New developments to fill the gap in high frequency data series and to integrate knowledge in Markov modeling of phytoplankton dynamics.

É. Poisson Caillault[1,2], T.T.H. Phan[1], A. Rizik[1], P. Ternynck[1], A. Bigand[1], A. Lefebvre[2].
[1] *Univ. Littoral Côte d'Opale, LISIC,* [2] *IFREMER LER Boulogne-sur-Mer* .
Email of corresponding author: emilie.poisson@univ-littoral.fr

The implementation of high-frequency measure systems produces large multivariate series with missing values. To take advantage of information at infra-day scale, numeric tools are required.
We present recent methodological developments, originally planned for Marel Carnot database and Pocket FerryBox series and extended to other observation data.
The unsupervised Hidden Markov Model approach (uHMM R-package) allows to define environmental states characteristic of a combination of physico-chemical and biological parameters and their dynamics. To improve state characterization and state prediction, semi-supervised machine learning techniques are investigated.
To deal with the drawback of missing values or intervals due to periods of sensor maintenance, failure, ..., we propose two automated imputation methods, one for monovariate series, the second for multivariate series. Available algorithms as interpolation or multiple imputation by prediction approaches (MI - Multiple imputation based Bayesian network, MICE - Multiple Imputation Chained Equations and Random Forest) are not efficient. The two proposed approaches are based on classification and dynamic time warping.
Then, in order to better understand the dynamics and the phytoplankton composition changes, we present a bloom event extraction and time characterization by a Gauss curve self-extraction approach using expectation-maximization algorithm and reconstruction criteria. The new knowledge, phenology and their taxonomic composition of each event/state is integrated in the uHMM building.

Keywords : Environmental State, Event Extraction, Semi-Supervised Clustering, Mono/multivariate Data Imputation, Phytoplankton Dynamics, Hidden Markov Models.

Nouveaux développements pour l'imputation des valeurs manquantes dans les séries HF et une modélisation markovienne semi-supervisée de la dynamique phytoplanctonique.

La mise en œuvre des systèmes de mesures automatisées à haute fréquence nécessite des développements numériques afin de pouvoir extraire de ces bases de données importantes, multiparamètres, à valeurs manquantes toute l'information motivant des mesures à fréquences infra-journalières.
Nous proposons de présenter les développements méthodologiques récents, initialement prévus pour (pré)traiter les données issues du système MAREL Carnot ou d'un Pocket Ferry Box et utilisables à ce jour pour d'autres jeux de données d'observation.
L'approche hybride non supervisée de modélisation Markovienne permet de définir des états environnementaux caractéristiques de la combinaison de plusieurs paramètres physico-chimiques et biologiques et de la dynamique de ces états. Afin d'améliorer la définition de ces états et d'envisager une meilleure prédiction de leurs occurrences, des approches semi-supervisées sont en cours de développement.
Pour pallier les valeurs manquantes liées aux périodes de dysfonctionnements ou maintenances des capteurs, …, nous proposons deux méthodes de complétion, l'une pour des séries mono-variable, la seconde pour des séries multivariables. Les algorithmes usuels tels l'interpolation, l'imputation multiple par prédiction (MI basé réseau bayésien, MICE basé Monte Carlo, Random Forest) ne sont pas efficaces. Les deux approches proposées utilisent la notion de classification et d'appariement élastique.
Ensuite, afin de mieux comprendre la dynamique et les changements de composition phytoplanctonique, nous présenterons une méthode d'identification/classification des efflorescences par extraction de courbes de Gauss selon une approche de maximum de vraisemblance couplé à des critères de reconstruction. La phénologie et composition taxonomique de chaque état sont intégrées dans le modèle de Markov Caché.