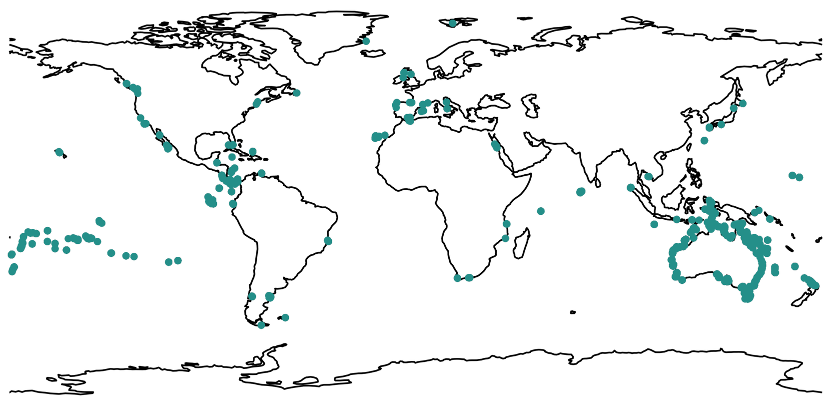
An interpretable clustering pipeline: UMAP/HDBSCAN identifies 24 marine habitats at the global scale

Motivation & Workflow

Standardised classification of marine habitats is critical for mapping and monitoring[1].

Establishing a global typology of coastal marine habitats requires to reconcile fine-scale habitat variability as well as broad-scale biogeographical patterns.

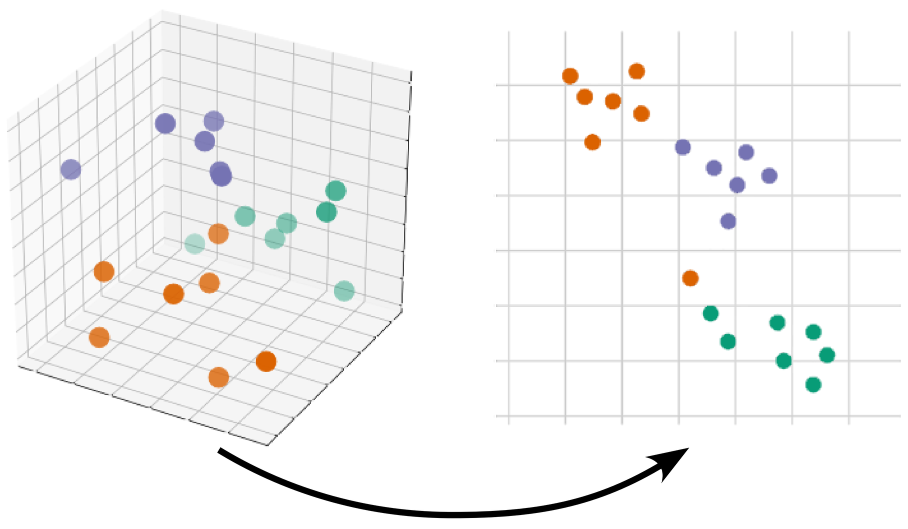
Here, we propose an innovative clustering pipeline to overcome these obstacles.



Data

The *Reef Life Survey*[2] is a citizen science program monitoring costal reefs around the world.

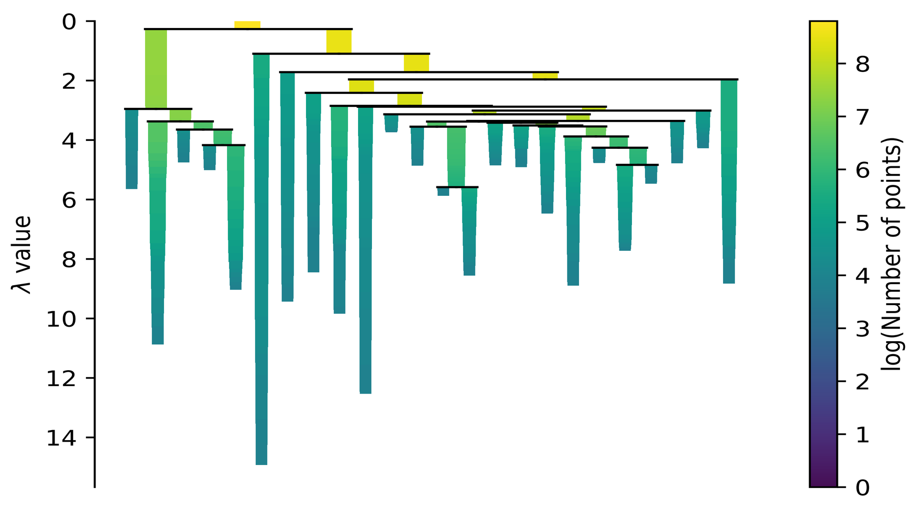
We focused on habitat cover from 6642 sites surveyed by scuba divers.



Dimension reduction

Challenge: Non-linearity in ecological data.

Solution: We used UMAP[3], a non-linear dimension reduction technique, prior to clustering global data into habitat types.



Clustering

Challenge: Citizen science data can be noisy.

Solution: We used HDBSCAN[4] a density-based clustering method. This algorithm makes no assumption about the size and density of the clusters. In addition, it excludes noisy observations from the clusters.

SHAP

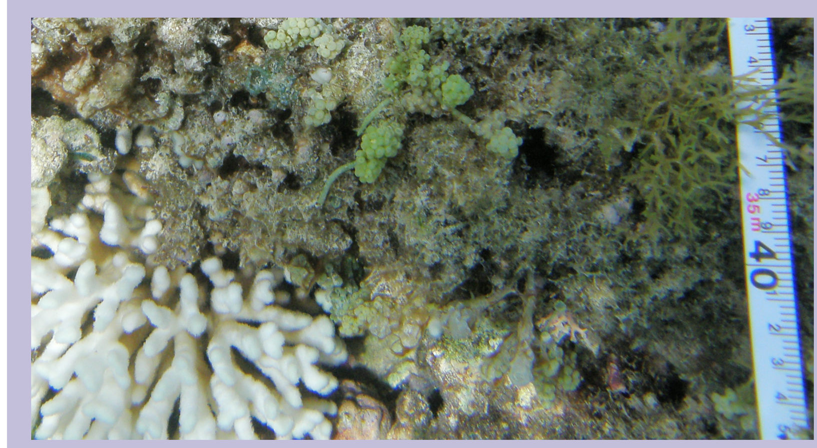
Interpretation

Challenge: Complex unsupervised clustering methods are hard to interpret.

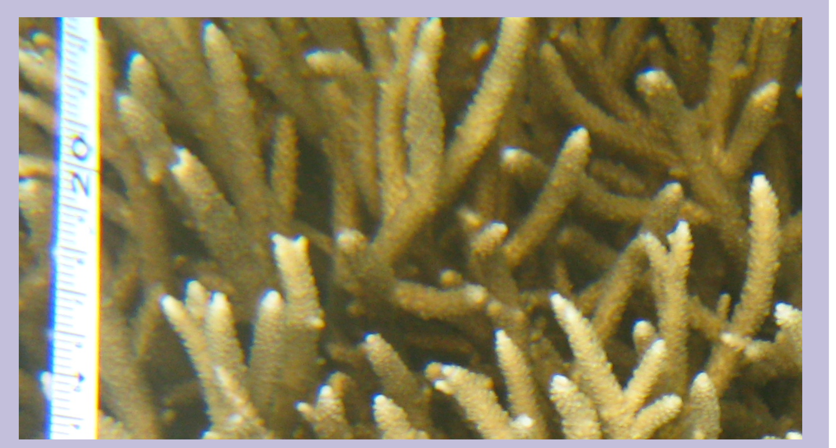
Solution: We used SHapley Additive exPlanations (SHAP)[5] to interpret clusters composition.

Results

Representative photo-quadrats of 4 out of the 24 habitat types identified.



Coral and turf cluster



Hard coral cluster

Fucoid and moderate kelp cluster

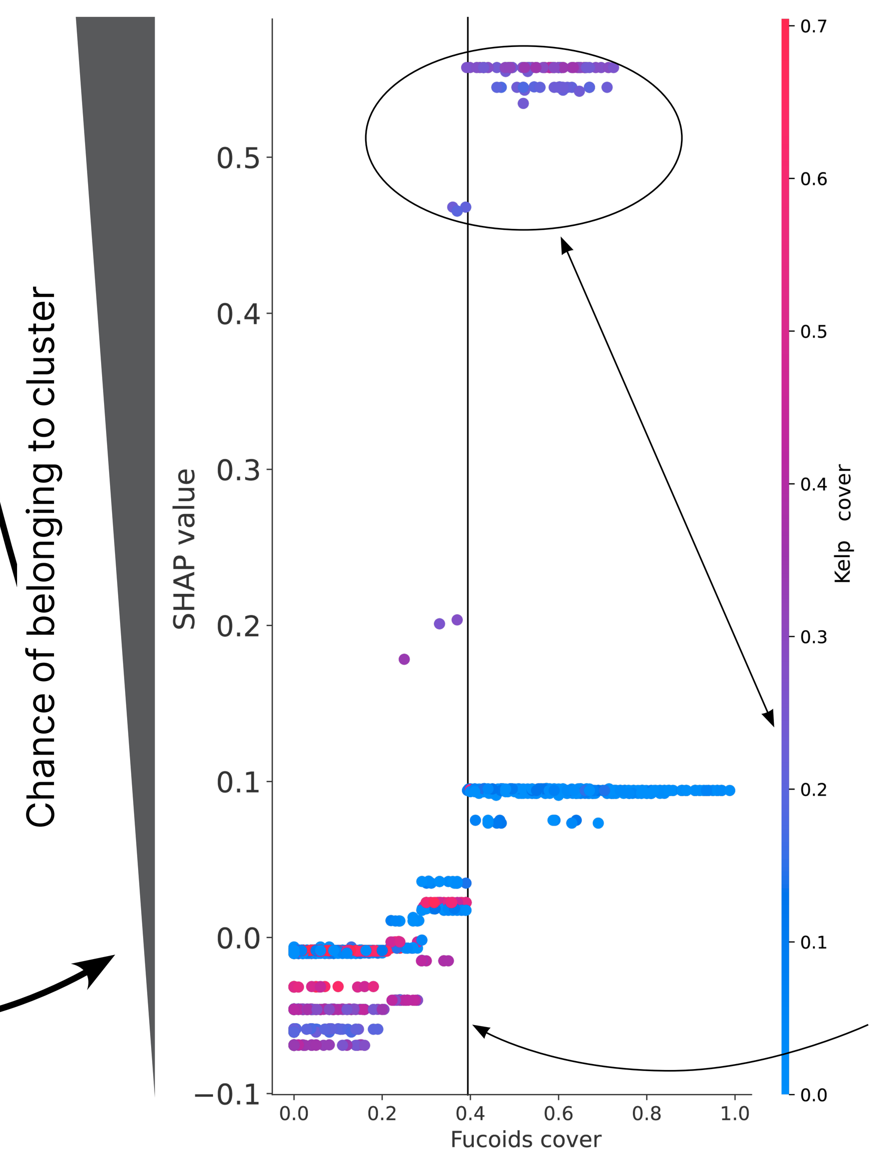
Seagrass meadow cluster

We identified 24 habitat types, which capture both biogeographic patterns (e.g. coral versus kelp-dominated reefs) and finer-scale variability in ecological states (e.g. gradient in healthy coral cover or ecotones).

This habitat typology discriminates both iconic marine habitats and ecotones. Our approach accurately captures regional habitat patterns as described by Cresswell *et al.* (2017)[6] in Australia.

Exploring our clusters

Example of cluster interpretation using SHAP: A combination of high cover of fucoids and moderate cover of kelp characterises this cluster.



This group is characterized by high fucal coverage (>40%) and moderate kelp coverage (≈ 20 %).

Kelp cover threshold

Fucoid cover threshold

Clément Violet, Aurélien Boyé, Graham Edgar, Rick Sutart Smith, Martin Marzloff

clement.violet@ifremer.fr @ClementVIOLET

Flash me to download the poster.

[1] Shumchenia and King (2010). Comparison of methods for integrating biological and physical data for marine habitat mapping and classification.
[2] Edgar and Stuart-Smith (2014). Systematic global assessment of reef fish communities by the Reef Life Survey program.
[3] McInnes and Healy (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.
[4] McInnes and Healy (2017). Accelerated Hierarchical Density Based Clustering.
[5] Ludber and Lee (2017). A Unified Approach to Interpreting Model Predictions.
[6] Cresswell et al. (2017). Translating local benthic community structure to national biogenic reef habitat types.

Ifremer