Ecological Indicators 34 (2013) 231–245

Contents lists available at SciVerse ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Seeking functional homogeneity: A framework for definition and classification of fish assemblage types to support assessment tools on temperate reefs

Miguel Pessanha Pais[a,*], Sofia Henriques[a], Marisa Isabel Batista[a], Maria José Costa[a,b], Henrique Cabral[a,b]

[a] Centro de Oceanografia, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
[b] Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 26 October 2012
Received in revised form 5 May 2013
Accepted 7 May 2013

Keywords:
Marine fish assemblages
Guild approach
Machine learning
Habitat classification
Typology definition
Portugal

ABSTRACT

Due to their important role in the ecosystem and high economic value, there is a need to assess the effect of anthropogenic impacts on marine fish assemblages. However, this can only be achieved if variations due to natural causes are known. Moreover, while most assessment tools rely on functional traits, bottom-up habitat classification frameworks tend to use species composition. The present study proposes an innovative framework to define fish assemblage types through metric pairwise constrained *k*-means (MPCK-means) clustering of sites based on functional guild categories and univariate metrics, an approach that takes into account within-site variability due to the sampling method and natural causes. This was followed by a label-based ensemble clustering approach, which finds patterns that minimise information loss when integrating clustering results from individual metrics. In order to test the method, fish assemblages on 14 nearshore rocky reefs along the Portuguese coast were sampled. The final typology configuration achieved through ensemble clustering consisted of three assemblage types and maintained an average normalised mutual information of 0.605 with the individual clustering results. Nested PERMANOVA found differences among types and the most variable metrics in the face of natural variation were identified. Ultimately, a *k*-nearest neighbours classifier is proposed to label new sites, based only on environmental variables that are unlikely to be directly affected by the presence of anthropogenic impacts. Optimal performance for the classification model was achieved with inverse distance-weighted voting of the 4 nearest neighbours with an average classification accuracy of 96.08%.

1. Introduction

Increasing pressure on the marine environment has led to several policies stressing the need to improve the state of marine ecosystems in the near future and ensure the sustainable use of resources, such as the European Marine Strategy Framework Directive (EC, 2008). Besides the alarming pressure of the fishing industry (Worm et al., 2006), marine fish assemblages, particularly in nearshore rocky reefs, are affected by many other pressure sources (Henriques et al., 2013a). Being in many aspects a highly valued resource (Holmlund and Hammer, 1999), it has become urgent to develop and apply methods that can enable scientists and managers to detect and act upon the sources of pressure affecting fish assemblages.

However, locations with long-term monitoring programmes that enable the comparison of conditions before and after the presence of a particular pressure source are the exception rather than the rule (Borja et al., 2012), and there is a need to develop tools that can signal managers when a fish assemblage has been or is being affected by human activities, without knowledge of the previous state of the system. For this purpose, functional guild approaches have been successfully used in streams and estuaries (Pérez-Domínguez et al., 2012; Roset et al., 2007), not only because they have a broader geographical application, but also because the response of functional guilds to pressure sources can be more predictable and easy to interpret than that of individual species (Elliott et al., 2007). However, changes due to anthropogenic pressures can only be detected when the range of variation due to natural causes is known (García-Charton and Pérez-Ruzafa, 2001; Osenberg et al., 1994).

There are several ways to minimise the effects of habitat in environmental monitoring, which usually involve the establishment of either type-specific or site-specific reference values that

* Corresponding author. Tel.: +351 217500826; fax: +351 217500207.
E-mail address: mppais@fc.ul.pt (M.P. Pais).

1470-160X/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecolind.2013.05.006

represent an ideal situation in the absence of impact or a goal for a given management programme (Borja et al., 2012). In site-specific approaches, each site has its own reference, which may be theoretical (e.g. modelled or historical data) or a direct comparison with one or more control sites with similar characteristics. On the other hand, type-specific approaches begin by establishing habitat types that share certain environmental characteristics and are assumed to support the same potential communities in the absence of impact. The delimitation of habitat types is usually achieved by either a bottom-up approach, where the communities themselves are divided into clusters with similar species composition, or a top-down approach, where sites with similar environmental features are grouped based on quantitative or qualitative criteria (Maxwell and Buddemeier, 2002).

While site-specific approaches allow for greater detail and precision, they are highly impractical at larger scales, thus the definition of habitat types is the most frequent method to support national monitoring programmes and international policy requirements (Borja et al., 2012). For this purpose, many national and international habitat classification frameworks have been established (Costello, 2009). However, the concept of "habitat" varies not only according to scale, but also according to the organisms in question, so top-down approaches may be useful for administrative purposes but are not guaranteed to delimit homogeneous communities for all organisms at the scale needed for a particular management objective (Costello, 2009). Moreover, habitat classification frameworks that use variables such as algal cover and the diversity of sessile fauna to classify sites at smaller scales are of little use in a monitoring context because these variables are also affected by impact sources (e.g. Arévalo et al., 2007) and thus site classification would be biased due to an already altered system.

There is still a discrepancy between species-based classification tools and guild-based assessment tools (Henriques et al., 2013a; Pais et al., 2012). This is an important issue, since functional guilds are more resilient to natural variation, as species are replaced by others from the same guild. This leads to areas with homogeneous guild abundance values tending to be larger than areas with homogeneous species composition (Pais et al., 2012), which is a desirable characteristic of a management-oriented habitat typology, as a large number of types can be impractical or even impossible for medium to large scale monitoring (Johnson et al., 2012). Moreover, unlike species that are either present or absent, the choice of functional guilds is arbitrary and can ultimately depend on management objectives, the expected response to impact sources, among other criteria (Elliott et al., 2007; Henriques et al., 2008, 2013a,b).

In the study of fish–habitat relationships, there is a long history of application of statistical methods that rely on assumptions regarding independence, linearity of responses or probability distributions (Knudby et al., 2010). In fact, ecological data is known to rarely satisfy such conditions (Olden et al., 2008) and fish species and guilds have been shown to have complex, non-linear responses to habitat variables (Friedlander and Parrish, 1998). All these constraints call for non-parametric methods that can deal with complex interactions, non-linearity and unusual distributions. Complex statistical tools that can find patterns and perform predictions based on empirical data have been developed in the field of artificial intelligence and experienced a huge progress in the last decade (Olden et al., 2008). These tools are known as machine learning (ML) techniques and rely on algorithms that are designed to deal with classic statistical problems, such as regression, clustering and classification, by interpreting complex (and often large) databases without having to comply with assumptions and yet outperforming classic procedures (Crisci et al., 2012).

Due to the potential of ML algorithms for interpreting patterns in ecological data, their use is steadily increasing. Nevertheless, when compared to other fields, ML applications in ecology are still at an embryonic stage, probably due to a language barrier between ecologists and computer experts (Olden et al., 2008), aggravated by the fact that some complex models may need very large datasets (Raudys and Jain, 1991) that are often nonexistent in ecology. Despite this, some techniques have shown promising results with ecological data (e.g. Crisci et al., 2012).

In the present study, machine learning algorithms are combined with permutation-based statistical tests to propose a bottom-up approach for the delimitation of reef fish assemblage types based on structural and functional metrics. Additionally, a quantitative model for the classification of new sites according to the established types is tested, by relying on a set of environmental variables that stay unaffected by most impact sources. During the process, the behaviour of several fish-based metrics in the face of natural variation is also assessed.

2. Materials and methods

2.1. Study area

In order to delimit assemblage types that reflect the potential assemblage characteristics associated with environmental variables, an effort was made to select sites without direct influence of impact sources in order to minimise their influence. A total of 14 sites covering a wide array of environmental conditions on nearshore temperate reefs were selected along a 300 km stretch of the Portuguese coast (Fig. 1). In order to optimise their potential to support fish assemblages, surveys were performed during summer, near the spawning season for many species (Henriques et al., 2013b).

2.2. Fish sampling method

Fish assemblages were sampled during daytime using SCUBA diving underwater visual census along 50 m strip transects. Each transect was travelled twice, with a first pass for demersal species (50 m × 2 m) and a second for cryptobenthic species (50 m × 1 m). A 50 m long thin rope was deployed while sampling demersal species, with cryptobenthic fish sampled while reeling the rope, by searching in crevices and under cobbles ≤20 cm in diameter (Henriques et al., 2013a; Pais et al., 2013).

Based on a pilot study to establish a representative number of replicates and calibrate size and abundance estimates between observers (see Henriques et al., 2013a for details), a total of 6 transects per site were performed, half by each observer (M. P. Pais and S. Henriques), starting each time at a random point allocated to one of two pre-determined depth intervals (0–5 m and 5–10 m), according to each site's characteristics.

2.3. Fish-based metrics and guild classification

Metrics were selected based on previous compilations that took into account their use in monitoring and assessment tools and programmes (Henriques et al., 2008, 2013a,b; Pais et al., 2012). A total of 47 metrics were calculated for each transect, representing a range of structural and functional fish assemblage characteristics including diversity, composition, abundance, trophic structure, habitat association, nursery function, mobility, resilience, spawning season and biogeographic affinity (Table 1). All metrics in the "univariate" group in Table 1 were treated individually in the analyses, while all other metrics were treated as part of their multivariate categories (for simplicity, the term "functional categories" will henceforth refer to both univariate metrics and multivariate categories). Species were classified into guilds based on previous studies (Henriques et al., 2007; Henriques et al., 2008, 2013a;