OS21C-1590: Classifying ocean profiles with machine learning algorithms

Kimmo Tikka, Antti Westerlund, Pekka Alenius and Laura Tuomi

Finnish Meteorological Institute

Objectives

Our objective was to answer the question: "*How do we identify the (dis-)similarities between multidimensional profiles of oceanographic data?*" For this, we applied unsupervised learning methods to recognize anomalies in the Argo and glider data. We may use the results to:

1. Quality check ocean profiles data,
2. Find anomalies using several parameters,
3. Recognize dynamical changes in ocean,
4. Evaluate ocean models, and
5. For automation of the Argo and glider piloting

Introduction

Today, online databases with historical and real-time data from research cruises, automatic profilers, floats, and gliders provide oceanographers vast opportunities and challenges. There are two problems: the ship data is sparse and the new automatic techniques, especially gliders, produce huge amounts of data. In both cases, the automatic classification of profiles could help to quality control and interpret the data.

Study area and data

Our research area is the Baltic Sea, which is a shallow (mean depth 54 m),seasonally and partly permanently stratified, brackish water sea with several sub-basins. The permanent halocline lies in 60-80 m depth and the seasonal thermocline reaches 15-30 m.

FMI operates several Argo buoys since 2012 (Siiriä &al. 2018) and one Slocum glider since 2016. Here, we used data from FMI's Argo floats in the Gotland Deep and from the GROOM2013 cruise with PLOCAN's (Gran Canary) Slocum glider in the Bothnian Sea.

Figure 1: The study areas in the Baltic Sea.

Methods

We applied unsupervised machine learning methods to classify ocean profiles into similar shape clusters. We used Dynamic Time Warping (DTW) algorithms to calculate the similarity of the profiles and K-Means to cluster them. These methods are widely used in time series analysis.

Seasonal variation of Gotland Deep

Our aim was to find a classification that describes the thermal seasons in the Baltic Sea. Such a classification would be useful in quality assurance of the data and in finding outliers.

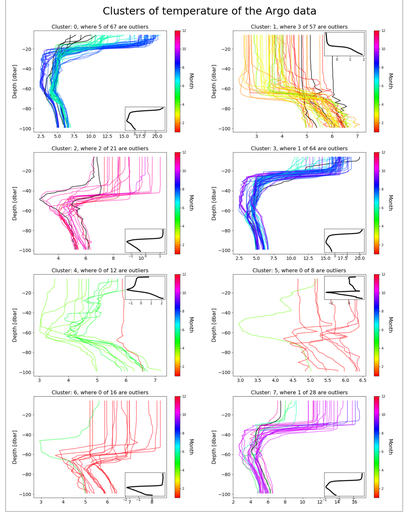

Figure 2: Argo temperature profiles of Gotland Deep classified in 8 clusters. The colors represents months and outlying profiles are drawn with black.

We used Argo profiles deeper than 100 m. The profiles were normalized and then clustered with DTW and K-means. By this, we could classify temperature profiles according to the season.

Outliers

In the manual QC of the Argo data some profiles where labeled as faulty because of clogged sensors. The clustering of the salinity profiles was able to recognize most of these.

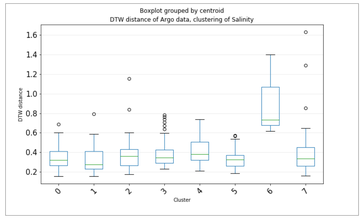

Figure 3: Boxplot of DTW distances of the Argo salinity profiles.

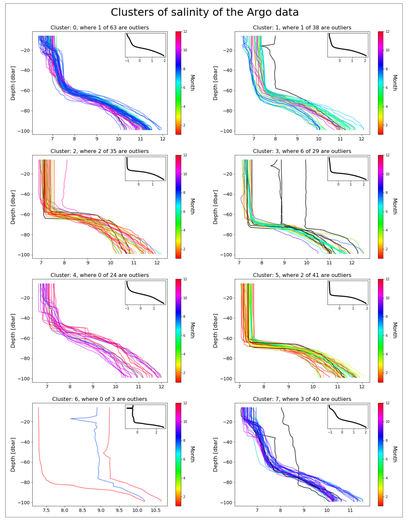

Figure 4: Argo salinity profiles of Gotland Deep classified in 8 clusters. The colors represents months and outlying profiles are drawn with black.

Of all 273 salinity profiles we labeled the three in cluster 6 and 15 in other clusters as possibly faulty.

As there are large horizontal and vertical gradients in the Baltic Sea temperature and salinity and large seasonal cycle in the surface layer temperature, the detection of outliers is not straightforward. Simple range check/scatterplot QC is insufficient. In our analysis, a few profiles where classified into the cluster of a 'wrong season'. Such profiles should also be considered as possible anomalies.

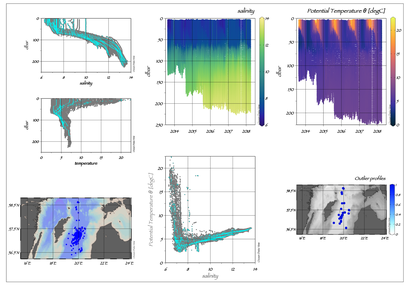

Figure 5: FMI-Argo profiles of Gotland Deep area. In scatterplots outlier profiles are on light blue.

Search for fronts and processes

In the glider data, we searched for profiles with subsurface chlorophyll-A maximum, as seen in Figure 6, from 71 profiles measured in 43 hours.

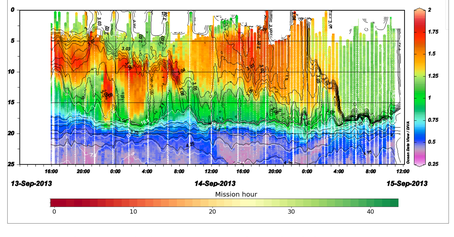

Figure 6: Chlorophyll-A in top layer of the GROOM2013 transect. Isobars are on black. The bottom scale refers to the time from the start of the segment.

The clustering found profiles with subsurface Chl-a maximum during the first 20 hours. At the end of the section, the glider faced a completely different water mass, which was recognized with the clustering of the profiles.

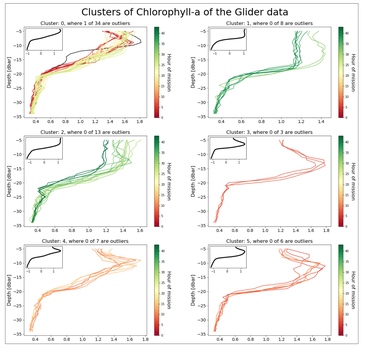

Figure 7: Glider profiles of chlorophyll-A clustered in 6 clusters. The color prepresents the time from the start of segment.

Conclusions

We applied unsupervised machine learning algorithms to classify ocean profiles. The method classified oceanographic profiles meaningfully, found profiles that need closer inspection and helps in interpretation of sparse data.

References

Maze &al. (2017). Profile Classification Models. Mercator Ocean Journal, 2017-04.

Siiriä &al. (2018). Applying area-locked, shallow water Argo floats in Baltic Sea monitoring. Journal of Operational Oceanography. DOI: 10.1080/1755876X.2018.1544783.

Tavenard, Romain (2017). tslearn: A machine learning toolkit dedicated to time-series data. https://github.com/rtavenar/tslearn

Corresponding author: Kimmo Tikka, email: kimmo.tikka@fmi.fi