Incremental clustering

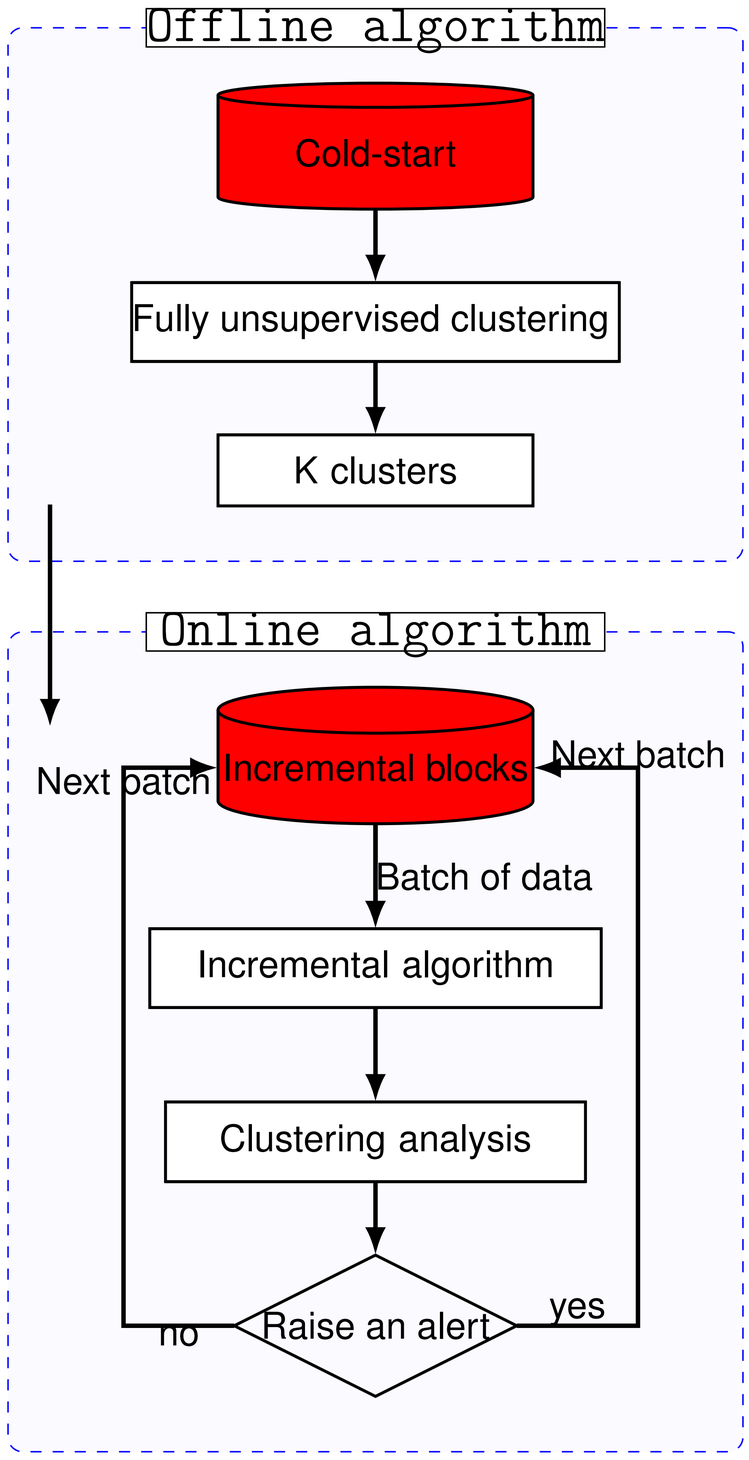
Boubacar SOW, Alain LEFEBVRE, Émilie POISSON-CAILLAULT,
Univ. Littoral Côte d'Opale, UR 4491 - LISIC - Laboratoire d'Informatique Signal et Image de la Côte d'Opale, F- 62228 Calais, France
Ifremer, Laboratoire Environnement et Ressources, F- 62321 Boulogne-sur-Mer, France

Introduction

We propose an approach to effectively monitor and manage phytoplankton populations in aquatic ecosystems using incremental clustering by developping a system that uses incremental clustering to raise alerts when unknown species are detected allowing for dynamic and effective management of phytoplankton populations.

General approach

- Two-stage approach: offline and online clustering.
- Offline: Cold-start batch to initialize clusters centers and find initial clusters.
- Online: apply incremental clustering algorithm to incoming data in batches.
- Online: Analyze clusters after each batch to to decide if an alert should be issued



Incremental clustering process

Labelling

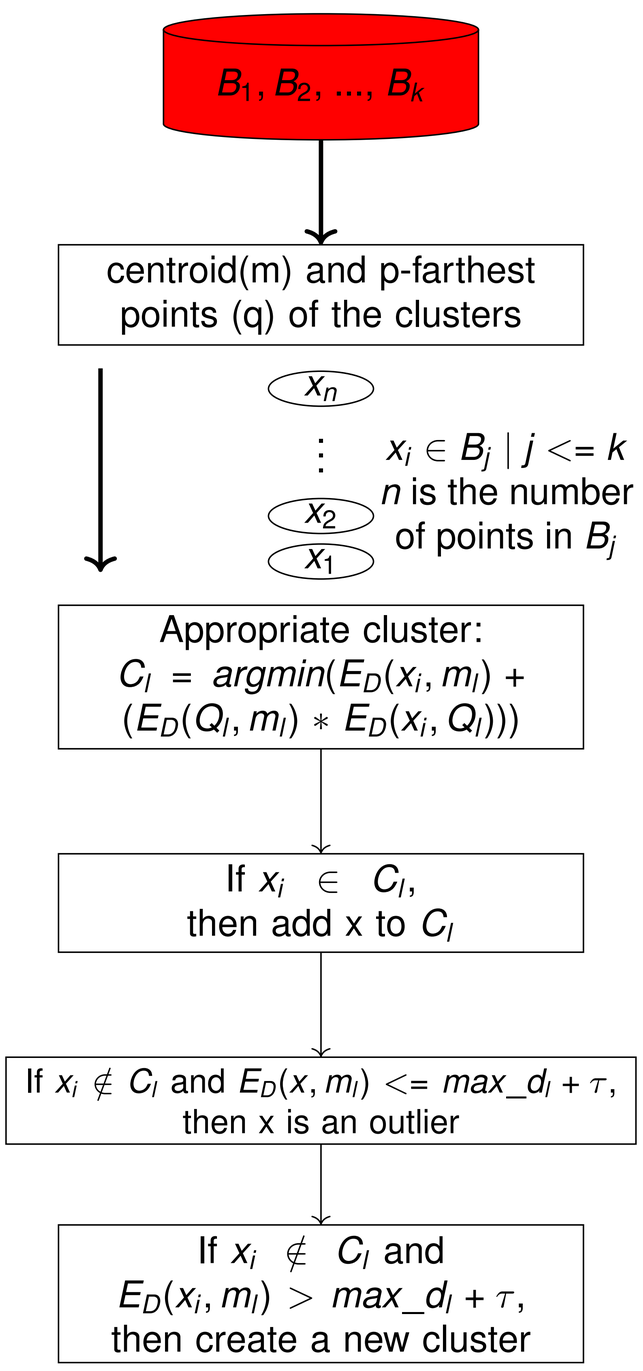
Labelling of the datasets (Dyphyma Leg 1 and Leg 2) to allow for clusters evaluation, see MEPS:

- By composition of the algaes
- By the number of algae types

We use the Adjusted Rand Index to evaluate the clustering algorithms

Incremental Clustering

- Distance based clustering algorithm



- Appropriate cluster: usage of the Euclidian distance, Q_l is the closest farthest point to x for the cluster l
- max_d_l is the distance from the centroid of C_l to the farthest point in the cluster C_l
- If the number of outliers exceed the threshold τ, apply an offline algorithm to cluster the outliers
- Merge close clusters after each batch.

Results

The results have been obtained on the Leg 2 dataset, with a merging threshold of 1 and a distance threshold of 2 for the incremental clustering algorithm.

- The best adjusted rand index score is 0.48
- The ARI is competitive compared to some implemented existing algorithms

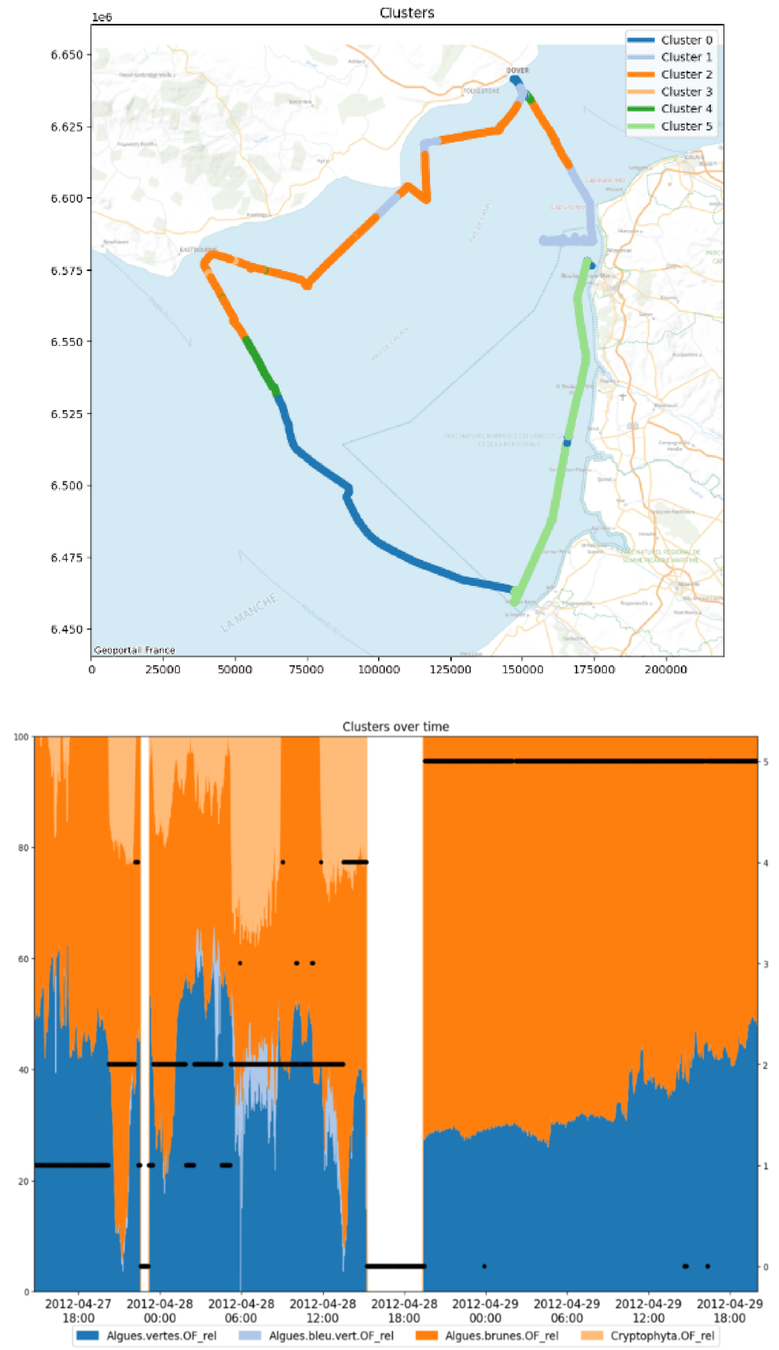
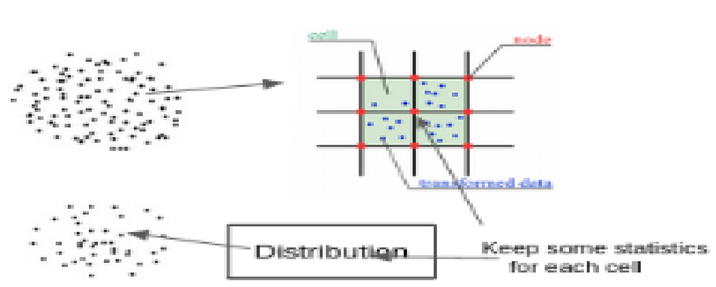


Fig. 1: Dyphyma Leg 2 clusters with the proposed approach

On-going research

- Challenge: not the best idea to always keep all already-clustered data points in memory.
- Strategy: vectorize clusters by constructing a grid for each cluster and keep some statistics for each cell
- Use this to approximate the original data whenever an offline algorithm is needed



- we choose to reconstruct a number of points that represent well the cluster
- The reconstructed data has the same mean for each cell, we can also choose to have the same covariance.

References

[1] Lefebvre Alain, Poisson-Caillault Emilie: MEPS, High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering, 2019, https://doi.org/10.3354/meps12781

[2] A. L. Felipe, "Dyphyma i et ii cruise, rv côtes de la manche," 2012, https://doi.org/10.17600/12480030

Partners