ICHA 2018 à Nantes (18th international conference on harmful algae)

Combination of "machine learning" methodologies and imaging-in-flow system to detect Harmful Algae semi-automatically.

Guillaume Wacquet(1), Alain Lefebvre(2), Camille Blondel(2), Arnaud Louchart(1),
Philippe Grosjean(3), Luis Felipe Artigas(1)

(1) CNRS, Laboratory of Oceanology and Geosciences (LOG), ULCO, Wimereux, France.
(2) IFREMER, LER, Environment and Resources Laboratory, Boulogne-sur-Mer, France.
(3) Univ. Mons, Laboratory of Numerical Ecology of Aquatic Systems, Mons, Belgium.

Abstract

In recent years, improvements in data acquisition techniques have been carried out in order to sample, characterize and quantify phytoplankton communities with a special focus on potential harmful algae during oceanographic campaigns or in the frame of monitoring networks. However, these acquisition and digitization techniques, including those concerning «imaging-in-flow» systems, still generate an important quantity of data in which the presence of target events might not be detected. Indeed, as for traditional samples analysis with inverted microscope, a full manual quantification of the particles based on a simple visual inspection can be time-consuming, tedious and consequently lead to erroneous or missing identifications. For this purpose, a specific R-package, named "zooimage", was and is still being developed to allow greater automation in data analysis and classification while permitting a limited user-interaction during the process. The proposed methodology consists in combining few expert knowledge and some "machine learning" algorithms at different levels: (i) to classify particles into different groups based on the definition of a specific training set; (ii) to detect and partially validate the "most suspect" predictions which can represent until 90% of the global error; (iii) to automatically estimate the number of cells for each colonial form. Moreover, in order to orientate the automated classification and consequently to reduce the global error rate, an active learning process consisting in adapting the training set to the phytoplankton communities generally encountered in the studied area, was developed. For this, some samples were chosen at regular time intervals, manually classified and used as "contextual data". Thanks to this technique, the initial recognition rate can be significantly improved, and even reach 95% when 20-25% of the particles are manually validated. These different semi-automated tools were applied on the *in vivo* image dataset acquired with the FlowCam system during the September-October CAMANOC 2014 (IFREMER) cruise in the English Channel, in order to evaluate their operational ability to automatically monitor the diversity of samples for the microphytoplankton, and especially to detect, track and count the most frequent potentially harmful algae found in this area at that period, like species belonging to the genera *Pseudo-nitzschia*, *Dinophysis, Prorocentrum* and *Phaeocystis*. A spatial distribution of these target groups was computed and could allow to highlight different sub-regions in the English Channel during the late summer-fall transition.

Keywords

Machine Learning; User-interaction; Semi-automated classification; R-package; Imaging-in-flow; English Channel; HABs.