Clustering: Fish Environment vs Fish Stocks

Valeriia Bondarenko ; Émilie Poisson Caillault ; Nicolas Andrialovanirina ; Kélig Mahé

Université du Littoral Côte d'Opale LISIC - IFREMER LRH

Contact: emilie.poisson@univ-littoral.fr

France 2030 « IFSEA » (ANR-21-EXES-0011)

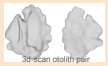
CPER CornellA

Journées IA-Optim Juin 2024, Calais

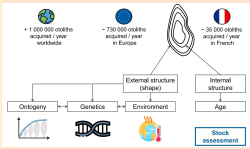
Context

Otolith: calcified structure in the inner ear of vertebrates, allow to :

- identify the fish species, its life area/population
- identify the fish eaten by other fish
- understand the fish life history



3d scan otolith pair



Questions

E1- Do physical environments coincide with different fish stocks ?

E2- What is the influence of changes in environmental habitat over the past one year, three years or five years ?

1y=short-term fluctuations, weather events, fishing pressure
3y=reproductive success, migration, habitat changes
5y= population dynamics, genetic adaptation

Study Case: Red Mullet in Mediterranean sea

88 sites

Mediterranean Sea Biogeochemistry Reanalysis (product identifier MEDSEA_MULTIYEAR_BGC_006_008) on a grid with 1/24° x 1/24° horizontal resolution and 125 vertical levels of thickness increasing with depth.

5 years (2014-2018), 3 years (2016-2018), and 1 year (2018).

***a priori* knowledge:** 2 or 3 stocks on other species, Strait of Sicily and Egeen/Adriatic Sea

What ?

ML 1: Which **clustering** techniques can address the question ?

How many physical eco-regions ? i.e. K number of clusters ?

ML2 : Which **consensus** among these techniques ?

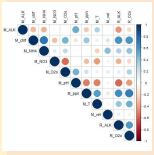
ML3 : Same conclusions with 1/3/5-year past physical information ?

(*E : Environment - ML: Machine Learning view)

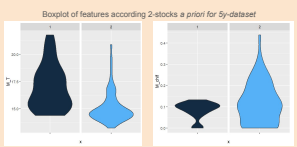
Parameter Name	Feature (MOI)	Unit
Alkalinity	M/R	mol eq kg^{-1}
Chlorophyll concentration (as carbon)	M	mg(C) m^{-3}
Molar ammonium concentration	M	mmol m^{-3}
Molar nitrate concentration	M	mmol m^{-3}
Molar dissolved molecular oxygen concentration	M/R	mmol m^{-3}
pH	M	
Molar phosphate concentration	M/R	mmol m^{-3}
Net primary production of biomass per day	M	mg m^{-2} day^{-1}
Salinity	M	psu
Temperature	M	°C
Velocity module	M	$m\ s^{-1}$

Table of available features (R: Range, M:median)

Good features ? 5-years



Correlation of retained features *for 5y-dataset*



Boxplot of features according 2-stocks *a priori for 5y-dataset*

Framework

Preprocessing: X datatable of features
remove feature fj with correlation(fi,fj)>75% => Xreduced
Xr=scale(Xreduced)
m=dist(Xr)
W=Local Zelnik-Perona Similarity, 7-neighbor
K=2 fixed (DBSCAN tuned for K=2)

Applied Clustering methods (R packages):

*** Hierarchical methods**
hclust Agglomeration by average and ward.D2 criteria
DIANA - Divise Analysis
AGNES - Agglomerative Nesting by Ward
pvclust - Agglomerative with p-values and multiscale bootstrap resampling, ward.D2

*** Crisp Expectation Maximization methods**
EM - VVV variable shape, volume and orientation
Kmeans - centroid-based, globular shape,
PAM - medoid-based
spectralPAM based on the Laplacian of W (diag=0) and its eigenvector space.

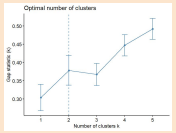
*** Fuzzy Expectation Maximization methods**
cmeans - kmeans, inertia minimization with observation membership weights
FKM - fuzzy kmeans
FKM noise - FKM + noise cluster
FANNY - membership exponent=1

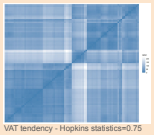
*** Density-based methods**
DBSCAN (eps=72, MinPts = 5)
HDBSCAN(m,minPts=10)

Labels used as *a priori* information
x12- 2-stocks *a priori*
x13- 3-stocks *a priori*
x14 - GSA, Geographical sub-area

Main Used Criteria:

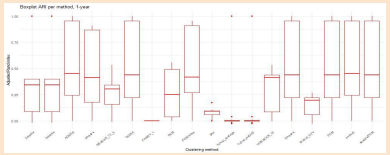
- visualization (ML1)
- Rand index and Adjusted Rand index (ML1,ML2)
- Mean and Fuzzy Silhouette (ML1)
- Re-Assignment of Labels according to 2-stocks *a priori (ML3)*
- Class membership percentage over all methods (ML3)

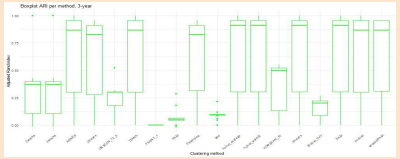


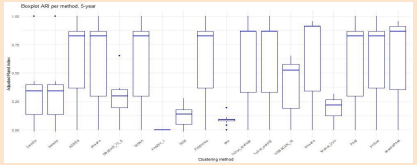


VAT tendency - Hopkins statistics=0.75

Results







Boxplot of Adjusted Rand index per method and per past-year information

ARI >0 in almost cases

Mean(ARI (2-stocks, Methods)) > 0.3

We clearly saw that the analysis can be influenced by the choice of clustering algorithm, distance or similarity measure, and parameters. Different combinations of these factors can produce different clustering results, and there is no universal or optimal choice that will work for every data set.

Having considered the Rand indices as verification criteria for fifteen methods, we believe that the values are quite identical for one year, three and five years. There is no unique or objective way to measure the quality or confidence of clustering, and we need to use both internal and external criteria, as well as visual and qualitative methods, to evaluate and compare clustering results. Moreover, we need to provide meaningful and understandable labels and descriptions for clusters, and explain the implications and applications of clustering for our problem or domain. It should be noted that analysis can be quite labor-intensive for multidimensional data sets.

Discussion

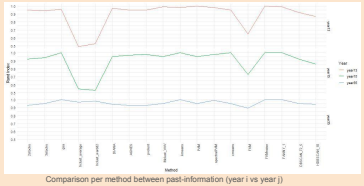
First Answers

E1/ML1- Do physical environments coincide with different fish stocks ?

Mean(ARI (2-stocks, Methods)) > 0.3 - a link but K number could be higher than 2 -> (4-7).

E2/ML3- What is the influence of changes in environmental habitat over the past one year, three years or five years ?

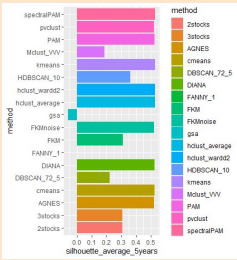
Clearly, ARI per method and per year for ⅗ years are higher



Comparison per method between past-information (year i vs year j)

ML2- Consensus ? close rand index

ML3- Same conclusions 1y-3y-5y? for ⅗years past physical information (blue line above)



References

⅗-stocks a priori:
B. Morales-Nin et al. "European hake (Merluccius merluccius) stock structure in the Mediterranean as assessed by otolith shape and microchemistry". In: Fisheries Research 254 (Oct. 1, 2022), p. 106419. ISSN: 0165-7836.
F. Morat et al. "Discrimination of red mullet populations (Teleostean, Mullidae) along multi-spatial and ontogenetic scales within the Mediterranean basin on the basis of otolith shape analysis". In: Aquatic Living Resources 25.1 (Jan. 2012), pp. 27–39. ISSN: 0990-7440, 1765-2952.

N. Andrialovanirina, E. Poisson Caillault, Kelig Mahé. Red mullet stock identification without any reference using otolith shape. RFIAP 2024