Remote Sensing of Environment 114 (2010) 1230–1241

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques

Anders Knudby *, Ellsworth LeDrew, Alexander Brenning

Department of Geography, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L3G1

ARTICLE INFO

Article history:
Received 24 August 2009
Received in revised form 14 January 2010
Accepted 16 January 2010

Keywords:
Predictive modeling
Machine-learning
IKONOS
Coral reefs
Habitat
Fish community
Ecology

ABSTRACT

During the last three decades, the large spatial coverage of remote sensing data has been used in coral reef research to map dominant substrate types, geomorphologic zones, and bathymetry. During the same period, field studies have documented statistical relationships between variables quantifying aspects of the reef habitat and its fish community. Although the results of these studies are ambiguous, some habitat variables have frequently been found to correlate with one or more aspects of the fish community. Several of these habitat variables, including depth, the structural complexity of the substrate, and live coral cover, are possible to estimate with remote sensing data. In this study, we combine a set of statistical and machine-learning models with habitat variables derived from IKONOS data to produce spatially explicit predictions of the species richness, biomass, and diversity of the fish community around two reefs in Zanzibar. In the process, we assess the ability of IKONOS imagery to estimate live coral cover, structural complexity and habitat diversity, and we explore the importance of habitat variables, at a range of spatial scales, in the predictive models using a permutation-based technique. Our findings indicate that structural complexity at a fine spatial scale (~5 to 10 m) is the most important habitat variable in predictive models of fish species richness and diversity, whereas other variables such as depth, habitat diversity, and structural complexity at coarser spatial scales contribute to predictions of biomass. In addition, our results demonstrate that complex model types such as tree-based ensemble techniques provide superior predictive performance compared to the more frequently used linear models, achieving a reduction of the cross-validated root-mean-squared prediction error of 3–11%. Although aerial photographs and airborne lidar instruments have recently been used to produce spatially explicit predictions of reef fish community variables, our study illustrates the possibility of doing so with satellite data. The ability to use satellite data may bring the cost of creating such maps within the reach of both spatial ecology researchers and the wide range of organizations involved in marine spatial planning.

1. Introduction

Local communities of coral reef fishes are influenced by a complex mix of global biogeographic patterns (Harrison & Cornell, 2008), stochastic recruitment pulses (Sale, 1977), and interactions between the fishes and their habitat operating at a range of spatial scales. This multi-scale complexity presents a challenge for the development of spatially explicit predictive models of the fish community, which could greatly aid spatial planning and conservation management of coral reef environments (Mellin et al., 2009; Pittman et al., 2009). Despite the complexity, field studies have established statistical relationships between several aspects of the local habitat and its fish community. The structural complexity of the substrate positively correlates with fish species richness, abundance and biomass, while live coral cover influences fish abundance, particularly of species directly dependent on the corals for food or shelter (Jones et al., 2004). Depth influences the composition and abundance of the fish community (Lara & González, 1998), and habitat complexity positively correlates with species richness (Pittman et al., 2007a; Purkis et al., 2008). In addition, location influences the fish community through factors such as proximity to the reef edge or nearby rivers (Friedlander & Parrish, 1998) and proximity to nursery habitat such as seagrass beds and mangrove stands (Dorenbosch et al., 2005; Mumby et al., 2004a). The strength of the specific statistical relationships varies between studies, as does the spatial scale at which the relationships are strongest (Mellin et al., 2009). Nevertheless, the existing body of field studies suggests a list of habitat variables that, if estimated by remote sensing instruments, would enable the development of spatially explicit predictive models of fish communities (Knudby et al., 2007). Five of these variables deserve further attention as they are as possible to map with remote sensing: depth, structural complexity, substrate type, habitat diversity, and live coral cover. Methods for deriving water depth from both passive optical (Lyzenga, 1978; Maritorena et al., 1994; Stumpf et al., 2003) and lidar

* Corresponding author. Tel.: +1 613 992 8348.
E-mail address: knudby@gmail.com (A. Knudby).

0034-4257/$ – see front matter
doi:10.1016/j.rse.2010.01.007

remote sensing (Guenther, 2001) are well established, although the maximum depths at which both methods can be applied are limited by water turbidity. Digital elevation models can be created from such spatially distributed depth measurements (Storlazzi et al., 2003), and measures of structural complexity can be derived (Pittman et al., 2009). Such measures of structural complexity have already been used to produce spatial predictive models of fish species richness (Pittman et al., 2007b). Substrate maps are the standard products from remote sensing of coral reefs (Mumby et al., 1997), and have been combined with depth to produce spatial predictions of fish community variables, both on coral reefs (Mellin et al., 2007) and elsewhere (Herold et al., 2007). Habitat diversity measures are straight-forward to calculate from substrate maps, but their relationship with fish community variables has been little explored. One study has found a significant relationship with fish species richness (Purkis et al., 2008), while others have found the same relationship to be non-significant (Pittman et al., 2004, 2007b). Live coral cover has been more difficult to map with remote sensing, because of sub-pixel heterogeneity and high spectral similarity between corals and other substrate types such as algae and seagrass (Hochberg et al., 2003). This causes live coral cover estimates to be highly dependent on conditions such as water depth, turbidity, and the presence/absence of spectrally similar substrate types (Hochberg & Atkinson, 2003; Mumby et al., 2004b). Nevertheless, a few studies have had some success with mapping live coral cover empirically, using either spectral indices (Joyce, 2004) or classification schemes based on the percentage of live coral cover (Isoun et al., 2003; Newman et al., 2007). Despite field studies having demonstrated relationships between live coral cover and fish community variables, remotely sensed live coral cover has not yet been used for predictive mapping of fish community variables.

In this study, we derive remotely sensed estimates of depth, structural complexity, habitat diversity and live coral cover. We derive these estimates at a range of spatial scales and use them, along with a substrate classification, a map of geomorphological reef zones, and a map of boundaries of a local marine protected area, to develop spatial predictive models for fish species richness, diversity, and biomass. We compare the predictive performance (error) of six statistical and machine-learning techniques used to model the fish–habitat relationships, and we assess the influence of each predictor variable, and its spatial scale, on model predictions. Finally, we use the best overall model type to produce maps of each fish community variable, and discuss the application of such maps for marine spatial planning.

2. Methods

2.1. Study area

Our study site covers the reefs around Chumbe Island (3.4 km^2) and Bawe Island (16 km^2), two raised Pleistocene reefs located in Zanzibar, Tanzania. Chumbe Island is located 12 km from Zanzibar town, a distance that historically has limited fishing pressure. The fringing reef on the western side of the island has been effectively protected as a marine park since 1994 (Muthiga et al., 2000), whereas the lagoon and scattered coral and seagrass areas on the other sides of the island are open to fishing. The reef around Bawe Island experiences substantial fishing pressure, as it is located only 5 km from Zanzibar and is legally open to fishing. The location of the two reefs, and the IKONOS imagery covering each, is shown in Fig. 1. The tidal range in the area is 4.2 m at spring tides.

2.2. Field data

This study relies on three distinct data sets: fish data, habitat data, and two IKONOS satellite images. All data were collected from mid-September to mid-December 2007, except the satellite image of Bawe Island which is from October 2005.

The fish community was surveyed at a total of 144 sites on the two reefs. For each reef, sites were stratified by geomorphologic zone and substrate type, with the number of sites in each stratum roughly proportional to its observed variance in fish species richness, biomass and diversity. The sampling strategy thus aimed at covering the feature space — i.e. the existing combinations of environmental situations — as far as possible, which is required for predictive modeling but came at the cost of some degree of spatial clustering in heterogeneous areas, such as along the edges of the reefs. The exact location of each site around Chumbe Island was determined by snorkelling in a random direction for a random number of fin kicks (minimum 50) from the previous site. Around Bawe Island, logistical constraints lead to a clustered sampling design, with points in each cluster distributed across the environmental gradient from land to the reef edge. Sites shallower that 8 m were surveyed while snorkelling; SCUBA equipment was used for deeper sites.

Upon arrival at a site, the observer (AK) would wait, passively, for a period of 5 min. Subsequently, the fish community was surveyed using the point count method of Bohnsack and Bannerot (1986), with the radius of observation limited to 5 m due to the combination of limited visibility and the need to identify small fish to the species level. While recording data, the observer rested on the surface (or on the bottom if using SCUBA) at the centre of the site. Observations were separated into two five-minute intervals. For the first five-minute interval, during which the observer slowly rotated to look in all directions, all fish species with individuals of >5 cm fork length, observed within a radius of 5 m of the centre of the site, were noted on a dive slate. During the second five-minute interval, the number and average fork length of individuals was noted for each species. If a species was observed as present during the first 5 min but could not be found during the second 5 min, the number and average fork length was retrieved from memory. Species observed only during the second five-minute interval were ignored.

These data were used to calculate species richness, biomass, and diversity for each site. Species richness values follow from a simple summation of the species seen at a given site. Biomass was derived from individual fish lengths using the formula Biomass $= A\,L^B$, where L is fish fork length, and A and B are constants that depend on fish shape. For each species, values of A and B published on Fishbase (Froese & Pauly, 2009) were used if available; values for the genus were used if species-specific values have not been published. Total biomass values for each site were log-transformed before analysis in order to reduce the influence of positive outliers caused by passing schools of large-bodied fishes. Diversity was calculated using Shannon's diversity index (Shannon & Weaver, 1963). Calculations were based on biomass per species rather than abundance, in order to reduce the bias caused by occasional large schools of small-bodied *Chromis* sp. fishes and provide a more ecologically relevant diversity measure (Wilhm, 1968).

Habitat data were collected at each site ($n = 144$) following the fish surveys, and at additional sites ($n = 583$) for use in image classification. The addition habitat sites were determined based on local knowledge of the area, in order to cover all major substrate types and geomorphologic zones. The maximum and minimum depths within a site were measured with a dive computer resting on the substrate, and used to calculate average depth and depth range for each site. Five orthogonal substrate photos from each site, taken at a distance of approximately 2 m from the substrate and each covering approximately 4 m^2, were processed in the software CPCe (Kohler & Gill, 2006) to derive the percentage cover of live coral and the dominant substrate type used in image classification. In addition to these data, 1015 depth measurements were made around the two islands, also using a dive computer resting on the surface. All field sites were geolocated using a tracking GPS towed in a water-proof bag, time-synchronized with either the substrate photos or times noted for each depth measurement.