Clustering hydrographic conditions in Galician estuaries

David E. Losada[1], Pedro Montero[2], Diego Brea[1], Silvia Allen-Perkins[2], and Begoña Vila[2]

[1] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela, Spain
david.losada@usc.es, diebrea@gmail.com
[2] Instituto Tecnolóxico para o Control do Medio Mariño de Galicia (INTECMAR), Vilagarcía de Arousa, Spain
pmontero@intecmar.gal, scaceres@intecmar.gal, bvila@intecmar.gal

Abstract. In this paper we describe our endeavours to explore the role of unsupervised learning technology in profiling marine conditions. The characterization of the marine environment with hydrographic variables allows, for example, to make technical and health control of sea products. However, the continuous monitoring of the environment produces large amounts of data and, thus, new information technology tools are needed to support decision-making. We present here a first contribution to this area by building a tool able to represent and normalize hydrographic conditions, cluster them using unsupervised learning methods, and present the results to domain experts. The tool, which implements visualization methods adapted to the problem at hand, was developed under the supervision of specialists on monitoring marine environment in Galicia (Spain). This software solution is promising to early identify risk factors and to gain a better understanding of sea conditions.

1 Introduction

The Ría de Vigo is the southernmost of Galician Rías (NW Spain), several inlets placed at the northern boundary of the NW Africa upwelling system [21]. The Ría de Vigo is a 32 km long v-shaped, 40 m average depth estuary connected to the shelf by a 52 m deep southern channel and a 23 m deep northern mouth, separated by the Cies islands. In the inner part of the ria, the main river, Oitaven-Verdugo is placed. The runoff of this river is mainly seasonal with high flow in winter and low in summer [14].

The wind-driven barotropic flow is the main driving force of the residual circulation [17], with time responses of local winds and remote winds within 6 h and 12 h [7]. From March-April to September-October, due to southward winds, cold and nutrient-rich Eastern North Atlantic Central Water (ENACW) upwells onto the shelf and is introduced to the ria [6, 1]. This entrance of ocean water mainly comes from the bottom. During the rest of the year, SW winds provoke the entrance of warm and nutrients-depleted water by the surface, blocking the circulation of the ria. This condition is related to a high runoff of the river and an exit of surface fresh water, producing a convergent front in the ria. However, this stational scheme gives only a general picture since the events with frequencies <30 days explains >70% of the variability [13]. As a

general picture, the winter condition consists in a stratified column maintained by the fresh water discharge, in spite of the thermal inversion by heat losing through the surface. In summer, the upwelling colds down the bottom layers. This fact along the warm surface water (because of radiation) causes a strong thermal stratification. The rest of the year, the column mix is dominant because of downwelling events.

Galician Rias are one of the most productive oceanic regions of the world [3]. Subtidal dynamics is important since it is the main responsible for the net export and import of water, nutrients, contaminants, plankton, to and from the ria of Vigo [8]. Hydrography is a fundamental tool to understand the dynamics of the ria.

The key contribution of this research is to design a tool able to cluster and visualize hydrographic conditions in Galician rias. An automatic categorization of hydrographic conditions is fundamental to determine the target *typical* state of the ria and its anomalies, which could be used as descriptor 7 of European Marine Strategy Framework Directive (MSFD). Any permanent alteration of these conditions could be used as an indicator of loss of good environmental status. Moreover, the determination of these typical conditions is the first stage to obtain an environmental impact assessment. As a matter of fact, it is important to understand how anthropogenic activity results in deviations from the usual marine conditions. On the other hand, the behavior of most phytoplankton species is influenced by the physical conditions, and knowing how much conditions deviate from the average could be a proxy of the bloom or decay of these species. There is also a direct relationship between the subtidal circulation and the hydrography and, thus, the knowledge derived from our tool can be also indicative of the capability of natural cleaning. The clustering of marine conditions can also help to understand where the marine litter will go to.

This is a preliminary research project that aims at exploring the possibilities of unsupervised learning technology. We therefore selected an initial sample of Galician stations (see Fig. 1) and we represented the data extracted from these stations at different points in time.

2 Materials and Methods

2.1 Collection of observations

In order to control the quality of the water in the Galician shellfish harvesting areas, INTECMAR[1] weekly monitors the hydrography of Galician coast. These weekly campaigns have been running since 1992. The current oceanographic network is formed by 43 oceanographic stations distributed along Rias Baixas and the Ría de Ares. Eight of these stations are located in the Ría de Vigo. Among other measurements, salinity and temperature profiles are recorded using a SBE25 CTD (conductivity-temperature-depth) profiler. Conductivity measurements are converted into salinity values using the UNESCO equation [20]. Every week, the obtained raw CTD data are processed, filtered and bin averaged using the standard prescriptions of the CTD manufacturer [9, 10]. All data are downloaded, processed and saved on the INTECMAR data center (and distributed through www.intecmar.gal).

[1] www.intecmar.gal

In order to make an initial prototype, only the samples of two years (2015-2016) obtained from the Ría de Vigo were considered. The profiles of temperature and salinitiy were used to represent the hydrographic conditions.

2.2 Data Representation

The main aim is to automatically discover associations among campaigns and, thus, for each campaign, the information collected from all the stations is represented into a single *campaign representational unit*. This is a vector of numerical values (temperature and salinity) obtained from all the stations at different levels of depth.

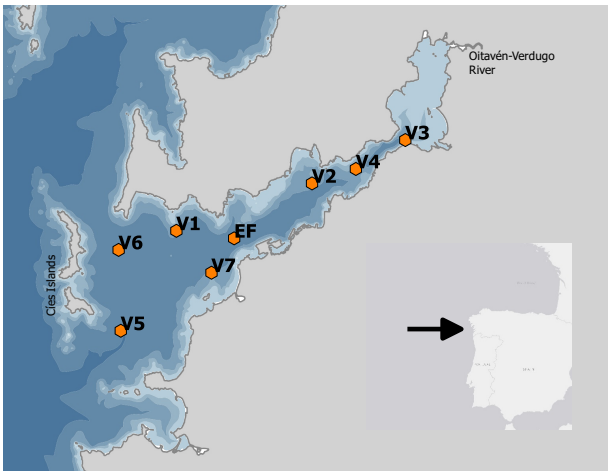

Fig. 1. Stations in one Galician estuary ("Ría de Vigo")

The information obtained from each station comes originally in the form of triples: $(depth, temperature, salinity)$ from INTECMAR data center. For example, $(2.35\ m, 12.34\ ^oC, 35.28)$, $(2.47\ m, 12.38\ ^oC, 35.48)$, and so forth[2]. The depth values are not uniform over campaigns. For example, campaign 1 might have measures at depth levels of $2.35\ m$, $2.47\ m$, ... while campaign 2 might have measures at depth levels of $2.15\ m$, $2.87\ m$, and so forth. Such inconsistencies come from the characteristics of the measuring devices and the type of bin average routinely used in the procedure. From a data representational perspective, this demands some normalization strategy. Following domain expert knowledge, superficial measures (all measures taken at depths lower than 2 metres) were discarded and the remaining measures were bin averaged at 1 meter intervals. This leads to the following intervals: $[2m, 3m)$, $[3m - 4m)$,... For each interval, all measures whose depth falls in the interval are aggregated by weighted average. This results in a representation that has two values for each interval: weighted average

[2] Temperature is measured in degrees centigrades and practical salinity is dimensionless.

of temperatures and weighted average of salinities in the interval. Given l_i, the number of intervals for station i, the overall vectorial representation of a given campaign is:

$$station_i = (T_{1,i}, S_{1,i}, T_{2,i}, S_{2,i}, \cdots, T_{l_i,i}, S_{l_i,i}) \quad (1)$$

$$campaign_k = (station_1, station_2, ...) \quad (2)$$

where $T_{n,i}$ ($S_{n,i}$) is the temperature (salinity) of the n-th interval in station i. The geographical locations of the stations (located at different points of the estuaries) make that the maximum depths are different and, thus, different stations contribute with different number of values to this representation. Additionally, since the maximum depth of each cast can vary among campaigns (e.g., due to the sea-weather, tide and surveyor skill), the deepest measures of each cast were discarded.

In this way, $campaign_k$ represents the hydrographic condition of the estuary at the k-th campaign (on a given date). This condition is modelled by the sequences of temperatures and salinities obtained at different depths from stations located at strategic points in the estuaries. By automatically associating $campaign_k$ with other campaigns, we can relate current campaigns with conditions seen in the past, we can profile marine conditions and we can try to anticipate risk factors.

2.3 Clustering campaign data

Each campaign is represented by a vector of values (eq. 2) and the main purpose of this new marine application is to cluster campaigns into groups. To meet this aim, we employ k-means [11, 12]. K-means is a well-known clustering algorithm that finds clusters and cluster centroids in a set of unlabelled data. The number of desired clusters (k) has to be chosen in advance and k-means proceeds iteratively by moving the cluster centroids in order to minimize the total within cluster variance. Given an initial set of centroids, k-means alternates between i) identifying the data points that are closer to each cluster centroid and ii) updating the centroids by computing the average of the points in each cluster (each cluster centroid is the vector of the feature means for the points in the cluster). The algorithm iterates until convergence. K-means aims to find a good set of non-overlapping clusters. And the main intuition is that a good cluster is one for which its points do not differ much from each other.

Cluster quality. There are different methods for choosing the optimal number of clusters. Next, we describe some of them. The Elbow method [19] runs k-means for a range of values of k and for each value of k computes the total within-cluster sum of squares (WSS). Such an approach estimates the compactness of the clustering from the pairwise squared Euclidean distances between the points in the cluster. The Elbow method plots WSS against the number of clusters and suggests to choose the number of clusters so that adding another cluster does not reduce much the total WSS. To meet this aim, the presence of an elbow or knee in the plot is considered as an indicator of the ideal number of clusters.

Silhouette plots [18] are alternative displays for interpreting and validating clusterings. They graphically represent clusters by a *silhouette*, which depicts the tightness and

separation of the clusters. For each point, its silhouette score measures how similar the point is to its own cluster (cohesion) compared to the other clusters (separation). This score ranges into $[-1, +1]$ and a score close to 1 means that the point fits well with its cluster and it is dissimilar to the other clusters.

Caliński and Harabasz [4] proposed another criterion to evaluate the quality of clusterings. It evaluates cluster validity based on the mean between- and within-cluster sum of squares. Davies and Bouldin [5] presented a measure that can be also used to infer the appropriateness of cluster partitions. Their measure incorporates well-accepted features employed in cluster analysis and its design was driven by certain heuristic criteria.

Our tool implements the four clustering quality measures described above. These four estimates can be used to automatically filter out bad partitions. However, the output of a given clustering configuration requires human interpretation and, thus, our tool is flexible and allows the user to specify the number of clusters, analyze the results, visualize the campaigns associated to each cluster, etc. As a matter of fact, this subjective analysis, done by the domain expert, should shed light on what is to be considered a good cluster of marine conditions.

Dimensionality reduction. The high number of dimensions or features in the campaign vectors makes it difficult to visualize clustering results. We therefore adopted some standard dimensionality reduction methods that are used for presenting the output of the clustering in three-dimensional graphs.

Principal Component Analysis (PCA) [15] is a traditional way to do dimensionality reduction. It is a statistical method based on orthogonal transformation that converts a set of points represented with possibly correlated features into a set of points represented with a set of linearly uncorrelated features (known as principal components). The transformation is performed in such a way that the first component accounts for as much of the variability in the data as possible, the next component has the highest variance under the constraint that it is orthogonal to the first component, and so forth.

In exploratory data analysis, PCA is often employed for visualization purposes. High-dimensional datasets cannot be easily explored and analyzed by humans. PCA supplies the user with low-dimensional representations. These representations, which can retain as much of the variance of the original representation as possible, can be plotted on informative graphs. Our tool uses PCA to generate visually amenable graphs that better communicate the clusterings to the domain expert.

3 Experiments

The dataset was built from the measures obtained from eight stations in one Galician estuary ("Ría de Vigo"). More specifically, we got data from the stations labelled as EF, V1, V2, V3, V4, V5, V6, and V7 in Figure 1. We analyzed the database provided by INTECMAR and selected an initial sample of dates (years 2015 and 2016). The overall number of campaigns in this sample (e.g. the number of points to be clustered) is 80. This is a small sample but it helps us to make initial tests with the tool. In the future, we plan to extend this cluster analysis to many more data points (larger range of dates, more campaigns and more stations from other Galician locations).

Given the characteristics of the eight stations (maximum depths of the measuring exercises), each data point was represented by a vector with 254 features (127 temperatures + 127 salinities). On average, each station contributed with about 16 depth levels.

The tool we developed is written in Python. This facilitates the incorporation of multiple data analysis libraries and toolkits. Furthermore, it is a language that is currently employed in several INTECMAR projects and, thus, the tool can be later adapted and maintained by INTECMAR analysts. Unsupervised learning is driven by a number of libraries and classes from scikit-learn [16].

3.1 Preliminary Tests

We first performed a number of experiments designed to set the main configuration options of the clustering algorithm. We worked with two versions of k-means, k-means and MiniBatchKMeans[3]. The first is a standard k-means implementation, while the latter is a variant that uses mini-batches to reduce the computation time. MiniBatchK-Means optimizes the same objective function as k-means but drastically reduces the computational effort required to converge to a solution. Mini-batches are sub-samples of the input data that are randomly selected at each training iteration. In contrast to other solutions that reduce k-means' computational time, mini-batch k-means outputs results that are generally only slightly worse than k-means' results. With the current dataset, computational time is not a major concern and, therefore, we did not observe substantial differences between both algorithms. We decided to adopt k-means for the subsequent experiments. However, our tool can be easily configured to work with Mini-BatchKMeans (if needed for performing large-scale experiments with massive datasets of marine conditions).

Next, we varied a number of parameters and observed the results obtained. More specifically, we tested some initialization parameters and a parameter related to the maximum number of iterations. K-means finds a local optimum rather than a global optimum. As a result, the final output depends on the initial set of centroids. For this reason, it is customary to run k-means multiple times from different initial configurations. This is governed by the parameter `n_init`, which we set to 1000. This means that the final results reported will be the best output (minimum within-cluster sum of squares) of 1000 consecutive runs of k-means. For each execution of the algorithm, the maximum number of iterations was set to 1000 ($\texttt{max\_iter} = 1000$). We also experimented with different initialization choices: i) a random selection of centroids, which chooses k data points at random for the initial centroids, and ii) `k-means++` [2], a more sophisticated selection of seed centroids, which selects initial centroids in a smart way to speed up convergence. Although there was no much overall difference, we finally selected the following configuration for the subsequent experiments:

```
KMeans(init='k-means++', n_init=1000, max_iter=1000)
```

[3] The corresponding scikit-learn classes are sklearn.cluster.KMeans and sklearn.cluster.MiniBatchKMeans, respectively.

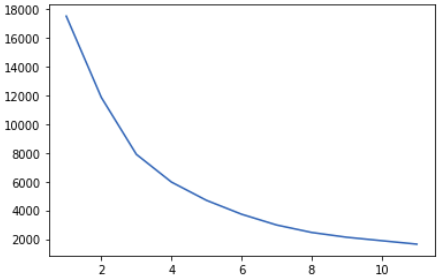

Fig. 2. Elbow method. The X axis represents the number of clusters and the Y axis represents the corresponding sum of squared errors.

Table 1. The three best configurations according to Silhouette, Caliński-Harabasz and Davies-Bouldin metrics. The figures in the table are numbers of clusters.

	best configuration	2nd best configuration	3rd best configuration
Silhouette	8	5	10
Caliński-Harabasz	6	7	8
Davies-Bouldin	5	10	11

3.2 Ideal number of clusters

First, the experimentation focused on selecting the number of clusters. To meet this aim, we experimented with clusterings with up to 12 clusters and we computed the metrics described in section 2.3. Figure 2 depicts the results of the Elbow method. Although there is not a clear knee, it appears that the most consistent solutions are clusterings with a number of clusters in the range from 3 to 7. Next, we proceeded to compute the Silhouette, Caliński-Harabasz and Davies-Bouldin metrics. Table 1 reports the suggested number of clusters of the three best configurations according to these metrics. The results are a mixed bag. They seem to suggest a high number of clusters (greater than 5) but there is not a clear choice.

Figures 3, 4 and 5 show the Silhouette graphs for clustering configurations from 2 to 10 clusters. This graphical presentation helps to shed light on the ideal number of clusters. For each plot, the X axis represents the Silhouette scores of the data points (the higher the better). Data points with Silhouette scores close to 0 are on the border between two clusters. The dashed vertical line represents the average silhouette score of all the values in the plot. Each plot contains a certain number of clusters, where each cluster is represented by a bar graph with the Silhouette scores of the data points in the cluster. The bar graph of each cluster is labeled with a number (from 0 to the number of clusters minus 1). The thicker the bar graph, the larger the cluster (more data points were assigned to the cluster). A number of observations can be derived from this visual analysis. First, cluster configurations with 5 or more clusters tend to produce clusterings where some clusters have very few data points. For example, with 5 clusters, cluster #3 is tiny. The same happens for cluster #5 (6 clusters plot) or cluster #0 (8 clusters plot). These partitions do not seem reliable as these tiny clusters hardly

reflect a real group of similar marine conditions. Furthermore, many of these cluster configurations (for example, all cases with 7 or more clusters) show data points with negative Silhouette scores, reflecting far from ideal cluster partitions (some points do not fit well with their cluster). Clusterings with 2, 3 or 4 clusters look better. However, the 2 and 4-cluster plots show also some negative scores. This suggests that the 3-cluster partition is the most natural choice. This outcome was discussed with a domain expert, who analyzed the groups and confirmed that a 3-cluster solution is rational and the three groups could be associated with three typical environmental conditions. We therefore adopted 3 clusters as our configuration for the rest of the analysis.

3.3 3D Visualization

To further analyze the quality of this clustering, we proceeded to apply PCA on the data. This transformation was done for visualization purposes and, thus, it is performed *after clustering* the original dataset (the dataset is clustered using the original set of features and, next, the data points are transformed into a reduced PCA space of features). With three principal components, the PCA transformation of the dataset maintains 86% of the variance. This suggests that visualizing the clusters using the three PCA principal components does not lose much information. Figure 6 shows the results. This graph confirms that a 3-cluster configuration partitions the data in a rational way, with no much overlapp among the three cluster *regions*.

3.4 Domain-adapted visualizations

Although the visual analysis described above helps to understand the relative quality of the clusterings, it is still not very informative to domain experts. To better communicate the results obtained, we designed a number of domain-adapted visualizations. First, we focused on understanding the main characteristics of the groups found. To meet this aim, we obtained the three cluster centroids and we created centroid visualizations amenable to analysis by domain experts. More specifically, we employed the centroids as obtained in the original set of dimensions (no PCA) and we separately presented the patterns of temperatures and salinities. The resulting visualization is shown in Figure 7. The vector was divided by station, and sorted from the inner stations (top profiles), to the outer stations (bottom profiles). The casts are presented as usual in the oceanographic style, where depth is shown in the y-axis. Left profiles correspond to temperature and right ones to salinity. Each cluster is represented by a line. This graph was analyzed by the domain expert and he found it highly informative. For example, we found that cluster #2 (green dotted line) represents a typical winter condition, where temperatures are lower and constant for each profile. The salinity is lower for the surface layers, mainly in the inner stations, due to the high river discharges during this season. The orange dashed line (cluster #1) can be representing upwelling events, with temperatures in the bottom layers very low (even lower than in the winter cluster #2 condition in the deepest stations: EF and V5) and high temperatures at the surface (because of sun radiation). The salinity profiles are mostly constant and high as corresponding with a lower river runoff. Only in the inner stations, the surface fresh water signal is significant. The blue solid line (cluster #0) is related to a downwelling/mixed condition, warmer

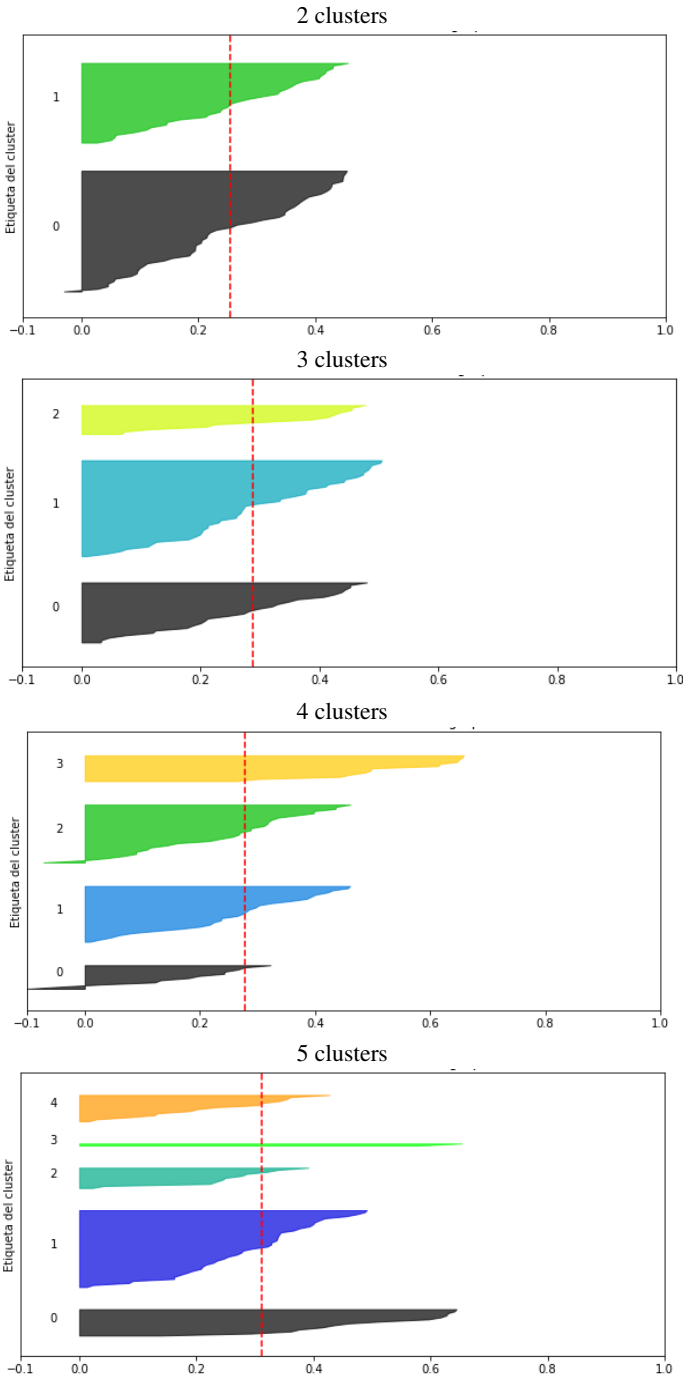

Fig. 3. Silhouette graphs (2-5 clusters)

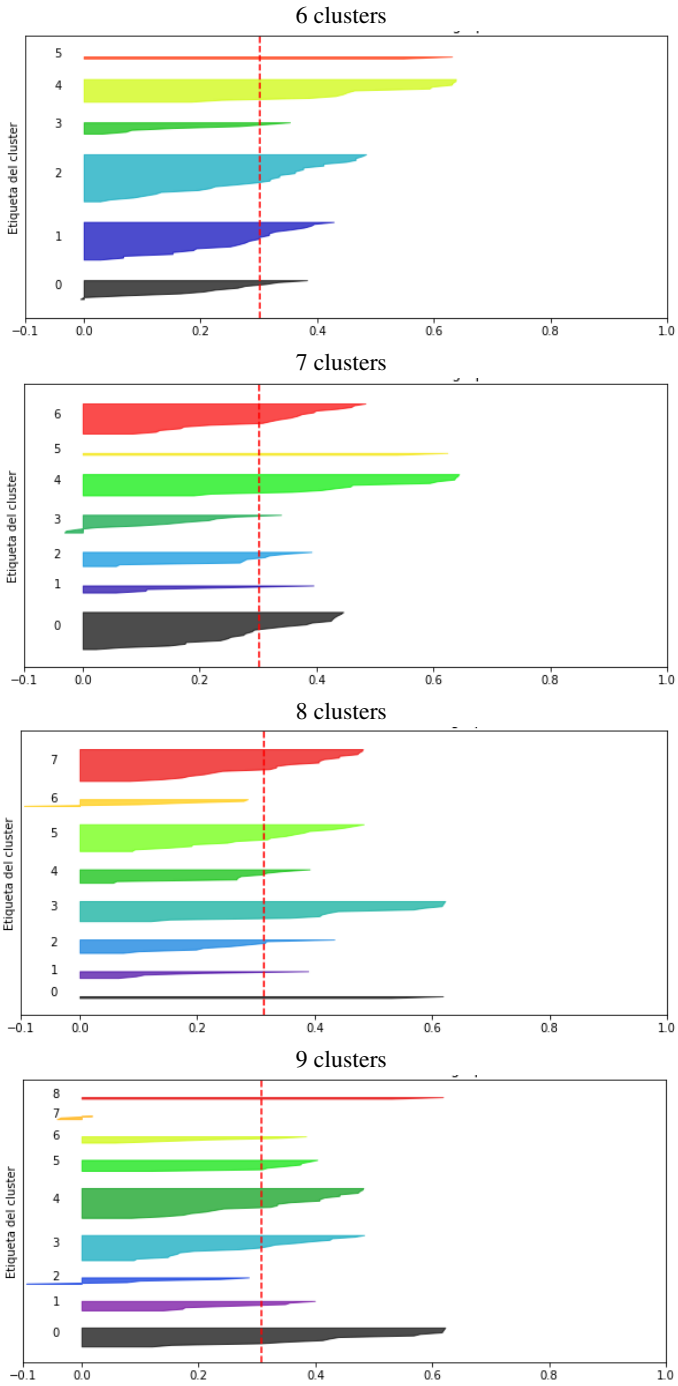

Fig. 4. Silhouette graphs (6-9 clusters).

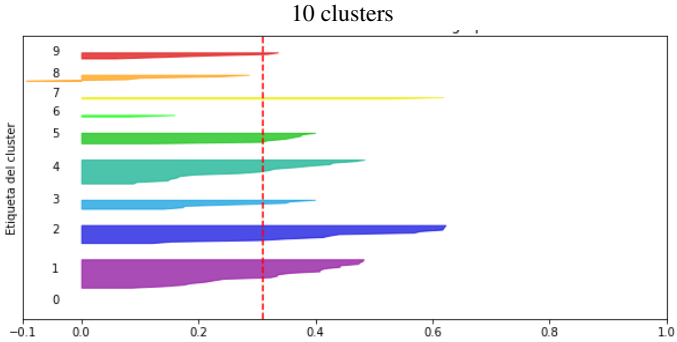

Fig. 5. Silhouette graphs (10 clusters).

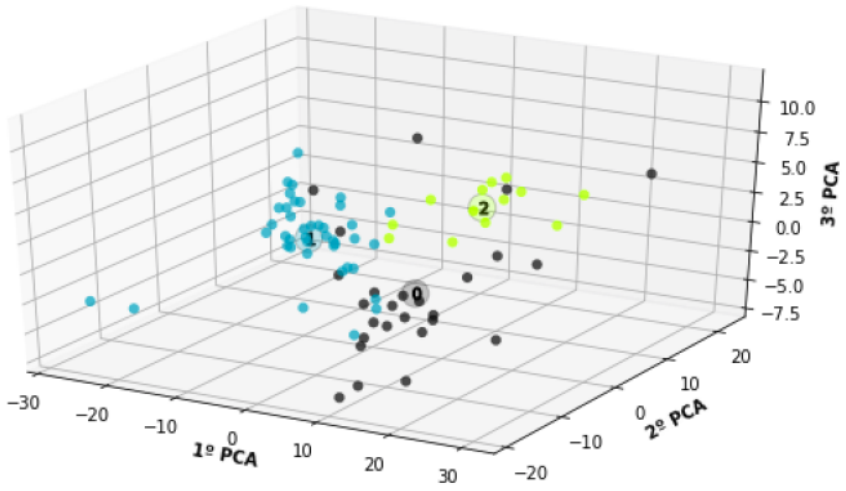

Fig. 6. 3D visualization of the proposed clustering

than the condition described above, mainly in the surface layers and with significant presence of freshwater in the inner stations.

In order to explore the correspondence of these clusters with the individual campaigns, figure 8 shows the cluster assignment (Y-axis) for each campaign date. Cluster #2 is associated to winter dates (February and March, which are months with high river runoffs). In 2015, upwelling-downwelling conditions alternated over the rest of the months. In 2016, upwelling events were grouped mostly during summer and early autumn, and downwelling during the last months of the year.

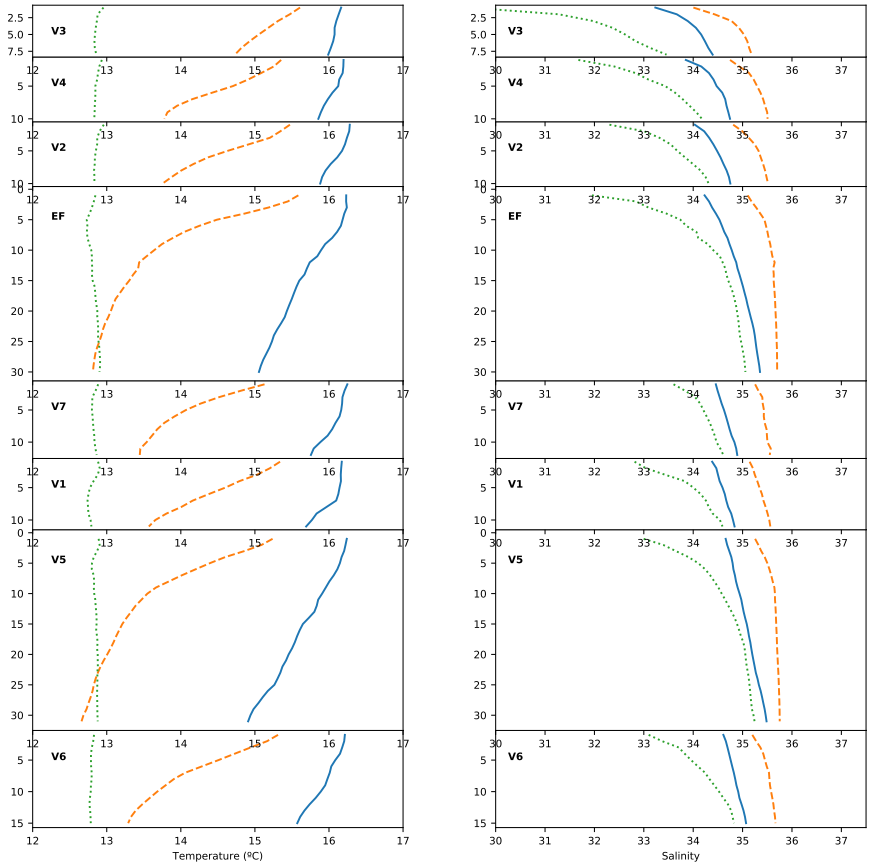

Fig. 7. Temperature (left graph) and Salinity (right graph) patterns associated to the three cluster centroids.

4 Conclusions

We designed and developed a new tool that employs clustering and PCA to group hydrographic events. Using this new tool, K-means clustering effectively organized data measures obtained from marine campaigns. With a 3-clustering setting and using data from two years of CTD weekly campaigns in the Ría de Vigo, the tool recognized typical conditions described in the literature (upwelling, donwelling events and transition between them [13]). Furthermore, the influence of the river runoff during winter and the surface radiation during summer was also detected and visualized.

This work represents our first attempt to use machine learning to understand hydrographic conditions. This research needs to be extended in a number of ways: a) to enlarge the study to datasets associated to larger periods of time; b) to apply this

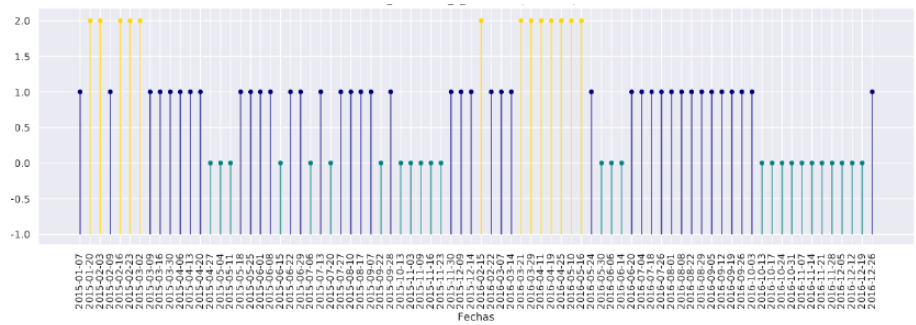

Fig. 8. Chronological visualization of the clusters. The X axis represents the dates of the campaigns and the Y axis represents the cluster assignment (0, 1 or 2)

methodology to other rias and c) to test a higher number of clusters. By considering and analyzing a larger number of clusters, we might be able to discover specific conditions that are not fully described in the literature. Furthermore, the centroids of these classes could be indicative of unknown circumstances, such as deviations or defects in CTD campaigns or new trends in the ria. The potential identification of unknown conditions is a promising feature that can help to early identify risk factors and to further understand the conditions of the sea.

Acknowledgements

This research has received financial support from the Galician Ministry of Education (grants ED431C 2018/29 and ED431G/08, co-funded by the ERDF), and the European Union MarRISK project: “Adaptación costera ante el Cambio Climático: conocer los riesgos y aumentar la resiliencia” (0262_MarRISK_1_E), through EP-INTERREG V A España-Portugal (POCTEP) program. www.poctep.eu/es/2014-2020/marrisk. The third author thanks the support obtained from Xunta de Galicia through the "BECAS EXCELENCIA XUVENTUDE EXTERIOR" program.

References

1. X.A. Alvarez-Salgado, J. Gago, B.M. Miguez, M.Gilcoto, and F.F. Perez. Surface waters of the nw iberian margin: Upwelling on the shelf versus outwelling of upwelled waters from the rías baixas. *Estuarine, Coastal and Shelf Science*, 51(6):821 – 837, 2000.
2. David Arthur and Sergei Vassilvitskii. K-means++: the advantages of careful seeding. In *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.
3. J.O Blanton, L.P Atkinson, F.F Castillejo, and A. Lavin-Montero. Coastal upwelling off the rias bajas, galicia, northwest spain i: Hydrographic studies. *Rapp. P.- v. Rèun. Cons. Int. Explor. Mer.*, 183:79–90, 1984.
4. T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

5. D.L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227, February 1979.
6. F. Fraga. *Upwelling off the Galician Coast, Northwest Spain*, pages 176–182. American Geophysical Union (AGU), 2013.
7. M. Gilcoto, J. L. Largier, E. D. Barton, S. Piedracoba, R.o Torres, R. Graña, F. Alonso-Pérez, N. Villacieros-Robineau, and F. de la Granda. Rapid response to coastal upwelling in a semienclosed bay. *Geophysical Research Letters*, 44(5):2388–2397, 2017.
8. M. Gilcoto, P. C. Pardo, X. A. Álvarez Salgado, and F. F. Pérez. Exchange fluxes between the ría de vigo and the shelf: A bidirectional flow forced by remote wind. *Journal of Geophysical Research: Oceans*, 112(C6), 2007.
9. Seabird Inc. *SBE 25 Sealogger CTD, user's manual*. Seabird Inc., Bellevue, WA, USA, 13 edition, 2005.
10. Seabird Inc. *Seasoft V2: Sbe Data Processing, Software Manual*. Seabird Inc., Bellevue, WA, USA, 7.26.8 edition, 2017.
11. S. P. Lloyd. Least squares quantization in PCM. Technical report, Bell Laboratories, 1957.
12. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.
13. E. Nogueira, F.F. Pérez, and A.F. Ríos. Seasonal patterns and long-term trends in an estuarine upwelling ecosystem (ría de vigo, nw spain). *Estuarine, Coastal and Shelf Science*, 44(3):285 – 300, 1997.
14. P. Otero, M. Ruiz-Villarreal, Á. Peliz, and J.M. Cabanas. Climatology and reconstruction of runoff time series in northwest iberia: Influence in the shelf buoyancy budget off ría de vigo. *Scientia Marina*, 74(2), 2010.
15. K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
16. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
17. S. Piedracoba, X. A. Álvarez Salgado, G. Rosón, and J. L. Herrera. Short-timescale thermohaline variability and residual circulation in the central segment of the coastal upwelling system of the ría de vigo (northwest spain) during four contrasting periods. *Journal of Geophysical Research: Oceans*, 110(C3), 2005.
18. Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.
19. Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, Dec 1953.
20. UNESCO. The international system of units (si) in oceanography. *UNESCO Technical Papers in Marine Science*, 45:131, 1985.
21. Warren S. Wooster, Andrew Bakun, and Douglas R. McLain. Seasonal upwelling cycle along the eastern boundary of the north atlantic. *Journal of Marine Research*, 34(2):131–141, 5 1976.