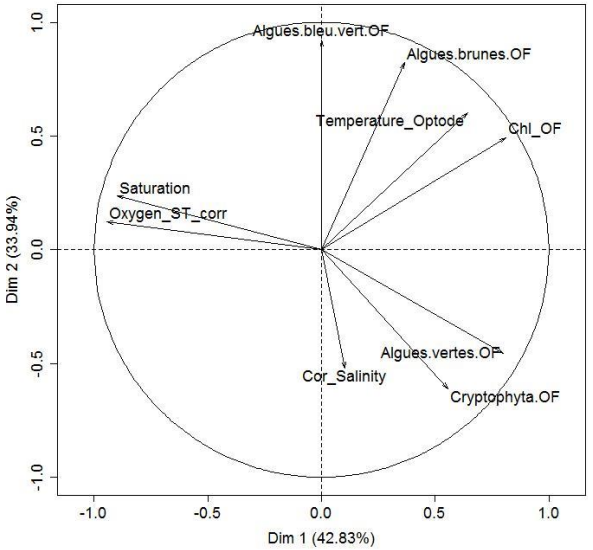


To understand the hierarchical organization with the boxplot analysis (summary statistics) and the Principal Component Analysis

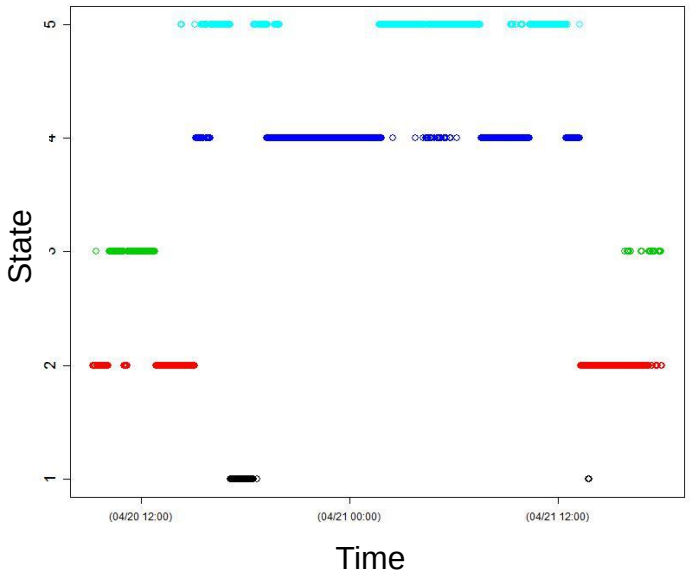
PCA for State 2



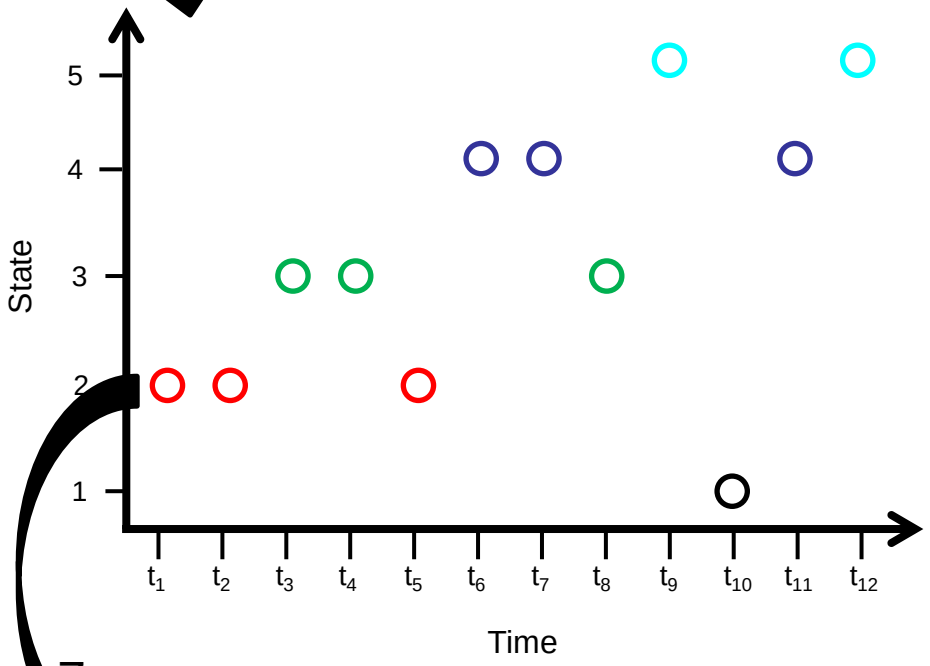
PCA for State 5



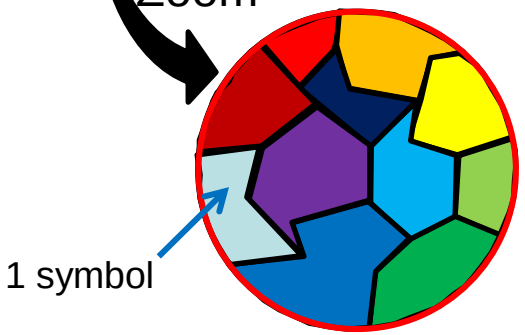
States characterization and hierarchical organization of the key parameters by cluster/state (2/2)



Zoom



Zoom

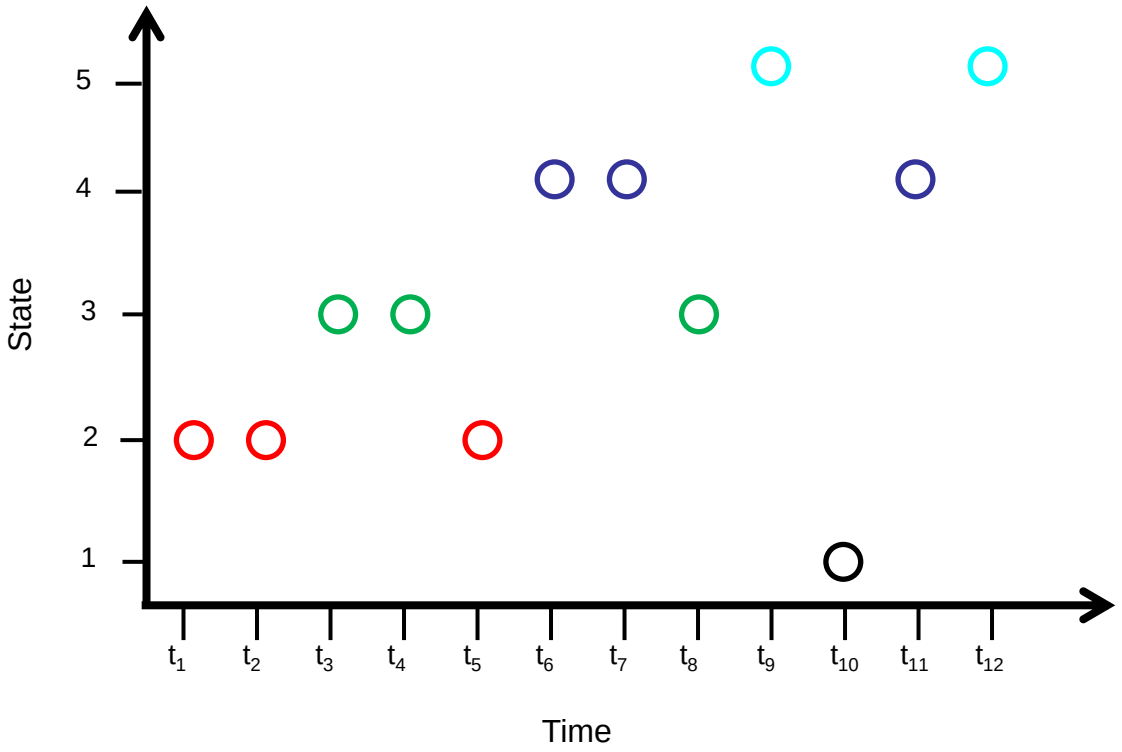


Temporal states visualisation = from this sequencing, you can analyse dynamics of the studied system and its determinism

Every state is characterized by a combination of groups/symbols = codebook (set by user)

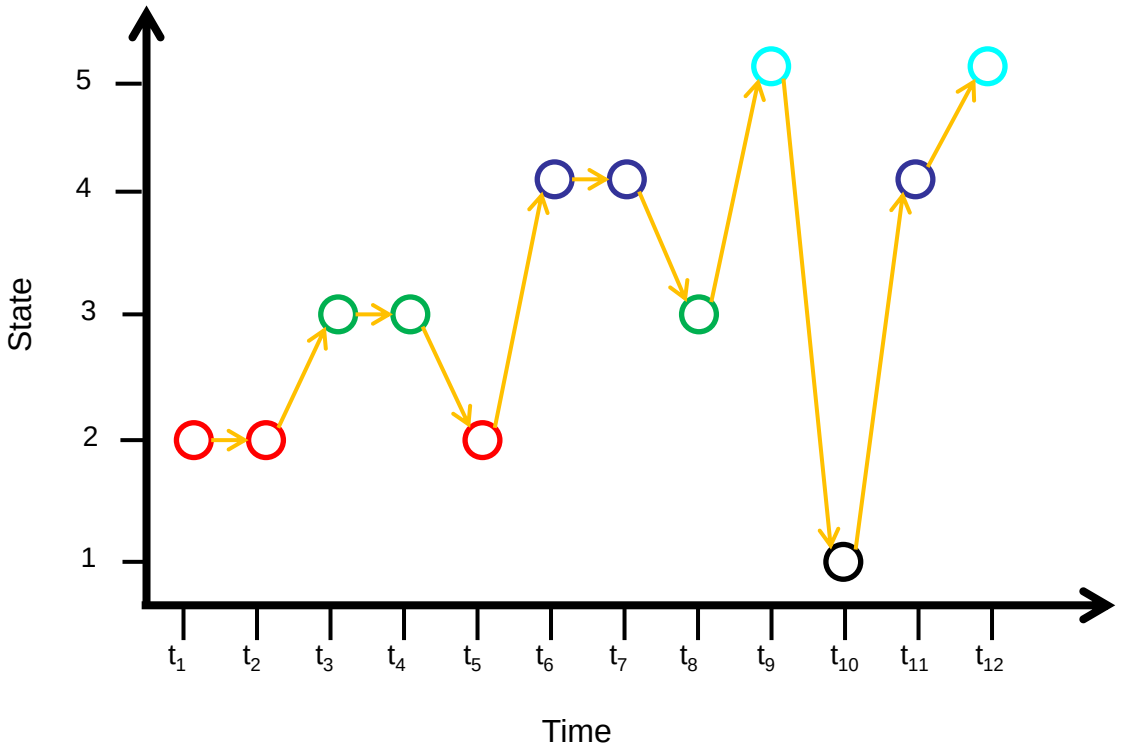
Environmental states dynamics

- Hidden Markov Model



Environmental states dynamics

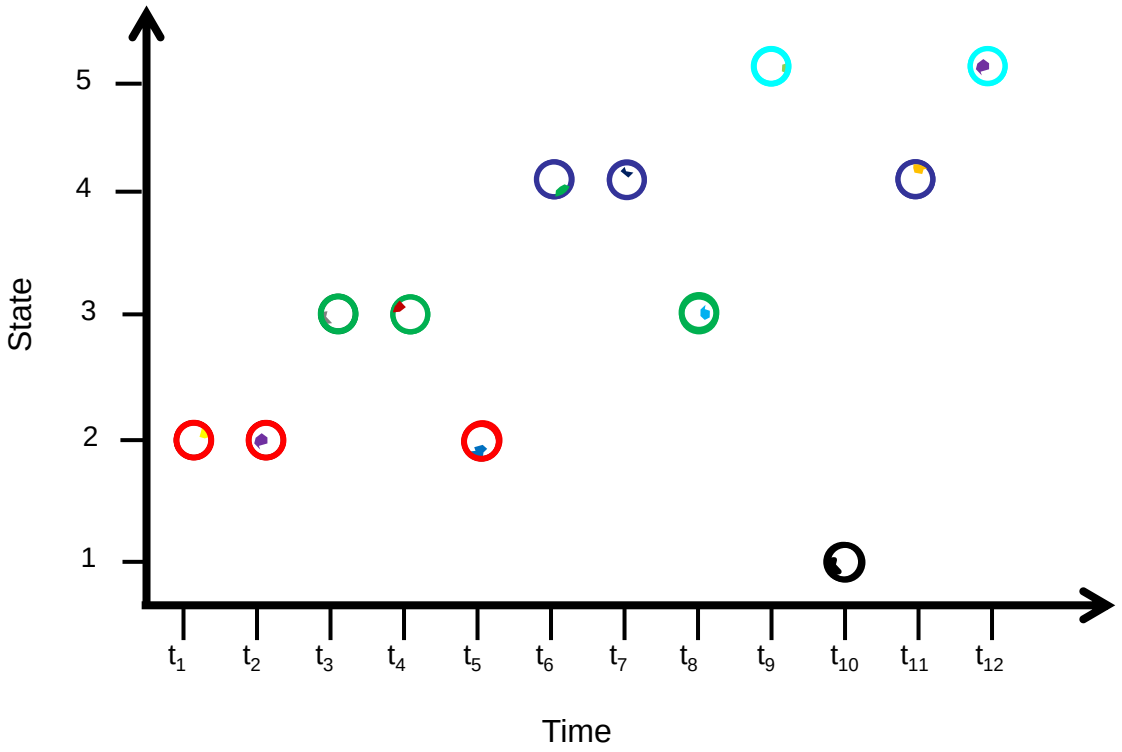
- Hidden Markov Model



Calculate the transition matrix = probability to move from a state to another one

Environmental states dynamics

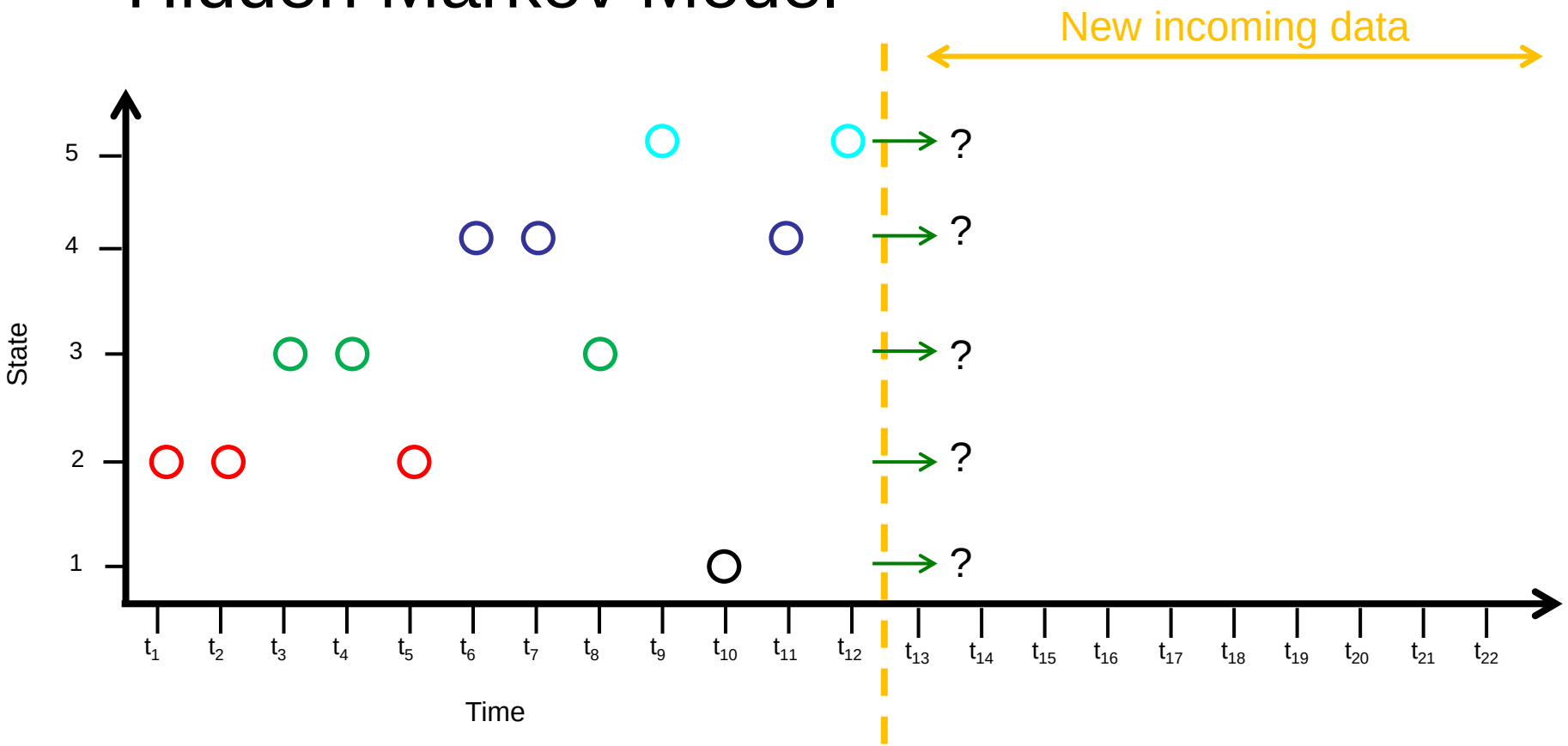
- Hidden Markov Model



Calculate the emission matrix = probability to be both in a state and in a group/symbol

Environmental states dynamics

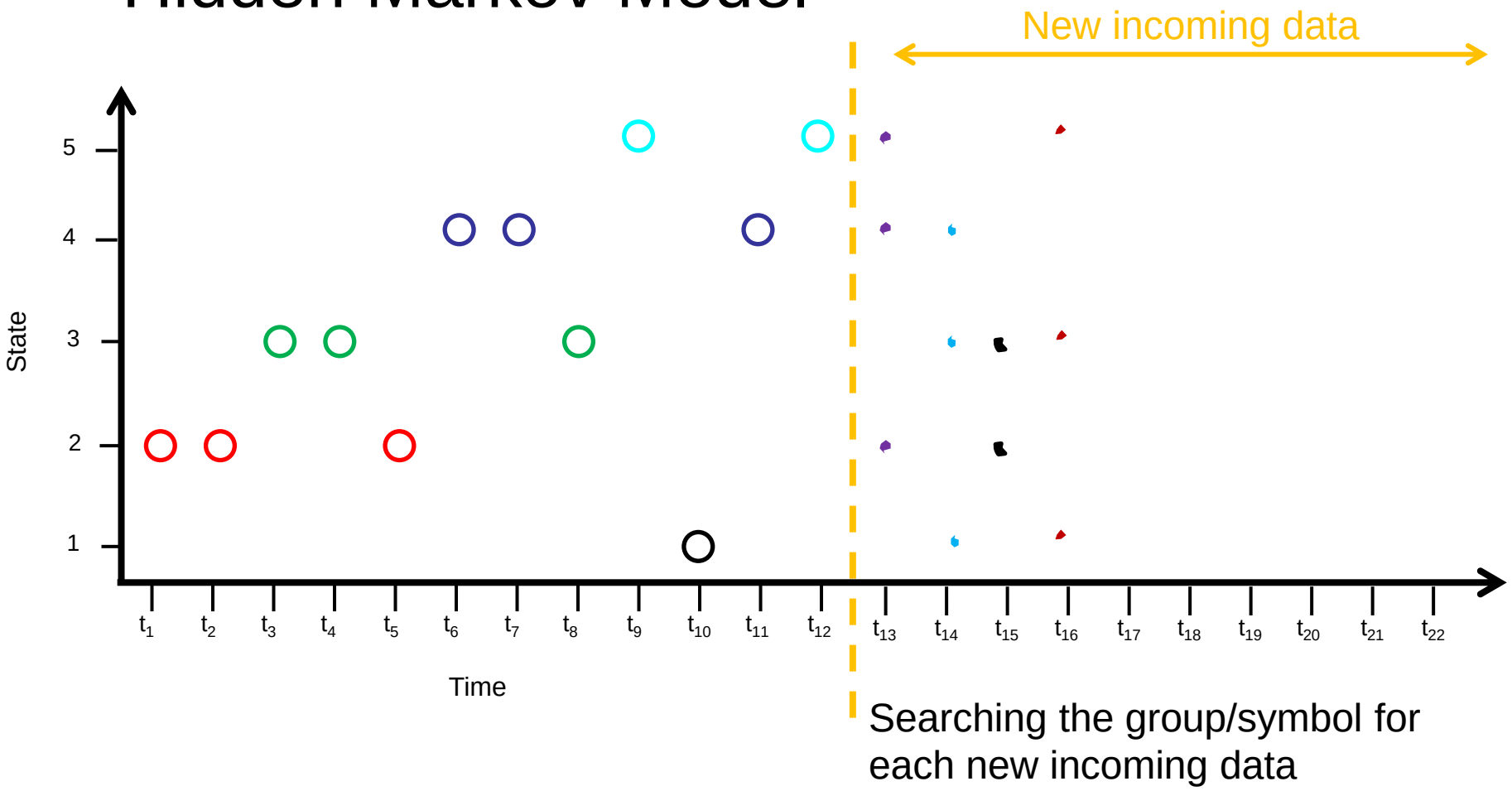
- Hidden Markov Model



→ ? Initial vector of distribution = probability to begin with a given state

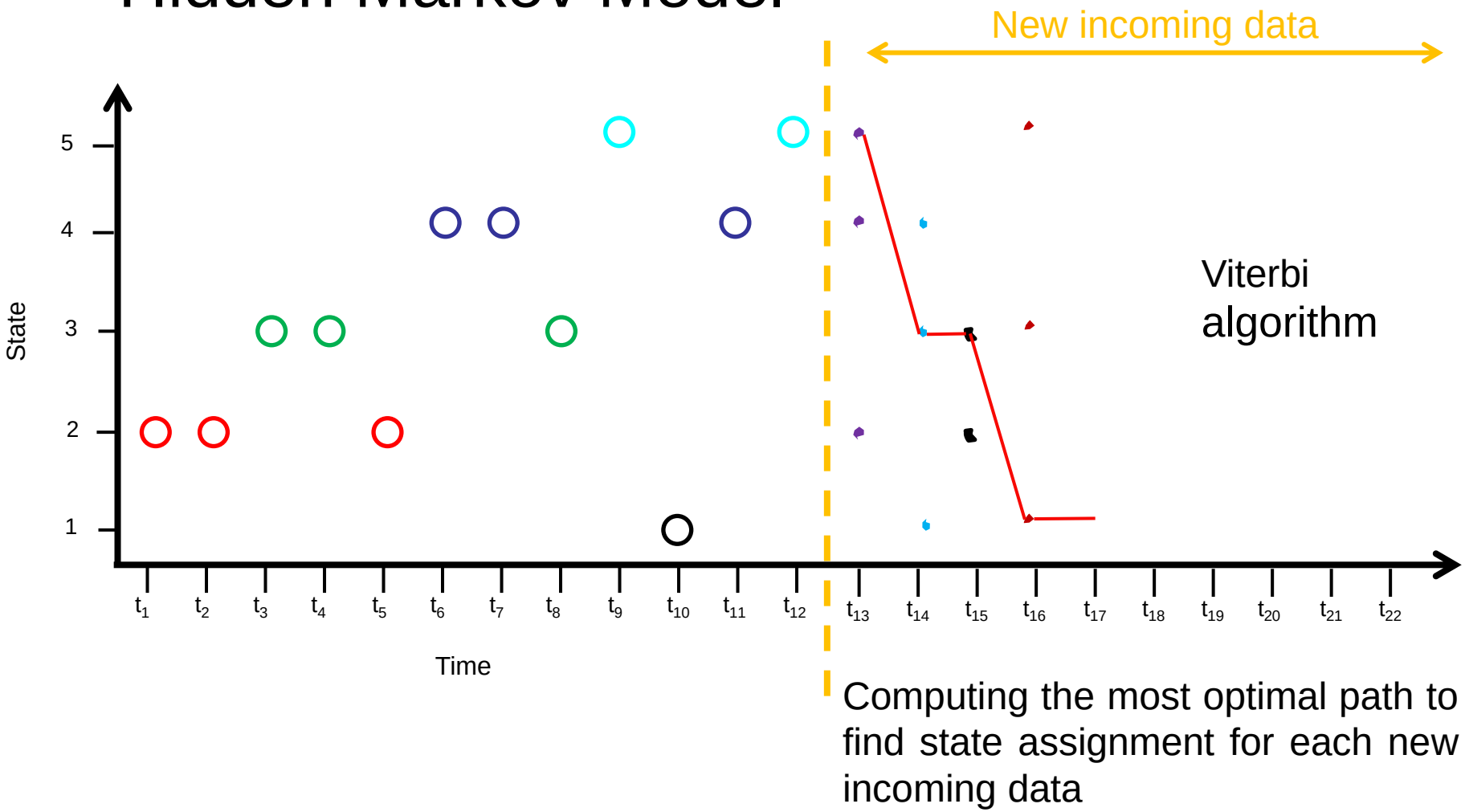
Environmental states dynamics

- Hidden Markov Model

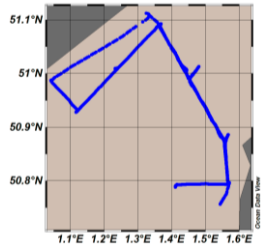


Environmental states dynamics

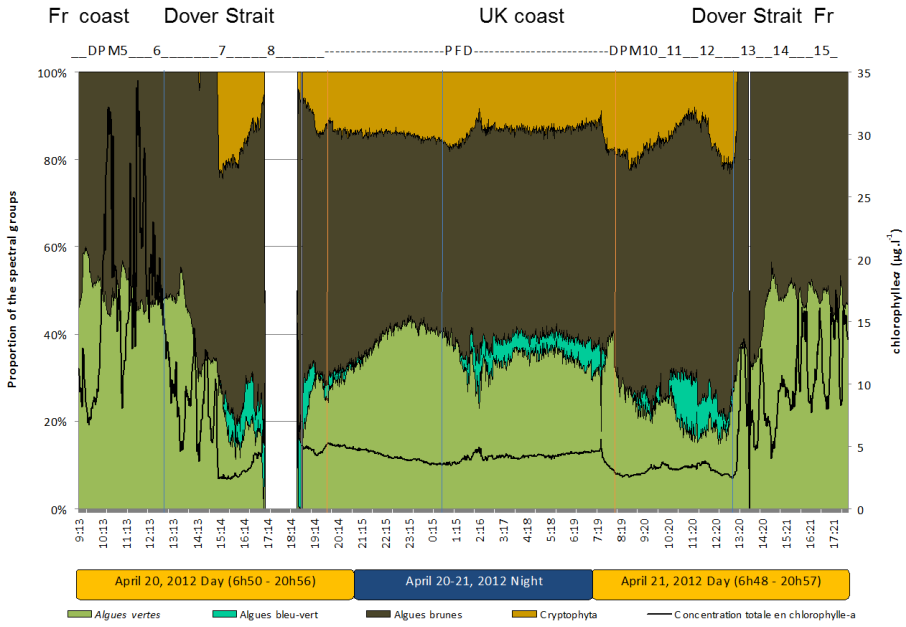
- Hidden Markov Model

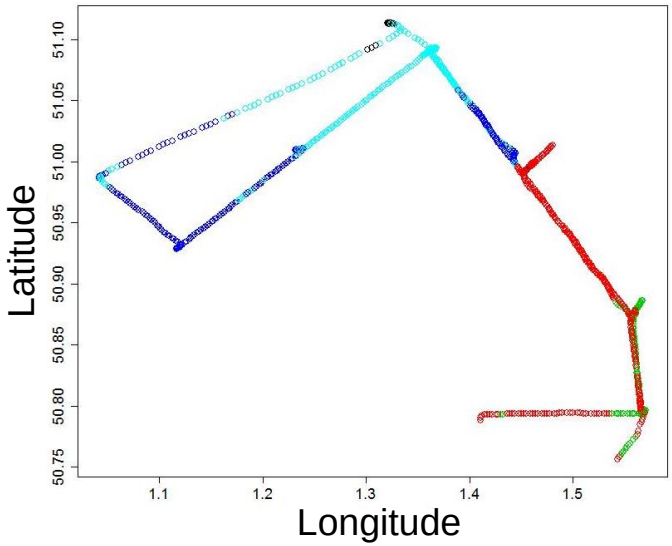


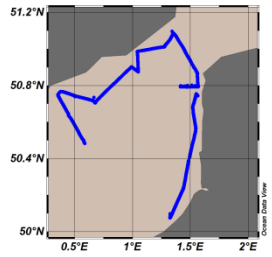
Results



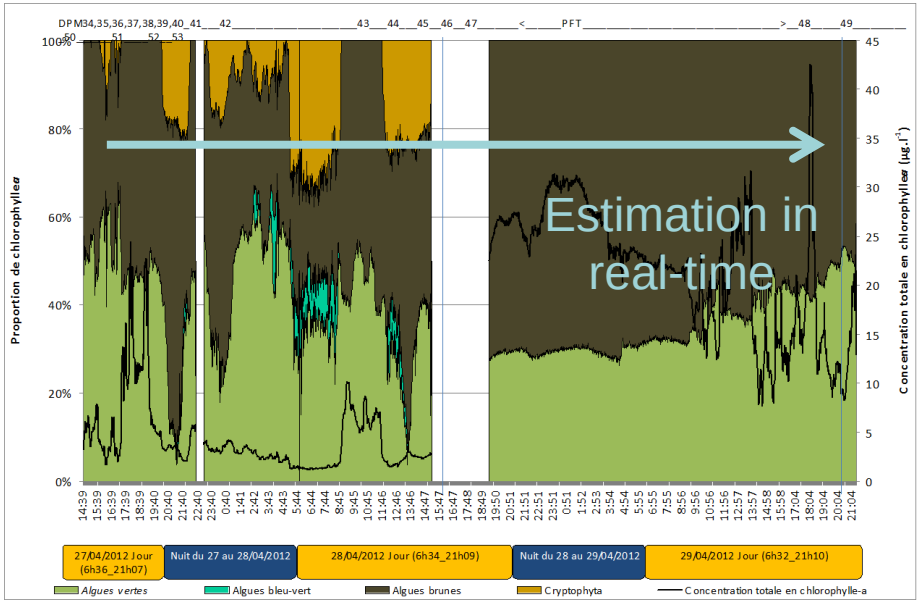
Learnt Leg1

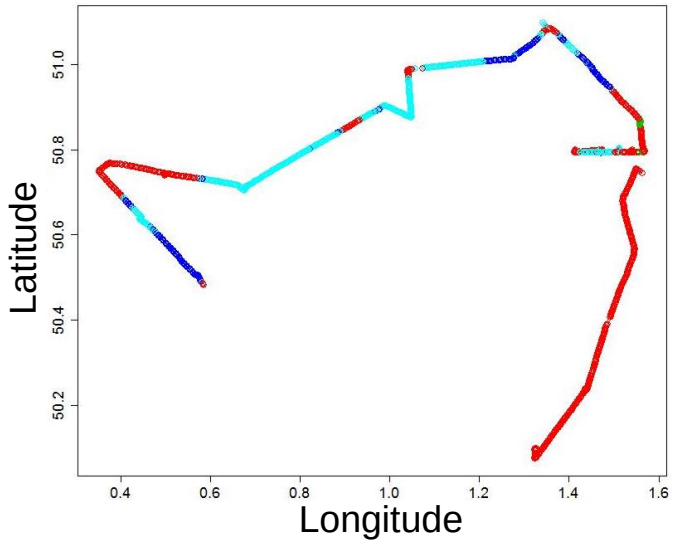




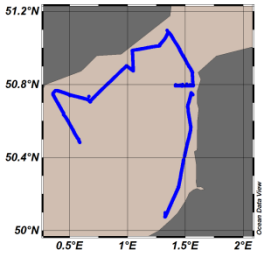


Estimated Leg2

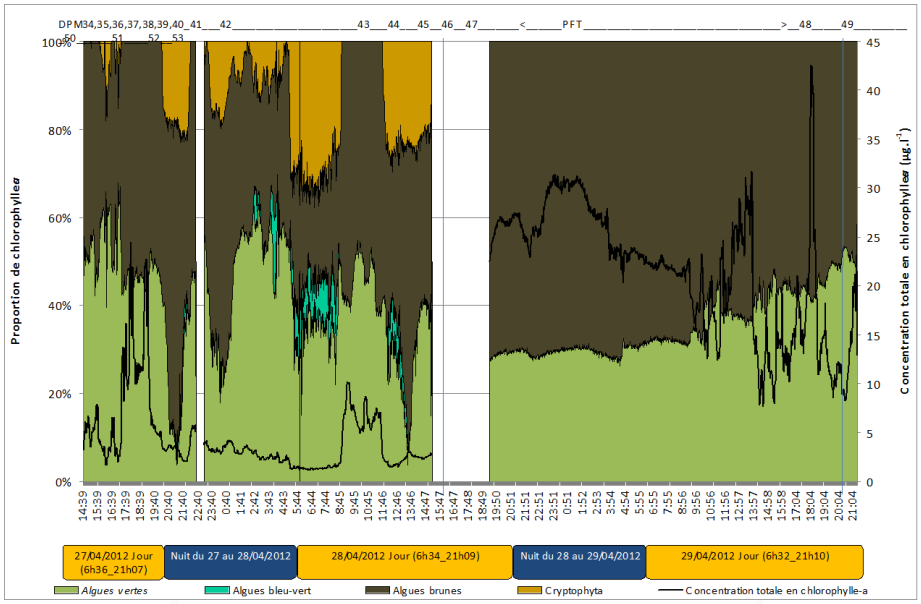


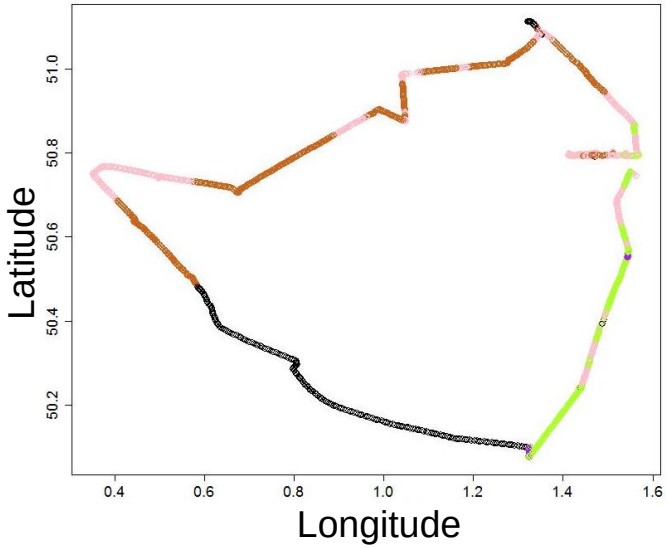


Results

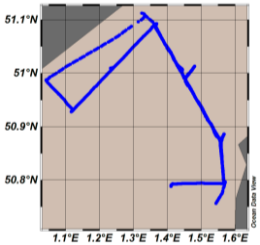


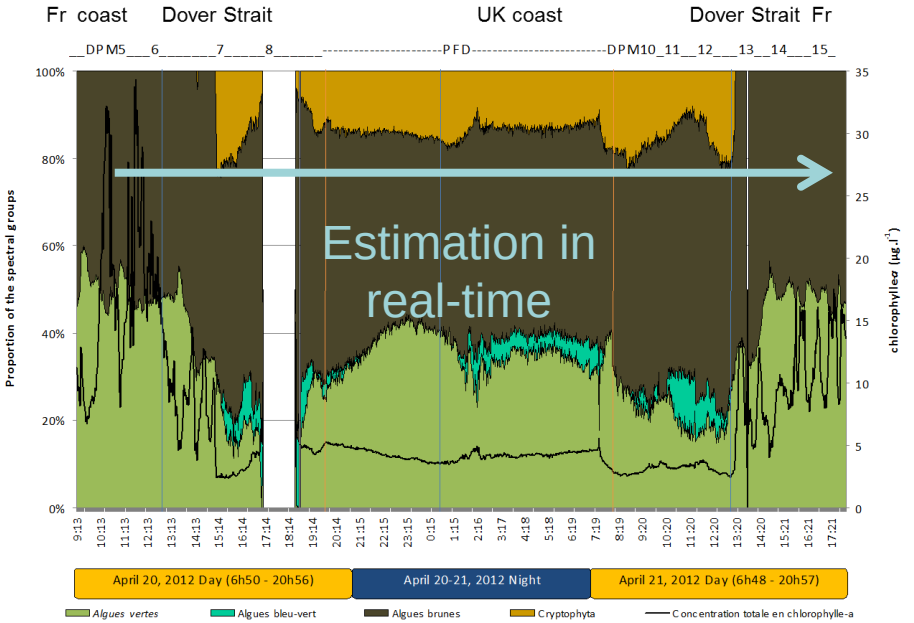
Learnt Leg2

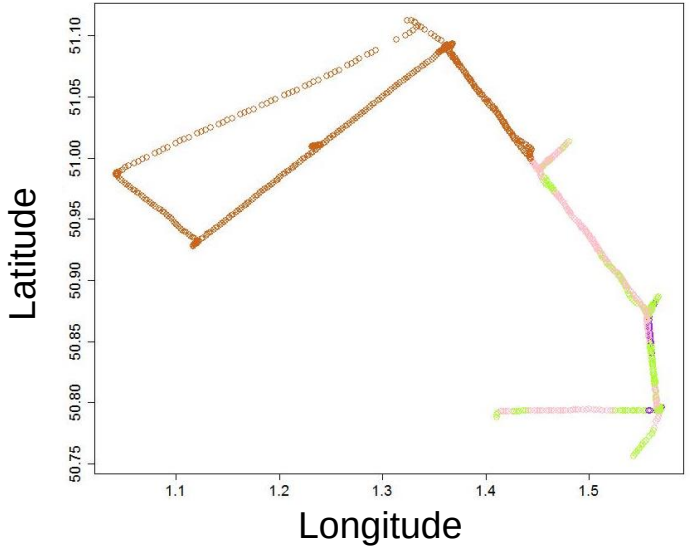




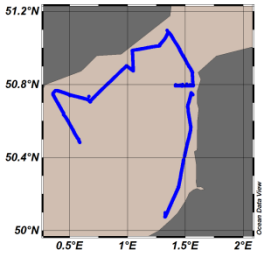
Estimated Leg1

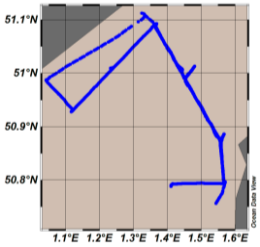




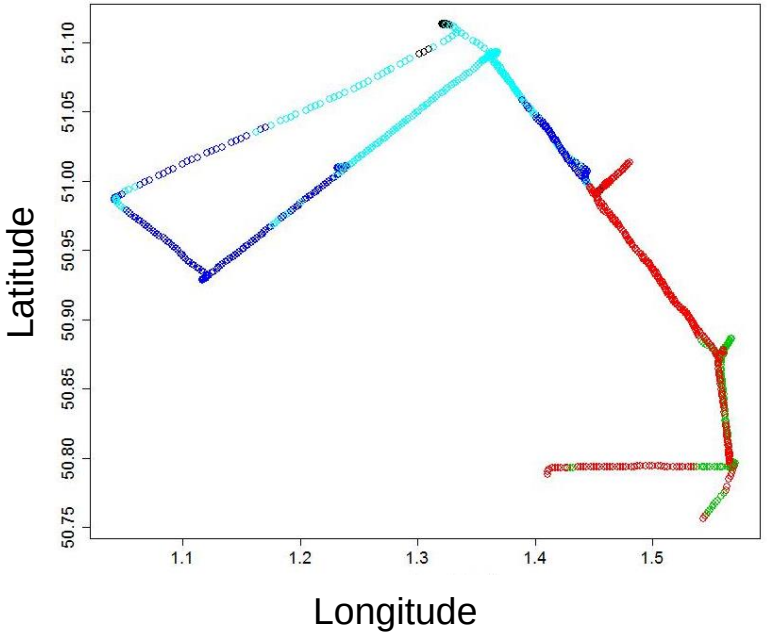


Results

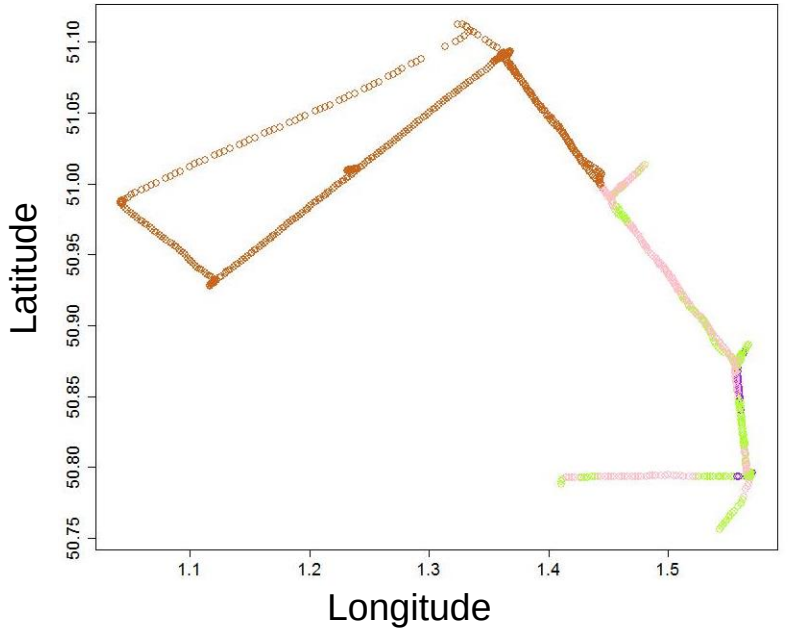




Leg1 learning



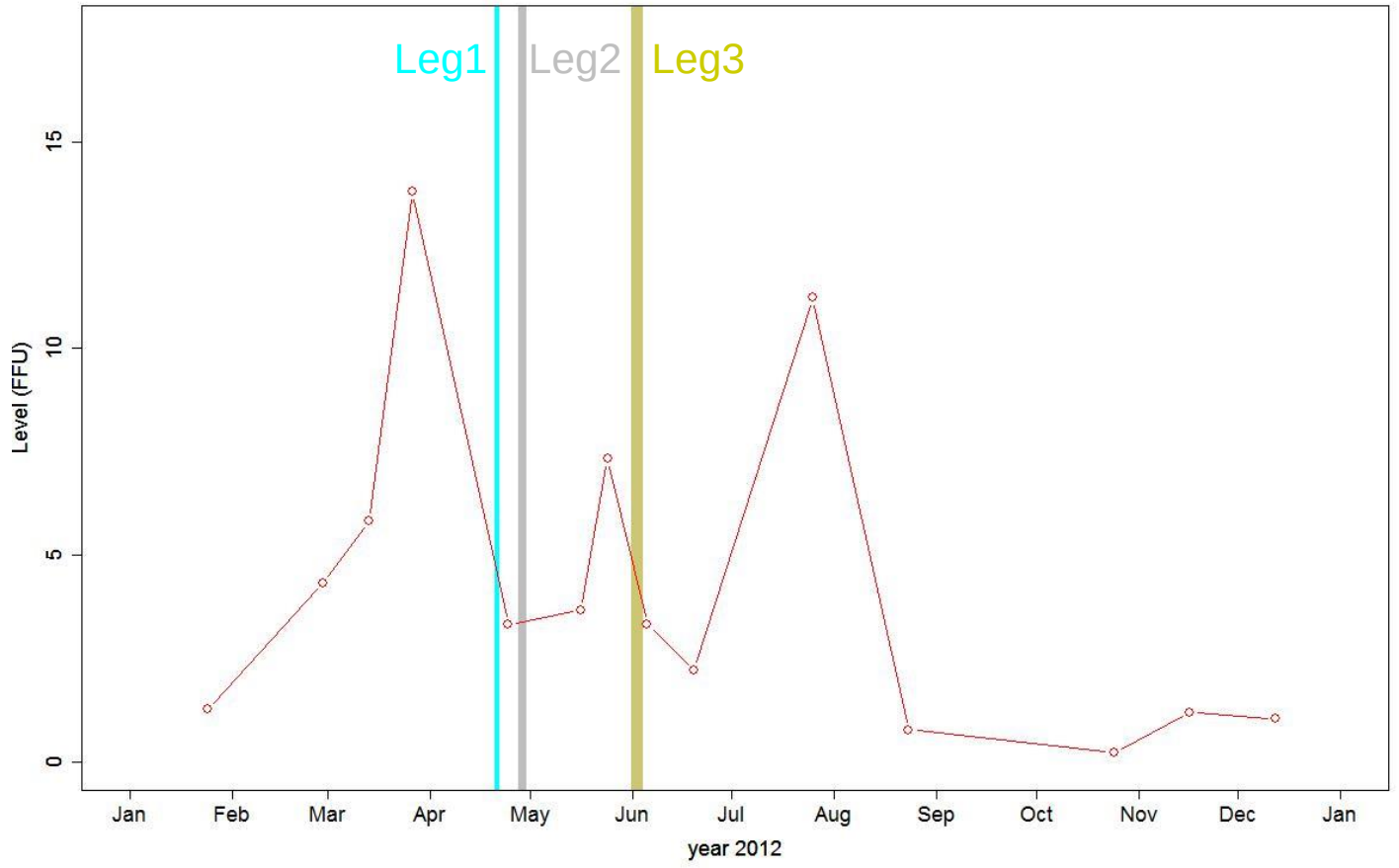
Leg1 estimating



74% of classified data in same or distinct states (Rand Index)

Results: why not 100%?

Change in the Chlorophyll-*a* concentration in the coastal area off Boulogne-sur-Mer in 2012 (SRN monitoring network)



Legs 1, 2 and 3 are not sampled in the same condition so learning reference does not integrate all variability (seasonal – inter-annual).

There will always be missing data.

Conclusions - Perspectives

- **Conventional approach**: build partition of time series and then build HMM on each cluster by learning data
- **Our approach**: build one HMM with unsupervised classifier for symbol and state (no knowledge)

- This monitoring system allows to:
  - Model dynamics of one time series by an HMM built with unsupervised classifier
  - Understand blooms dynamics from environmental states sequencing
  - Forecast Harmful Algal Bloom (early warming system to help environmental management)
  - Adapt sampling strategy in real time during scientific cruises

- Perspectives:
  - The automation of HMM algorithm building (number of states and symbols)
  - Management of missing data and metrology issues (sensors shift or offset)

INTERREG IVA 2 Mers Seas Zeeën Crossborder Cooperation Programme 2007-2013
Part-financed by the European Regional Development Fund (**ERDF**)

Thank you

Any questions do you have?

This work was supported by IFREMER and the Artois-Picardie Water Agency with a PhD grant. This project is partially financially supported by the DYMAPHY Interreg IV A “2 Mers Seas and Zeeën” program.