Mixed Deep Gaussian Mixture Model: A clustering model for mixed datasets
Published in *Advances in Data Analysis and Classification* (2021)

Coauthors: Robin Fuchs, Denys Pommeret, Cinzia Viroli

Abstract: Clustering mixed data presents numerous challenges inherent to the very heterogeneous nature of the variables. A clustering algorithm should be able, despite this heterogeneity, to extract discriminant pieces of information from the variables in order to design groups.
In this work we introduce a multilayer architecture model-based clustering method called Mixed Deep Gaussian Mixture Model (MDGMM) that can be viewed as an automatic way to merge the clustering performed separately on continuous and non-continuous data. This architecture is flexible and can be adapted to mixed as well as to continuous or non-continuous data. In this sense, we generalize Generalized Linear Latent Variable Models and Deep Gaussian Mixture Models. We also design a new initialization strategy and a data-driven method that selects the best specification of the model and the optimal number of clusters for a given dataset.
Besides, our model provides continuous low-dimensional representations of the data which can be a useful tool to visualize mixed datasets. Finally, we validate the performance of our approach comparing its results with state-of-the-art mixed data clustering models over several commonly used datasets.

MIAMI: MIxed data Augmentation MIxture
Submitted at the ICML conference 2022

Coauthors: Robin Fuchs, Denys Pommeret, Samuel Stocksieker

Performing data augmentation for mixed datasets remains an open challenge. We propose an adaptation of the Mixed Deep Gaussian Mixture Models (MDGMM) to generate such complex data. The MDGMM explicitly handles the different data types and learns a continuous latent representation of the data that captures their dependence structure and can be exploited to conduct data augmentation. We test the ability of our method to simulate crossings of variables that were rarely observed or unobserved during training. The performances are compared with recent competitors relying on Generative Adversarial Networks, Random Forest, Classification And Regression Trees, or Bayesian network on the UCI Adult dataset.

Automatic recognition of flow cytometric phytoplankton functional groups using Convolutional Neural Networks
Accepted in Limnology & Oceanography: Methods (2022)

Coauthors: Robin Fuchs, Melilotus Thyssen, Véronique Creach, Mathilde Dugenne, Lloyd Izard, Marie Latimier, Arnaud Louchart, Pierre Marrec, Machteld Rijkeboer, Gérald Grégori, Denys Pommeret

Abstract: The variability of phytoplankton distribution has been unraveled by high-frequency measurements. Such a resolution can be approached by automated pulse-shape recording flow cytometry (AFCM) operating at hourly sampling resolution. AFCM records morphological and physiological traits as single-cell optical pulse shapes that can be used to classify cells into Phytoplankton Functional Groups (PFG).
However, the associated manual post-processing of the data coupled with the increasing size and number of datasets is time-consuming and error-prone.
Machine learning models are increasingly used to run automatic classification. Yet, most of the existing methods either present a long training process, need to manually design features from the raw optical pulse shapes, or are dedicated to images only. In this study, we present a Convolutional Neural Network (CNN) to classify several PFGs using AFCM pulse shapes.
The uncertainties of manual classification were first estimated by comparing experts' recognition of six PFGs.
Consensual particles from the manual PFG classification were used to train and validate the CNN.
The CNN obtained competitive performances compared to other models used in the literature and remained robust across several sampling areas, and instrumental hardware and settings.
Finally, we assessed the ability of this classifier to predict phytoplankton counts at a Mediterranean coastal station and from a cruise in the South-West Indian Ocean, providing a comparison with the manual classification over three-month periods and a two-hour frequency. These promising results strengthen the near real-time observation of PFGs, especially required with the increasing use of AFCM in monitoring research programs.

Intermittent wind-driven upwelling events temporary enhance pico-nanophytoplankton biomass during stratified periods in coastal oligotrophic biomes
In Preparation for Geophysical Research Letters

Coauthors: Robin Fuchs, Chloé Caille, Vincent Rossi, Nathaniel Bensoussan, Christel Pinazo, Olivier Grosso, Melilotus Thyssen

Abstract: Pico-nanophytoplankton organisms are dominant in oligotrophic areas of the ocean thanks to competitive skills in poor nutrient areas. Their contribution to the oceanic component of the global carbon cycle is difficult to estimate because of the overlapping of constraints, mainly the small sizes of the cells, the fast-changing surroundings, and their highly adaptive growth rates. Environmental shifts due to wind-driven upwelling modify the conditions in which the microbial communities used to evolve. To study the shifts and responses capacities of the pico-nanophytoplankton, a high-frequency continuous sampling strategy was deployed in a coastal station of the bay of Marseille (NW Mediterranean Sea). The area is influenced by intermittent north-westerlies causing upwelling events that result in nutrient pulses and seawater temperature drops up to 9°C in four days. Using a CytoSense flow cytometer continuously operating at a two-hour frequency from September 2019 to November 2021, we monitored the abundances and estimated cell biomass of five phytoplankton functional groups over two complete annual cycles, and focus on events happening when the water column is stratified in late Spring and Summer. We show that despite their short duration, these wind-induced upwelling events trigger temporary increases in abundances and biomass for most phytoplankton groups that often overpass values observed during the Spring blooms. These positive reactions are immediately followed by an overall drop in abundance and biomass. Given the magnitude of the biological reactions observed in stratified periods, not including these events may significantly bias Carbon budgets.

A RUpture-BAsed detection method for the mesopeLagic active horIZon (RUBALIZ) : a crucial step towards rigourous C budget assessments
In Preparation for Limnology & Oceanography: Methods

Coauthors: Robin Fuchs, Chloé M.J. Baumas, David Nerini, Marc Garel, Frédéric A.C. Le Moigne, Christian Tamburini

Abstract: Carbon budgets are key indicators to assess the role of oceans as carbon dioxide sinks or sources. Measured carbon budgets often present a discrepancy with the carbon demand being deemed higher than the carbon supply. This underlines either major methodological issues in the budget calculations or incomplete knowledge of the mesopelagic carbon cycling with potentially large unknown carbon sources. Such carbon budgets are estimated by partitioning the ocean into homogeneous vertical depth layers. In this respect, the mesopelagic layer is known as the scene of intensive biological processes acting on carbon fluxes and is therefore of particular interest for the air-sea carbon dioxide balance. However, the determination of the vertical boundaries of the mesopelagic layer is conventionally performed using simple heuristics or thresholds and lacks robust methodology. Here, using a statistical rupture detection method applied to CTD-cast variables (fluorescence, oxygen, potential temperature, salinity, and density), we provide independent estimates of mesopelagic boundaries. We show that these boundaries are remarkably consistent with variations in mesopelagic biological fluxes. We found that the depths of the mesopelagic layer depend strongly on the ocean region considered. This contrasts with numerous studies assuming a fixed depth for all oceans, typically between 200m and 1000m. The identified layer corresponds to the most active part of the usual mesopelagic layer and we call it the “active mesopelagic layer”. Our results demonstrate that the mesopelagic carbon budget discrepancy can vary three folds depending on the boundaries chosen and provides novel grounds to reassess the mesopelagic carbon budget.

High-resolution description of a storm-induced phytoplankton bloom in the north-western Mediterranean Sea
In Preparation

Coauthors: Stéphanie Barrillon, Robin Fuchs, Anne Petrenko, Caroline Comby, Anthony Bosse, Christophe Yohia, Jean-Luc Fuda, Nagib Bhairy, Léo Berline, Frédéric Cyr, Andrea Doglioli, Gérald Grégori, Roxane Tzortzis, Francesco d'Ovidio, Melilotus Thyssen

Abstract: The FUMSECK cruise was conducted in Spring 2019 in the Ligurian Sea (north-western Mediterranean Sea), to explore fine-scale dynamics and their effect on microorganisms. During the cruise, physics and biogeochemistry measurements were performed at high resolution. After a strong storm, the data revealed a zone with particular physics characteristics, implying a striking phytoplankton reaction. Our results show that this storm physical forcing leads to a deepening of the mixed layer depth and dilution of the deep chlorophyll maximum with a resulting increase in biomass in the surface layers. We observed the short-term physical and biological reactions to this event. This observational evidence of an immediate reaction of phytoplankton to impulse physical forcing enlightens the need for high-resolution coupled physics-biology measurement.