96

Improving ecological clustering with UMAP and HDBSCAN: an application to the Reef Life Survey datasets

Clustering, i.e. defining groups of similar objects, is a central task of ecology in order to simplify and structure high-dimensional observations of nature's complexity. Clustering methods are fundamental in taxonomy and phylogeny are at the core of the definition of habitat typologies, of biogeographical regions and functional groups, and of the identification of structures in ecological networks, for example. Yet the very definition of a cluster is a complex issue, especially when dealing with high-dimensional data, as is often the case in ecology. This is why special attention must be paid to the choice of the clustering pipeline.

Among the long list of clustering methods that exist, popular ones in ecology, like K-Means or Ward's hierarchical classification, have some limitations: (1) they can only identify spherical shaped clusters (2) are sensitive to outliers (3) cannot distinguish between core cluster members and noisy observations (e.g. transition between typologies). Recently developed clustering approaches can potentially overcome these limitations.

Here we present an application of a clustering pipeline based on a non-linear reduction dimension technique, UMAP, and on the density-based clustering algorithm HDBSCAN on the Reef Life Survey data. This participatory science program surveys the fauna and habitat of coral and rocky reef ecosystems all around the world. Based on more than 7000 underwater diver-based visual censuses across 40 countries, we optimize clustering pipeline for 3 data types : habitat, benthic fauna, fish fauna. UMAP + HDBSCAN often outperform classical approaches. Furthermore, we used UMAP to integrate the three datasets and provide a combined typology, improving our knowledge of the biogeographical patterns across rocky reefs at a global scale and allowing the delineation of core cluster areas, most representative sites and transition zones. This case study demonstrates the usefulness of integrating non-linear density reduction techniques like UMAP and density-based clustering algorithms like HDBSCAN into the ecologists' toolbox to better understand biodiversity.