Oecologia (2020) 194:283–298
https://doi.org/10.1007/s00442-020-04753-2

CONSERVATION ECOLOGY – ORIGINAL RESEARCH

Application of machine learning algorithms to identify cryptic reproductive habitats using diverse information sources

Jacob W. Brownscombe[1,2] · Lucas P. Griffin[3] · Danielle Morley[4] · Alejandro Acosta[4] · John Hunt[4] · Susan K. Lowerre-Barbieri[5,6] · Aaron J. Adams[7,8] · Andy J. Danylchuk[3] · Steven J. Cooke[1]

Received: 25 March 2020 / Accepted: 8 September 2020 / Published online: 1 October 2020

Abstract

Information on ecological systems often comes from diverse sources with varied levels of complexity, bias, and uncertainty. Accordingly, analytical techniques continue to evolve that address these challenges to reveal the characteristics of ecological systems and inform conservation actions. We applied multiple statistical learning algorithms (i.e., machine learning) with a range of information sources including fish tracking data, environmental data, and visual surveys to identify potential spawning aggregation sites for a marine fish species, permit (*Trachinotus falcatus*), in the Florida Keys. Recognizing the potential complementarity and some level of uncertainty in each information source, we applied supervised (classic and conditional random forests; RF) and unsupervised (fuzzy *k*-means; FKM) algorithms. The two RF models had similar predictive performance, but generated different predictor variable importance structures and spawning site predictions. Unsupervised clustering using FKM identified unique site groupings that were similar to the likely spawning sites identified with RF. The conservation of aggregate spawning fish species depends heavily on the protection of key spawning sites; many of these potential sites were identified here for permit in the Florida Keys, which consisted of relatively deep-water natural and artificial reefs with high mean permit residency periods. The application of multiple machine learning algorithms enabled the integration of diverse information sources to develop models of an ecological system. Faced with increasingly complex and diverse data sources, ecologists, and conservation practitioners should find increasing value in machine learning algorithms, which we discuss here and provide resources to increase accessibility.

Keywords Marine biology · Spawning aggregations · Ecology · Conservation · Machine learning

Introduction

Remote measurement of the biotic and abiotic conditions of natural ecosystems provides essential insights in the fields of ecology and conservation. However, collecting information

Communicated by Yannis Papastamatiou.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00442-020-04753-2) contains supplementary material, which is available to authorized users.

✉ Jacob W. Brownscombe
jakebrownscombe@gmail.com

[1] Fish Ecology and Conservation Physiology Laboratory, Department of Biology, Carleton University, 1125 Colonel by Drive, Ottawa, ON K1S 5B6, Canada

[2] Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax, NS B4H 4R2, Canada

[3] Department of Environmental Conservation, University of Massachusetts Amherst, 160 Holdsworth Way, Amherst, MA 01003, USA

[4] Florida Fish and Wildlife Conservation Commission, 2796 Overseas Highway, Suite 119, Marathon, FL 33050, USA

[5] Florida Fish and Wildlife Conservation Commission, Fish and Wildlife Research Institute, 100 8th Avenue Southeast, St. Petersburg, FL 33701, USA

[6] Fisheries and Aquatic Science Program, School of Forest Resources and Conservation, University of Florida, 7922 Northwest 71st Street, Gainesville, FL 32653-3071, USA

[7] Bonefish and Tarpon Trust, 135 San Lorenzo Ave., Suite 860, Coral Gables, FL 33146, USA

[8] Florida Atlantic University Harbor Branch Oceanographic Institute, 5600 North Highway A1A, Fort Pierce, FL, USA

on wild organisms in natural ecosystems presents exceptional logistical and analytical challenges due to technological limitations of study methods, uncontrolled conditions, complex interactions between biotic and abiotic factors, biases and uncertainties inherent to measurement, and data deficiencies. Scientific advances are increasing our capacity to study wild animals and ecosystems using tools such as telemetry (Cagnacci et al. 2010; Hussey et al. 2015), stable isotopes (West et al. 2006; Michener and Lajtha 2008), genetics/genomics (Manel et al. 2003; Hanski and Gaggiotti 2004), as well as access to local ecological knowledge (Menzies 2006; Brook and McLachlan 2008). Yet, our knowledge of natural ecosystems is still incomplete, and often suffers from various biases and uncertainties (Ludwig et al. 1993; Ascough et al. 2008; Fulton et al. 2019), including new challenges related to translating complex data collection techniques and datasets into actionable knowledge (Young et al. 2013; Nguyen et al. 2018). Cohesive and comprehensive data collection and analysis techniques; therefore, play an essential role in converting data into actionable knowledge (Nguyen et al. 2017). There is great potential to overcome the data deficits and uncertainty (Green et al. 2007; Michener and Jones 2012) and avoid analytical errors that can result in inaccurate interpretations and ineffective conservation actions (Johnson 1981; Zuur et al. 2010).

Increasingly complex datasets on ecosystems necessitate the development and application of sophisticated analytical techniques. Emerging from the field of machine learning, various statistical learning algorithms (i.e., machine learning algorithms) have gained popularity for modeling ecological systems in the recent decades due to their ability to handle diverse data types, large numbers of correlated predictors, and model complex hierarchical relationships with high predictive accuracy relative to other statistical techniques (Cutler et al. 2007; Elith et al. 2008; Olden et al. 2008). Some algorithms are particularly well suited for dealing with challenges in ecology, such as 'large p small n' problems, where there are a large number of potential predictors relative to the sample size (Oppel et al. 2009). Statistical learning algorithms are highly numerous and diverse (see Kuhn et al. 2019 for some examples) and are generally implemented in two ways—supervised or unsupervised. In supervised learning, the model is trained to predict a known response variable, for example, to identify specific animal behaviors from positioning or sensor data based on direct observations of tagged animals (Guilford et al. 2009; Brown et al. 2013). With unsupervised learning there is no response variable specified; algorithms identify patterns or groupings in the data based on its structure alone, for example, to explore biological community dynamics (Chon et al. 2000). These statistical learning algorithms are often applied with a focus on prediction accuracy, and while predictor importance scores can often be generated, they do not typically provide statistical *p* values to indicate 'significant predictors', which ecologists often focus on, sometimes to their peril (Halsey 2019). Olden et al. (2008) suggested it is perhaps this philosophical difference in analytical approach that has constrained the application of these algorithms in ecology. Yet, these approaches are becoming more accessible (Elith et al. 2008; Christin et al. 2019) and increasingly applicable to large and complex ecological datasets. There may be some confusion about machine learning terminology because there is a distinction between what are often referred to as machine learning algorithms (i.e., a specific set of statistical learning techniques), and the field of machine learning, which focuses on the application of algorithms or statistical models to perform a task (e.g., robotic movements, pattern recognition, data prediction) without specific instruction. This distinction is illustrated by the fact that logistic regression is commonly utilized as a statistical model in the field of machine learning but is not generally referred to as a machine learning algorithm.

Across a large range of animal taxa, certain biological events such as mating or spawning disproportionately shape populations, food webs, and ecosystems (Nowlin et al. 2008; Archer et al. 2015; Mourier et al. 2016). In aquatic ecosystems, there is growing recognition that protection of fish spawning sites, particularly spawning aggregations, is of major importance due to the role these sites play as productivity and biodiversity hotspots (Sadovy and Domeier 2005; Sadovy De Mitcheson and Colin 2012; Lowerre-Barbieri et al. 2016; Erisman et al. 2017; scrfa.org). Fish spawning aggregations are also highly vulnerable to fisheries overexploitation due to the concentration of individuals at predictable locations and times (Sala et al. 2001; Erisman et al. 2017). Population depletion also often occurs cryptically with aggregate spawning species (Erisman et al. 2011). For these reasons, protection of spawning fish and their habitats is a major focus of aquatic conservation efforts. However, identifying critical spawning locations presents many challenges related to monitoring mobile animals in expansive aquatic systems such as open oceans. Efforts to identify fish spawning sites can be informed by diverse sources, commonly from local ecological knowledge accrued from years of interactions between local people and the ecosystem (Silvano et al. 2006; Hamilton et al. 2012), through direct scientific studies using techniques such as individual tracking via telemetry (Zeller 1998; Danylchuk et al. 2011; Crossin et al. 2017; Binder et al. 2018; Brownscombe et al. 2020), or a combination of these approaches (e.g., Adams et al. 2019).

One such aggregate spawning marine fish species is the permit (*Trachinotus falcatus*), which lives throughout the tropical and subtropical western Atlantic Ocean, Caribbean Sea, and Gulf of Mexico, supporting popular recreational fisheries in many locales (Adams and Cooke 2015). Permit commonly feed on invertebrates in shallow

nearshore flats, moving further offshore to form spawning aggregations in proximity to reef promontories (Graham and Castellanos 2005; Bryan et al. 2015; Brownscombe et al. 2020). Gonadosomatic data suggest permit have an extended reproductive period in south Florida from March to September, potentially indicating multiple batch spawning (Crabtree et al. 2002). Permit are commonly targeted by recreational anglers at spawning aggregation sites, but harvest is limited by fisheries regulations, especially in the southernmost part of Florida in the Special Permit Zone where harvest is prohibited from April through July (Brownscombe et al. 2019a; www.myfwc.com/fishing/saltwater/recreational/Permit/). However, catch-and-release fishing of spawning aggregations is common, and may not be sustainable in some locations due to high rates of shark depredation (Holder et al. 2020).

As an aggregate spawning fish that supports popular recreational fisheries, the identification of permit spawning sites is of major importance for their conservation.

Further, spawning locations also dictate larval distribution patterns and population connectivity (Lowerre-Barbieri et al. 2017), as well as the spatial dynamics of predator–prey interactions (De Mitcheson and Colin 2012) relevant to biological community and ecosystem dynamics. Unlike many other aggregate spawning species, where aggregation sites are better studied and are known to occur at predictable locations (Kobara and Heyman 2008, 2010), less is known about permit aggregation sites. This may be in part because permit are wary of boats and divers, and rarely observed underwater. Despite this elusiveness, anglers commonly target permit spawning aggregations by visually observing them from boats or using advanced sonar technology (Holder et al. 2020). There is at least one known permit spawning aggregation site in south Florida, but tracking data indicate permit occupy numerous locations throughout the region during the spawning period, some of which may also be spawning sites (Brownscombe et al. 2020). Over the course of tagging permit to collect said tracking data, we also acquired hundreds of visual observations of permit behavior over space and time, including likely prespawning schooling behaviors. Yet, without clear evidence of spawning activity (e.g., observations of direct spawning or collection of released eggs), the links between behavioral observations, tracking data, and permit spawning are not concrete. Recognizing uncertainty and potential biases in these information sources, we applied a combination of supervised and unsupervised machine learning techniques to identify potential permit spawning sites in proximity to the Florida Keys. These findings may be of relevance to spatial management of fisheries and marine habitats, and also improve accessibility to tools (including R code; Appendix III) and insights for the application of machine learning to other ecological data.

Methods

Fish tracking

Acoustic telemetry was used to track permit movement patterns and space use. In August 2015, an array of 60 acoustic receivers (VR2W, Vemco Inc, Halifax, NS, Canada) was established in nearshore regions of the Florida Keys. Additional receivers were added to the tracking system over time, totaling 84 by May 2018 (See Appendix I: Fig. 1 for locations). Receivers were moored to the substrate via attachment to a 1 m length rebar extending from 30–50 kg cement bases, placed adjacent to popular permit fishing locations, informed by consultation with local fishing guides. This acoustic receiver array complimented others established in the region in 2014–2016 to track a diverse range of fish species, providing extensive coverage along the Florida Reef Tract. In addition to these arrays, another 1000 + receivers were maintained by numerous research groups throughout the southeastern United States, with data sharing facilitated through the FACT, integrated Tracking of Animals in the Gulf of Mexico network, and the Ocean Tracking Network (Appendix I: Fig. 1). Receiver coverage was variable amongst locations, which may have had an influence on detectability over space. These acoustic receiver arrays were used to track permit with internally implanted acoustic transmitters (V13-1x; high power, 80–160 s delay, 653 day life, 6.2 g in water, Vemco Inc), V13A-1x (low power, 80–160 s delay, 355 day life, 6.2 g in water, Vemco Inc), or V16-4x (high power, 60–120 s delay, 1910 day life, 11.7 g in water, Vemco Inc) from Mar 2016 to May 2019. Tagging locations spanned throughout the Florida Keys region (Appendix I: Fig. 2); see Brownscombe et al. (2020) for details on permit capture and tagging. All procedures were conducted in accordance with the Carleton University Animal Care Committee (application 11,473), as well as the American Association for Laboratory Animal Science (IACUC protocol 2013-0031, University of Massachusetts Amherst).

Visual surveys

Over the course of permit tagging efforts from Aug 2015 to May 2019, the research team spent over 2000 h searching for permit, successfully finding, visually observing, and capturing permit at 78 general locations throughout the study region. Efforts were concentrated in the months of March–October, which encompasses the primary spawning period (Crabtree et al. 2002; Brownscombe et al. 2020). The entire spatiotemporal structure of sampling was informed by

working closely with local fishing guides, especially those in the Lower Keys Fishing Guides Association and the Florida Keys Fishing Guides Association. Researchers were either fishing directly with fishing guides or acting on shared information about fishing opportunities. Visual surveys were therefore structured largely by local ecological knowledge.

Although some permit was located with sonar, the vast majority of successful permit tagging involved visual observation of groups of permit from a boat, sight casting to them via fishing rod and reel. During this process, distinct permit behaviors were observed over space and time. In shallow water flats, permit were observed either: (1) foraging in the benthos alone or in small groups (< 20 individuals); (2) moving in or out of the flats alone or in small groups; or (3) floating in channels or nearshore natural or artificial reef structures (2–7 m water depth) in small to large groups (10 to ~ 100 individuals). In deeper water sites > 5 m, permit were observed exclusively in proximity to natural or artificial reef structures, generally in larger groups of 20 to > 500 individuals (roughly estimated). In these habitats, permit were either: (4) floating in currents or moving slowly and independently with large spacing (> 3 m) between them; or (5) in tightly aggregated (~ 1 to 2 m spacing) groups of > 50 individuals either floating in currents or moving rapidly in a highly coordinated manner (see Appendix II for video of this behavior).

Of the five distinct behaviors observed and described above, behavior 5) is consistent with those associated with direct observations of permit spawning in Belize (Graham and Castellanos 2005) and prespawning behavior in the Dry Tortugas (Bryan et al. 2015). Aside from spawning, another major driver of fish schooling behavior is the presence of predators (Pitcher 2001). Behavior 5) was observed both with and without predators (large sharks) in the vicinity. Spawning aggregations often attract predators due to the high density of relatively stationary prey (Sancho et al. 2000; De Mitcheson and Colin 2012). In many cases, permit was observed exhibiting behavior 5) with no predators present, then predator(s) subsequently arrived, at which time the schooling behavior remained consistent but movements became more rapid. In these cases, the observed permit never strayed more than ~ 500 m from the putative point of attraction, which was either a natural or artificial reef promontory. The potential reasons why these types of locations may be popular spawning locations are discussed in Brownscombe et al. (2020). Lastly, behavior 5 was observed exclusively in the months of March–June, which coincides with peak gonad development in permit in this region (Crabtree et al. 2002). In many locations, behavior 5 was observed in those aforementioned spawning months, but at other times in July–October, behavior 4 was observed. Owing to the combination of empirical information on gonad development, empirical behavioral observations of prespawning and spawning, and the observations made here, behavior 5) was interpreted as probable prespawning behavior. Sites were therefore categorized into those where prespawning behaviors were observed, sites where only nonprespawning behaviors were observed, or sites where no behavioral observations were available.

Data analysis

Fish tracking data consisted of 1.53 million detections of 112 individual permit (682 ± 96 mm fork length; mean ± SD; 474–978 mm range). The majority of these individuals were likely mature (50% maturity at 486 mm in males and 547 mm in females; Crabtree et al. 2002). Tracking durations were variable amongst individuals (286 ± 177 days; mean ± SD, 1–658 day range; Appendix I: Fig. 3). Detections occurred at 205 acoustic receivers, which were grouped into 43 spatially distinct sites, or receiver nodes. To reduce the potential for false detections (see Simpfendorfer et al. 2015), detections were filtered in a two-stage process. First, detections that occurred prior to tag deployment were removed, as were duplicates that were spaced within a time period less than the minimum tag delay (60 s) at the same receiver. In the second stage, detections were removed if they consisted of a single detection at given site (i.e., receiver node) within a 2-h time period. Once filtered, total detections were calculated at each site, along with the number of detections during the known spawning based on gonad development period from March to July (Crabtree et al. 2002). These data were also used to calculate individual-level residency periods at each site, where permit was considered resident at a site when < 1 h lapsed between detections. Total residency, mean residency period (i.e., the mean period of time spent at the location per visit), the number of unique visits (i.e., individual residency periods) and the total number of individuals visiting each node were calculated. Site degree was also calculated as the number of other unique sites each site was connected to via permit movements, site strength was calculated as the total number of movements between each site and all other sites, and edge weights were calculated for each pair of connected sites as the total number of movements between them. All environmental and fish tracking data for each of the 43 sites surrounding the Florida Keys is reported in Appendix I: Table 1; however, generic site names are used to avoid any potential misuse of data to target spawning permit via fishing. Specific site information can be made available upon reasonable request.

Of the 43 sites at which permit were detected in proximity to the Florida Keys, visual observations of permit behaviors were available at 27. Although there is strong empirically based rationale for the association of behavior 5) with spawning, without direct observations or extensive site monitoring (aside from presence/absence from acoustic

detection data), these observations are not concrete evidence of spawning locations. Therefore, a combination of statistical learning approaches was applied to identify potential permit spawning sites in the Florida Keys using available data including environmental characteristics, permit tracking metrics, and visual behavioral observations. All analysis were conducted with R (R Core Team 2018) via RStudio (RStudio Team 2016), and all R code for analyses can be found in Appendix III.

Supervised learning

Operating on the assumption that visual observations of permit prespawning behaviors are reliable evidence of spawning sites, we aimed to develop a predictive model of the presence/absence of this behavior to make predictions about sites where no visual observations were available, but permit were detected via acoustic telemetry. We identified a range of potential metrics that could be related to permit spawning, including environmental characteristics (habitat type, water depth) and fish tracking metrics (number of detections, total residency, mean residency period, number of unique visits by individuals, number of individuals visiting the site, site degree, and site strength; see "Data analysis" section above). Unfortunately, there was a large number of predictors relative to sample size and the majority of these metrics are correlated with each other, display substantial heterogeneity and some outliers (Appendix I; Fig. 4, 5), posing challenges for developing a predictive model. Initial attempts to utilize logistic regression to model the presence/absence of prespawning behavior revealed that cross-validation (i.e., separating data into training and test sets to assess model predictive performance to ensure the model makes generalizable predictions and is not overfitting the training data) was not possible with any models with complexity beyond a single predictor due to convergence failure (although there are more advanced methods to implement logistic regression in similar scenarios; e.g., Piironen et al. 2020). This is indeed a 'large p small n' problem, that is, a large number of potential predictors relative to sample size, which occurs frequently with ecological questions and datasets.

To overcome the above-discussed analytical challenges, we applied decision tree algorithms to predict permit prespawning behavior. Basic versions of these models are classification and regression trees (CART), which generate recursive binary partitions in the data to optimize prediction of the response (Breiman et al. 1984; De'Ath and Fabricius 2000). This approach is fundamentally different from frequentist models like regression in that CART, like many learning algorithms, does not require a defined relationship between the predictors and response, rather it uses an algorithm to learn the relationship. These models have many advantages; they handle a wide range of variable types, are not majorly affected by variable distributions, outliers, missing data, or monotonic variable transformations, can model complex hierarchical interactions, and handle high-dimensional datasets (Olden et al. 2008). They are also highly tractable, providing intuitive and interpretable outputs. It is likely these qualities that have attracted ecologists to versions of these models (Cutler et al. 2007; Elith et al. 2008; Olden et al. 2008). Caveats of these models can include data overfitting, predictor biases, and poor performance on some types of datasets. Overcoming many of these caveats, random forests (RF) are a relatively recent extension (patented in 2009) of decision trees that fit numerous trees (often thousands) with random data subsampling and variable selection via bootstrapping and aggregate the trees via bagging to optimize the prediction of the response (Breiman 2001). RF can be biased towards utilizing certain types of predictors and correlated predictors (Strobl et al. 2007). These biases can be overcome with conditional RF (Strobl et al. 2009), which build ensembles of conditional inference trees, using statistical criteria to determine data partitions (Hothorn et al. 2006). Fitting conditional RF that subsample without replacement and a conditional permutation scheme enables unbiased predictor selection and importance measures (Strobl et al. 2008).

We applied both classic RF and conditional RF to observations of permit prespawning behavior (binary categorical) with the site-specific environmental and fish tracking predictions stated above. RF were fit with replacement and 1000 trees, although > 500 trees produced stable outputs (Appendix I; Fig. 6). The number of variables subsampled at each partition (mtry) was tested at every value; classic RF were sensitive to variation in mtry, but conditional were not (Appendix I; Fig. 7). Classic RF were therefore fit with mtry = 2; for conditional RF the default of the square root of the number of variables. Model performance was assessed based on prediction accuracy, accuracy balance between classes, Cohen's Kappa, sensitivity (proportion of presences accurately classified), and specificity (proportion of absences accurately classified) in out-of-bag samples, i.e., data not used for training each subsampled decision tree. Predictor importance was assessed for classic RF using permutation importance (Mean Decrease in Accuracy), while conditional permutation importance was used for conditional RF, calculated from out-of-bag data. To assess how the models were making their predictions, partial dependencies were calculated for the top predictors, which represent predicted values of the response (expressed as a probability in this case) across levels of each predictor while holding other predictors at their mean (i.e., the marginal effects of the predictor). These models were then used to predict on to sites that lacked visual observations of permit behavior, but tracking and environmental data were available. Classic RF were fit with the randomForest R package (Liaw and Wiener

2002) and conditional RF the cforest function in the partykit package (Hothorn and Zeileis 2015). Model performance metrics were calculated with the confusionMatrix function in the caret package (Kuhn et al. 2019).

Unsupervised learning

The supervised learning approach employed above operates on the assumption that visual observations are reliable evidence of potential permit spawning sites, using these observations as a primary basis for developing a model to predict additional potential sites in the region based on environmental and tracking information. Despite the fact that these visual observations are similar to those of permit spawning in other locations (Graham and Castellanos 2005; Bryan et al. 2015), without direct observations of actual permit spawning at these sites or continuous monitoring of sites for this behavior, it is not concrete evidence. For example, it is possible that permit were forming prespawning aggregations prior to moving to a nearby spawning location. Recognizing this uncertainty, an unsupervised learning algorithm was applied to explore potential clusters of sites with unique characteristics that may be indicative of spawning sites independent of observations of permit behavior. Specifically, fuzzy k-means (FKM) was applied using the 'fclust' package (Ferraro and Giordani 2015). Similar to most clustering algorithms, the FKM algorithm aims to identify a limited number of homogeneous groupings in the data. However, FKM accommodates both numerical and categorical predictors, and allows for some uncertainty in data clustering, assigning a probability of each data point belonging to each cluster (Hoppner et al. 2000). Variables included in FKM included habitat type, water depth, total permit residency, and mean permit residency period. All available model performance criterion were considered to identify the optimal number of unique clusters, including the partition coefficient, modified partition coefficient, partial entropy, silhouette, fuzzy silhouette, and Xi and Beni. The optimal number of clusters was considered that which the majority of performance criteria agreed upon.

Results

Permit were detected via acoustic telemetry throughout the broader Florida Keys region, moving frequently amongst shallow water flats habitats, nearshore, and offshore reefs, both natural and artificial (Fig. 1). There was high connectivity amongst the flats of the Lower Florida Keys, the Marquesas and adjacent sites on the Florida Reef Tract and artificial reefs west of the Marquesas. Of the 43 sites permit were detected, 27 had associated observations of permit behaviors

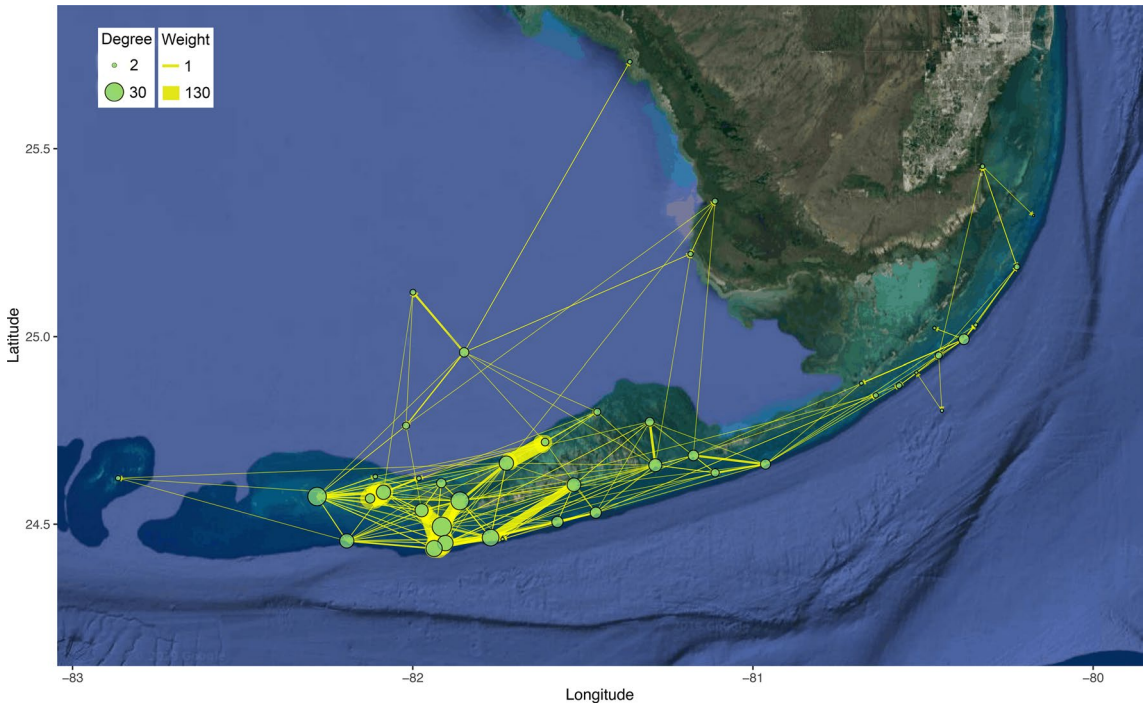

Fig. 1 Network map of permit movement patterns in proximity to the Florida Keys, with node (green circle; $N=43$) size representing its degree (number of unique connections to other nodes) and edge (yellow line) width representing its weight (number of movements between connecting nodes) (color figure online)

(categorized as spawning or nonspawning related). Within these sites, supervised learning via classic RF predicted sites where putative prespawning behavior occurred with 100% accuracy in training data and 85% accuracy in out-of-bag (nontraining) data, significantly higher than the no-information rate (Table 1). Conditional RF were less accurate in training data, at 89%, but had equal out-of-bag accuracy to classic RF and better accuracy balance between response categories (Table 1). The most important predictors, determined via permutation, differed between model types. Classic RF identified the most important predictors as depth, followed by detections during the known spawning period, and residence hours, while conditional RF identified habitat type, water depth, and mean residency period (Fig. 2). Both models agreed the number of individuals detected, number of unique site visits, site degree and strength were the least important predictors. Examining partial dependencies, both model types predict similar relationships between model predictors and permit spawning sites; however, classic RF predicted higher probabilities at greater water depths and outer reef habitats (Appendix I; Fig. 8). Predicting on to sites without behavioral observations, the two types of models produced different results (Fig. 3a, b). Classic RF identified 16 sites as probable spawning locations (9/9 with prespawning observations, 7/16 with no observations, 0/18 with nonprespawning observations; Fig. 4a). Conditional RF identified 9 sites as probable spawning locations (7/9 with prespawning observations, 1/16 with no observations, 1/18 with nonprespawning observations; Fig. 4b).

Unsupervised learning via FKM applied to environmental and permit tracking metrics identified five clusters based on the general agreement amongst the modified partition coefficient, partial entropy, silhouette, and fuzzy silhouette cluster validity indices (Appendix I: Fig. 9). Within these, clusters 2, 3, and 4 were comprises artificial and outer natural reef habitats with high mean permit residency values (Fig. 5). Cluster 5 included entirely natural reefs with moderate permit residency values, and cluster 1 was primarily shallow flats habitats with low residency values (Fig. 5). Comparing these characteristics to empirical knowledge of permit spawning sites, which consist of deeper, outer reef habitats, clusters 1 and 3 were considered improbable spawning locations, clusters 2, 4, and 5 as probable spawning locations. Based on the post hoc interpretation of cluster identity, 18 sites were considered probable spawning locations (9/9 with prespawning observations, 7/16 with no observations, 2/18 with nonprespawning observations; Fig. 3c).

Combining supervised and unsupervised learning by collating ensemble votes and calculating multiplicative probabilities of spawning locations from classic RF, conditional RF, and FKM, there were eight sites with the highest probability of being permit spawning locations. Five of these sites were located along the Florida Reef Tract, one cluster of shallow artificial reefs west of the Marquesas, and two artificial reef sites in the Gulf of Mexico (Fig. 3d). An additional eight sites throughout the region were potential spawning sites, while 30 sites were categorised as low probability. Sites generally had higher probability of being potential spawning locations which were deeper water, artificial or outer reef habitats with a higher mean permit residency period (Fig. 6).

Discussion

The field of ecology is fraught with challenging questions about how ecosystems function and the best approaches to achieve effective biological conservation, often requiring action based on the complex and diverse information sources. The case of identifying permit spawning aggregation sites in the Florida Keys is interesting due to the nature of available data and its applicability to marine conservation. We used a combination of individual tracking-based site use metrics, visual observations guided by local ecological knowledge, and habitat characteristics analyzed with a combination of supervised and unsupervised machine learning algorithms to identify potential permit spawning aggregation sites. With this approach, we aimed to integrate diverse information sources, avoiding reliance solely on any particular information source (e.g., visual observations of

Table 1 Model performance metrics for classic random forests and conditional random forests models predicting potential permit spawning sites in the Florida Keys

Model	Accuracy	Balance	Kappa	Sensitivity	Specificity	NIR	*p* value [Acc > NIR]
Classic random forests							
Training	1.00	1.00	1.00	1.00	1.00	0.67	*< 0.001*
Out of bag	0.85	0.81	0.65	0.67	0.94	0.67	*0.03*
Conditional random forests							
Training	0.89	0.86	0.74	0.78	0.94	0.67	*0.008*
Out of bag	0.85	0.83	0.67	0.78	0.89	0.67	*0.03*

Statistical significance at $p < 0.05$ is indicated by italics

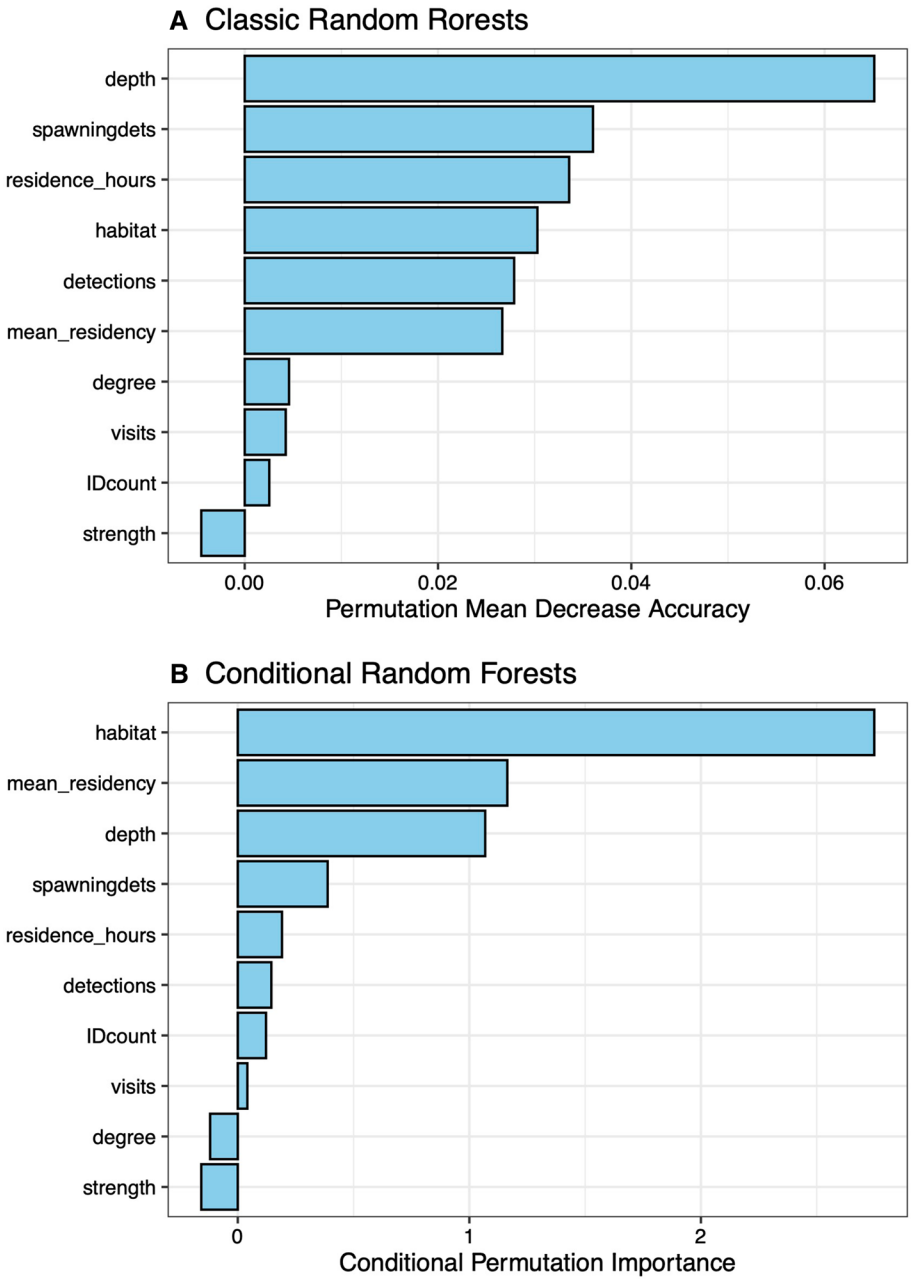

Fig. 2 Variable importance values for **a** classic random forests, and **b** conditional random forests for predictors of sites where permit exhibit prespawning behavior

aggregation behavior) to avoid confounding our findings with any potential biases or uncertainty. Through applying multiple statistical learning approaches to this problem, we have gleaned insights relevant to applied conservation of marine ecosystems and applications of these algorithms in ecology.

Conservation implications

As an aggregate spawning species that is subject to fisheries exploitation, the identification and protection of permit spawning sites is an essential component of their conservation strategy. The sites identified here in both the Atlantic Ocean and the Gulf of Mexico were generally in proximity to deeper water natural and artificial reef structures, where tracking data indicated longer mean residency periods. These characteristics are consistent with the site characteristics and observations of spawning and prespawning behaviors in other locations including Belize and the Dry Tortugas (Graham and Castellanos 2005; Bryan et al. 2015). All models combined identified eight locations with high probability of being spawning sites, and an additional eight with moderate probabilities (Fig. 3d). However, the results are presented from each model type with recognition that end users may decide (perhaps based on the new information in the future) that they trust the assumptions of one model

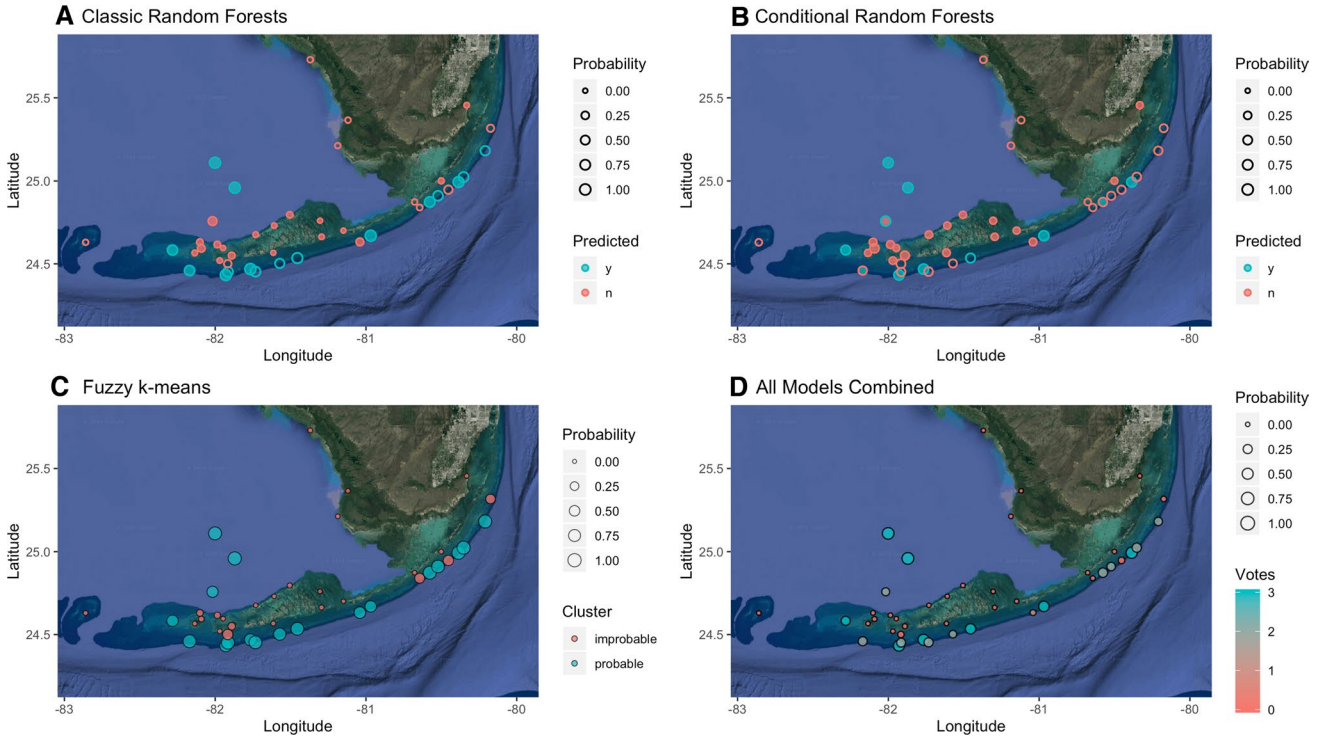

Fig. 3 Sites where permit were observed conducting prespawning behavior (filled blue circle; $N=9$), sites where nonprespawning behaviors were observed (filled red circle; $N=18$), and **a** classic random forests model predictions of spawning sites (open blue circles; $N=16$) on nonspawning sites (open red circles; $N=27$) and **b** conditional random forests predictions of spawning sites (open blue circles; $N=9$) and nonspawning sites (open red circles; $N=34$), **c** sites categorized into probable ($N=18$) and improbable ($N=25$) permit spawning sites based on unsupervised learning (fuzzy *k*-means) cluster assignment and post hoc interpretation of unique grouping representation, and **d** Multiplicative probability of sites representing permit spawning locations using all model predictions combined (from **a**, **b**, **c**) (color figure online)

more than another. For example, if visual observations of prespawning behavior are indeed reliable indicators of spawning sites (supported by Graham and Castellanos 2005; Bryan et al. 2015), the supervised model predictions may be most reliable. These predictions may serve to inform future research approaches, such as structured visual surveys to confirm spawning sites. They may also help to guide efforts to identify permit spawning locations in other locations with popular fisheries, such as Belize, Mexico, and Cuba. Our findings and approach here may be more broadly relevant in locating and characterizing fish spawning aggregations within the framework of the Science and Conservation of Fish Aggregations (https://www.scrfa.org).

The majority of probable permit spawning sites identified here are located within the Special Permit Zone (SPZ), where permit harvest is prohibited from April through July (recently extended; Brownscombe et al. 2019a); however, the northernmost site in the Gulf of Mexico is north of the SPZ, where harvest is allowed year-round. Importantly, even within the SPZ, harvest prohibition alone may not be sufficient to protect permit from fisheries overexploitation because precapture depredation rates can be extraordinarily high at certain locations (Holder et al. 2020). Although there has not been a dramatic collapse in the permit fishery like that of bonefish in the region (Santos et al. 2018; Brownscombe et al. 2019b), fishing guide and angler surveys indicate fishing quality has declined in recent decades (JW Brownscombe, unpublished data). Overfishing at spawning aggregation sites has linked with population declines in many marine fish species (Sadovy and Domeier 2005; Aguilar-Perera 2006; Graham et al. 2008; Erisman et al. 2011), and therefore should be considered a potential threat to the permit fishery and population sustainability. Protection of permit spawning aggregations from overfishing could potentially be accomplished through regulatory or voluntary approaches such as educational programs to promote cultural change (Cooke et al. 2013; Waterhouse et al. 2020). For example, encouraging a social awareness of permit depredation issues could facilitate cultural shifts in angler behavior, with anglers avoiding angling permit at spawning sites when sharks are observed. Such an approach may avoid the potential for conflict by restricting fishery access.

The locations of permit spawning habitats are not only directly relevant to fisheries management strategies, but

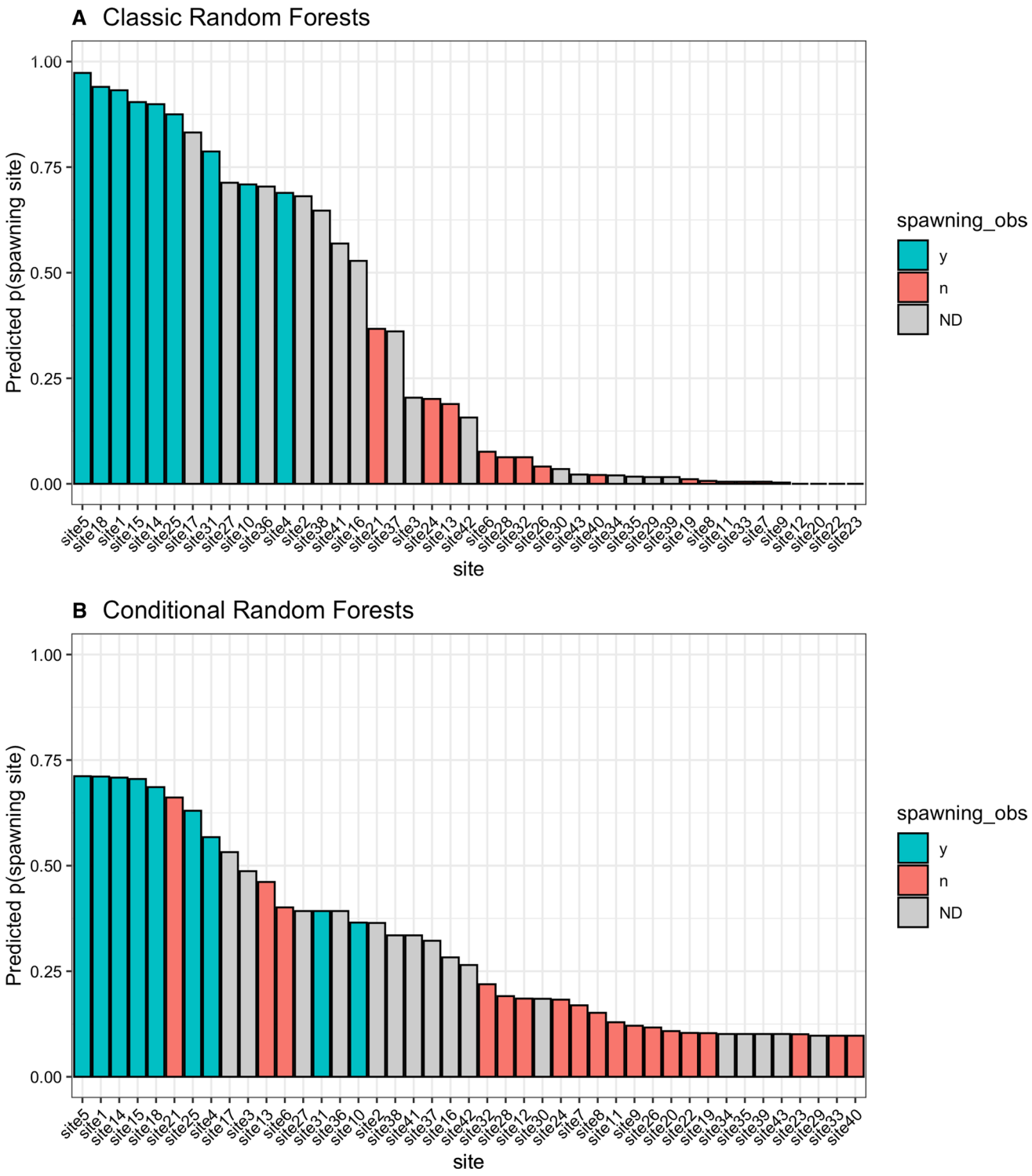

Fig. 4 Sites where permit were detected by acoustic telemetry in proximity to the Florida Keys categorised by whether prespawning observations occurred (y; $N=9$) or not (n; $N=18$) or no observations were available (ND; $N=16$), and predicted probability of being a spawning site by **a** classic random forests, and **b** conditional random forests

also more broadly to population connectivity and ecosystem dynamics. The complex current patterns surrounding the Florida Keys are such that spawning in certain locations, especially those further east along the Florida Reef Tract, may result in poor larval recruitment (Zeng et al. 2018) and hence fish spawning in these locations could represent

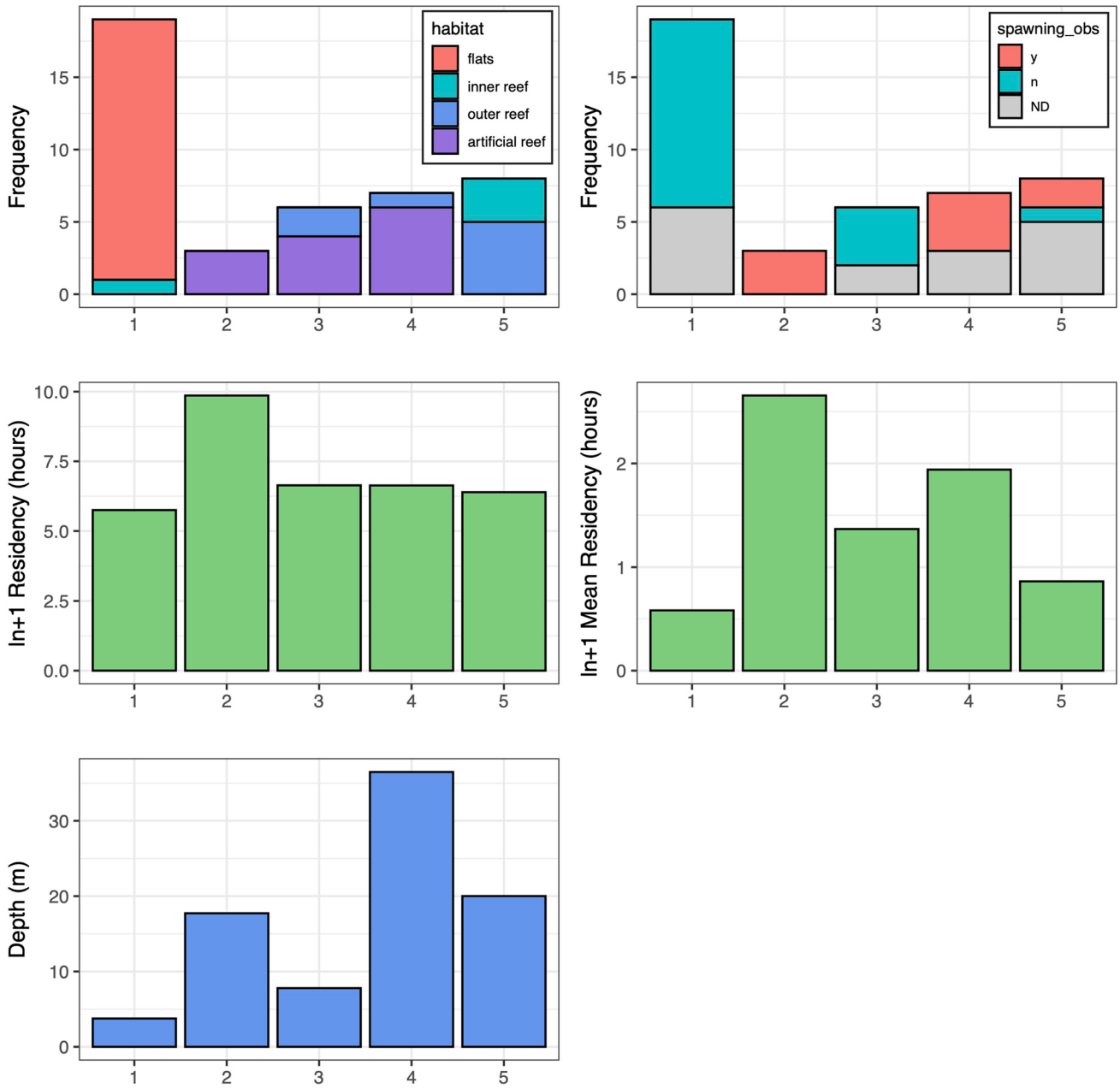

Fig. 5 Characteristics of locations grouped into five unique clusters using a fuzzy *k*-means clustering algorithm

population sinks. Indeed, Bryan et al. (2015) suggest that permit spawning in the Dry Tortugas could be a major source of permit larvae for the Florida Keys, while Adams et al. (2006) found juvenile permit settlement occurred year-round in nearshore habitats of the Florida Keys, which is suggestive of larval inputs from regions outside of the United States of America. Yet, as discussed in Brownscombe et al. (2020), complex current patterns including mesoscale eddies and tidal bores coincide closely with permit spawning activity and may be a mechanism for higher self-recruitment than larval drift models suggest.

Based on our observations and empirical reports from other regions, it appears permit likely spawn on the offshore side of prominent natural or artificial reefs. The mechanisms driving this pattern are worth considering given they may assist in finding additional spawning sites and assess potential impacts of rapid environmental change on permit populations. Permit larval periods typically span from ~ 15 to 20 days prior to settlement into windward sandy beaches, or more seldomly mangrove shorelines (Adams and Blewett 2004; Adams et al. 2006). Although larvae do not settle into reefs, permit spawning locations may enable larvae to

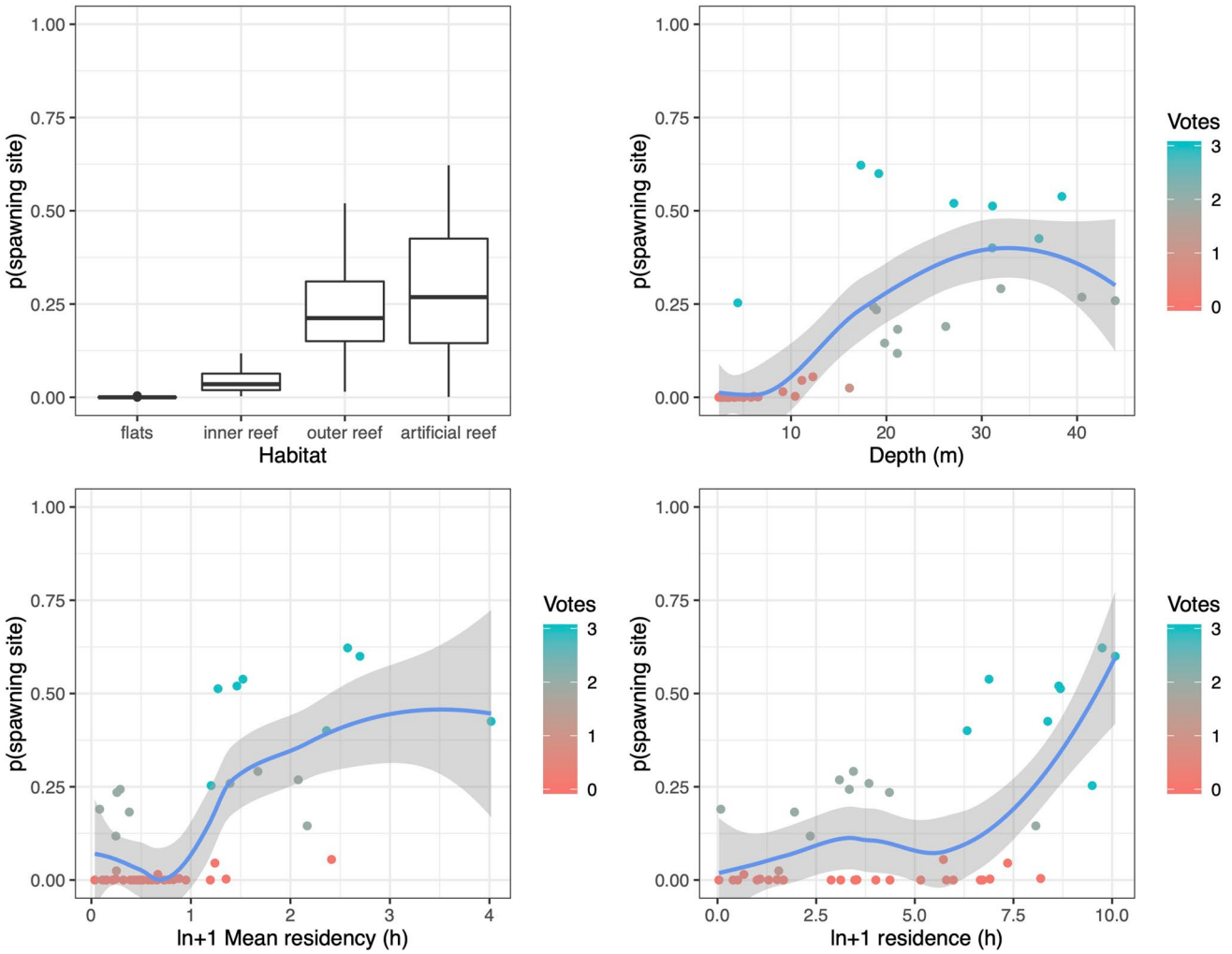

Fig. 6 The relationship between environmental and tracking data-based factors and the multiplicative probability of sites representation permit spawning locations combining supervised random forests models based on environmental data and tracking data, and an unsupervised fuzzy *k*-means clustering algorithm. Votes refers to the number of times a model predicted the location as a spawning location, lines represent a loess smoother. Mean residency reflects the mean period of time spent resident at the site per visit, residence refers to the total residency of permit at the site

navigate toward shorelines using sound or chemical cues from the reef, as is common with reef fishes to recruit into the reef (Leis et al. 2003; Paris et al. 2013). Reef noise is the highest seasonally during the permit spawning period (Brownscombe et al. 2020). Reef promontories may also serve as a 'common meeting point' for adults, serving as a reference that increases aggregation efficacy (Soria et al. 2009).

Applications of machine learning to ecological data

In an absence of direct observations of permit spawning behavior, potential aggregation sites were identified based on a combination of diverse information sources, integrated using multiple modeling techniques. The supervised modeling approaches, using classic and conditional RF operated on the assumption that observations of large aggregations of permit exhibiting schooling behaviors is a reliable indication of a spawning site, which is a logical assumption based on confirmed observations of permit spawning in proximity to reef promontories in Belize (Graham and Castellanos 2005). The use of RF models enabled us to overcome a number of challenges in this dataset including small sample sizes, unbalanced data, and correlated predictors (Breiman 2001; Liaw and Wiener 2002; Cutler et al. 2007) which are problematic for most other commonly employed models such as generalized linear models. However, importantly, small datasets can have higher potential of being unrepresentative of the system due to sampling biases, which are challenging to overcome analytically and must always

be considered in study interpretation. With more complex analytical approaches, interpreting performance and understanding how predictions are generated is essential for reliable application. Global model performance metrics are informative and well-developed (see Kuhn et al. (2019) for a range of available metrics). Beyond overall model accuracy in nontraining data, with classification problems it is important to consider accuracy balance amongst classes. This is especially true when the response variable is imbalanced, as models often favor overall accuracy by default and may neglect the minority class. Ecologists may commonly face this problem with zero-inflated presence/absence datasets, in which there is a particular interest in ensuring presences are accurately classified. This can be accomplished with RF models by including a weighting structure. Model sensitivity and specificity are useful model fit metrics in this regard. Generally speaking, when modeling species distributions, it is more conservative to ensure a model has high sensitivity (accurately classifying presences). There is also a growing suite of model agnostic (i.e., applicable across a range of model types) tools for understanding model function (e.g., see Molnar 2019). Our application was relatively simple, with a small dataset and a limited number of important predictors, hence partial dependencies were sufficient to understand model function. However, in more complex cases, tools such as Local Interpretable Model-agnostic Explanations (LIME; Ribeiro et al. 2016) may be employed to understand more nuanced relationships between predictors and a specific predictions to assess their validity, and also improve upon and choose between classifiers and subsampling regimes.

In the application of these models to identify permit spawning sites, classic and conditional RF identified different partial dependencies and variable importance structures; conditional RF are generally considered more robust in this respect (Strobl et al. 2008). Key predictors of potential reproduction sites included a combination of environmental characteristics (i.e., water depth and habitat type), and tracking metrics (i.e., mean permit residency duration). With the increasing applications of telemetry to track individual fish over expansive space and time, metrics can be developed that indicate certain behaviors and provide greater insights into habitat function and ecological interactions. For aggregate spawning species, it is intuitive that residency period metrics relate to spawning sites; however, fish may reside in a particular area for a variety of reasons (Lowerre-Barbieri et al. 2013). Here we found mean residency period was a better predictor of potential permit spawning sites than total summed residency values, which reflects that fish remained resident for longer periods per visit at spawning sites.

Despite the rational link between observations of large aggregations of schooling fish and spawning sites, we also explored potential permit spawning sites independent of this assumption using unsupervised FKM. This alternate approach requires few assumptions about the data; rather it allows the algorithm to explain the structure of the dataset through identification of unique clusters, or groupings. Advantages of FKM include the ability to include diverse predictor types (unlike k-means, which only accepts continuous predictors), and the fuzzy set assignment of data points, allowing for uncertainty in group assignment (Equihua 1990; Salski 2007). Overall, group assignment generally parsed out sites in a similar manner to the supervised RF approach, but additional inner reef sites were also assigned to probable spawning site groups. This indicates that certain permit space use characteristics (i.e., longer residency periods) were present at these inner reef sites, despite the fact that they do not fit the mold of deeper water structures where spawning aggregation schooling behaviors were observed (Fig. 3).

As ecological datasets continue to grow and become more complex, ecologists should find increasing value in machine learning algorithms for a range of applications. Here we provide an example of application of decision tree and clustering algorithms to a challenging dataset, along with associated R code (see Appendix III), which may help increase the accessibility of these models. There are numerous statistical learning algorithms that ecologists may find useful, others include boosted regression trees (Elith et al. 2008) and neural networks (Christin et al. 2019). In addition to specific applications of machine learning algorithms to individual datasets, they will become an increasing component of diverse aspects of research for ecologists. For example, learning algorithms can be used as an integral component of scientific methodology, exploring data patterns and hypotheses interactively (Peters et al. 2014). As datasets become increasingly large, machine learning will play an important role in extracting relevant patterns and answers from big data (Durden et al. 2017). As web scraping and natural language processing tools develop and become more accessible, information aggregation and synthesis will become increasingly automated, changing the roles that scientists play in the scientific process. It would therefore behoove ecologists to have some level of familiarity with machine learning and artificial intelligence as they become integral parts of the scientific process.

Summary

We applied multiple statistical learning techniques to a range of information sources to identify prespawning behavior, a potential indicator of spawning sites for permit in the Florida Keys. These findings may serve to guide permit conservation efforts by management and conservation organizations through educational or legal avenues, in particular to address issues related to angling depredation at spawning

aggregation sites. The predictions from the developed models may also serve to inform future research approaches, for example, guiding visual surveys to confirm spawning sites or structuring local ecological knowledge surveys. We also provide insights for identifying ecological phenomena, including cryptic spawning aggregations, using various information sources. The application of statistical learning techniques allowed us to overcome significant analytical challenges, and we provide relevant information and tools, including R code, to increase their accessibility.

Acknowledgements This project was funded by Bonefish and Tarpon Trust with support from Costa Del Mar, The March Merkin Fishing Tournament, Hell's Bay Boatworks, and private donations. Additional support was provided by a NASEM Gulf Research Program through a hurricane recovery grant, and the acoustic receiver array was partially supported by a loan from the Ocean Tracking Network. We thank the fishing guides and anglers who assisted with the telemetry array design and Permit tagging for this project including Captains Travis and Bear Holeman, Will Benson, Rob Kramarz, Zack Stells, Brandon, and Jared Cyr, Chris Trosset, Nathaniel Linville, Ian Slater, Augustine Moss, Sandy Horn, Ted Margo, Richard Berlin, and Jeff Rella. We also thank the researchers that shared permit acoustic telemetry detection data with us through integrated Tracking of Animals in the Gulf (iTAG) and Florida Acoustic Telemetry Network (FACT), in particular Harold "Wes" Pratt of Mote Marine Laboratory, funded by The Shark Foundation/Hai Stiftung, and Mike McAllister. Brownscombe is supported by a Banting Postdoctoral Fellowship, Dalhousie University, Carleton University, and Bonefish and Tarpon Trust. This research was conducted with permission of the Florida Keys National Marine Sanctuary under permit # FKNMS-2013-040-A2, and the Florida Fish and Wildlife Conservation Commission under permit # SAL-16-1205.

Author contribution statement JWB designed and conducted data collection, analysis, and manuscript writing. LPG, DM, AA, JH, SKL-B, AJA, AJD, SJC contributed to data collection and manuscript preparation.

References

Adams AJ, Blewett DA (2004) Spatial patterns of estuarine habitat type use and temporal patterns in abundance of juvenile permit, trachinotus falcatus, in Charlotte Harbor, Florida. Gulf Caribb Res 16:129–139. https://doi.org/10.18785/gcr.1602.01

Adams AJ, Cooke SJ (2015) Advancing the science and management of flats fisheries for bonefish, tarpon, and permit. Environ Biol Fishes 98:2123–2131. https://doi.org/10.1007/s10641-015-0446-9

Adams AJ, Wolfe RK, Kellison GT, Victor BC (2006) Patterns of juvenile habitat use and seasonality of settlement by permit, *Trachinotus falcatus*. Environ Biol Fishes 75:209–217. https://doi.org/10.1007/s10641-006-0013-5

Adams AJ, Shenker JM, Jud ZR et al (2019) Identifying pre-spawning aggregation sites for bonefish (*Albula vulpes*) in the Bahamas to inform habitat protection and species conservation. Environ Biol Fishes 102:159–173. https://doi.org/10.1007/s10641-018-0802-7

Aguilar-Perera A (2006) Disappearance of a Nassau grouper spawning aggregation off the southern Mexican Caribbean coast. Mar Ecol Prog Ser 327:289–296. https://doi.org/10.3354/meps327289

Archer SK, Allgeier JE, Semmens BX et al (2015) Hot moments in spawning aggregations: implications for ecosystem-scale nutrient cycling. Coral Reefs 34:19–23. https://doi.org/10.1007/s00338-014-1208-4

Ascough JC, Maier HR, Ravalico JK, Strudley MW (2008) Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecol Modell 219:383–399. https://doi.org/10.1016/j.ecolmodel.2008.07.015

Binder TR, Farha SA, Thompson HT et al (2018) Fine-scale acoustic telemetry reveals unexpected lake trout, Salvelinus namaycush, spawning habitats in northern Lake Huron, North America. Ecol Freshw Fish 27:594–605. https://doi.org/10.1111/eff.12373

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification algorithms and regression trees. Classification and regression tree. Wadsworth International Group, Belmont, California, pp 246–280

Brook RK, McLachlan SM (2008) Trends and prospects for local knowledge in ecological and conservation research and monitoring. Biodivers Conserv 17:3501–3512. https://doi.org/10.1007/s10531-008-9445-x

Brown DD, Kays R, Wikelski M et al (2013) Observing the unwatchable through acceleration logging of animal behavior. Anim Biotelemetry 1:20. https://doi.org/10.1186/2050-3385-1-20

Brownscombe JW, Adams AJ, Young N et al (2019a) Bridging the knowledge-action gap: a case of research rapidly impacting recreational fisheries policy. Mar Policy 104:210–215

Brownscombe JW, Danylchuk AJ, Adams AJ et al (2019b) Bonefish in South Florida: status, threats and research needs. Fish Res 102:329–348

Brownscombe JW, Griffin LP, Morley D et al (2020) Seasonal occupancy and connectivity amongst nearshore flats and reef habitats by permit (*Trachinotus falcatus*): considerations for fisheries management. J Fish Biol 96:469–479

Bryan DR, Luo J, Ault JS et al (2015) Transport and connectivity modeling of larval permit from an observed spawning aggregation in the Dry Tortugas, Florida. Environ Biol Fishes 98:2263–2276. https://doi.org/10.1007/s10641-015-0445-x

Cagnacci F, Boitani L, Powell RA, Boyce MS (2010) Animal ecology meets GPS-based radiotelemetry: a perfect storm of opportunities and challenges. Philos Trans R Soc B Biol Sci 375:2157–2162. https://doi.org/10.1098/rstb.2010.0107

Chon TS, Park YS, Park JH (2000) Determining temporal pattern of community dynamics by using unsupervised learning algorithms. Ecol Modell 132:151–166. https://doi.org/10.1016/S0304-3800(00)00312-4

Christin S, Hervet É, Lecomte N (2019) Applications for deep learning in ecology. Methods Ecol Evol 10:1632–1644. https://doi.org/10.1111/2041-210X.13256

Cooke SJ, Suski CD, Arlinghaus R, Danylchuk AJ (2013) Voluntary institutions and behaviours as alternatives to formal regulations in recreational fisheries management. Fish Fish 14:439–457. https://doi.org/10.1111/j.1467-2979.2012.00477.x

Crabtree RE, Hood PB, Snodgrass D (2002) Age, growth, and reproduction of permit (*Trachinotus falcatus*) in Florida waters. Fish Bull 100:26–34

Crossin GT, Heupel MR, Holbrook CM et al (2017) Acoustic telemetry and fisheries management. Ecol Appl 27:1031–1049. https://doi.org/10.1002/eap.1533

Cutler DR, Edwards TC, Beard KH et al (2007) Random forests for classification in ecology. Ecology 88:2783–2792. https://doi.org/10.1890/07-0539.1

Danylchuk AJ, Cooke SJ, Goldberg TL et al (2011) Aggregations and offshore movements as indicators of spawning activity of bonefish (*Albula vulpes*) in The Bahamas. Mar Biol 158:1981–1999. https://doi.org/10.1007/s00227-011-1707-6

De'Ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data

analysis. Ecology 81:3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2
De Mitcheson YS, Colin PL (2012) Reef fish spawning aggregations: biology, research and management. Springer, New York
Durden JM, Luo JY, Alexander H et al (2017) Integrating "big data" into aquatic ecology: challenges and opportunities. Limnol Oceanogr Bull 26:101–108. https://doi.org/10.1002/lob.10213
Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x
Equihua M (1990) Fuzzy clustering of ecological data. J Ecol. https://doi.org/10.2307/2261127
Erisman BE, Allen LG, Claisse JT et al (2011) The illusion of plenty: hyperstability masks collapses in two recreational fisheries that target fish spawning aggregations. Can J Fish Aquat Sci 68:1705–1716. https://doi.org/10.1139/f2011-090
Erisman B, Heyman W, Kobara S et al (2017) Fish spawning aggregations: where well-placed management actions can yield big benefits for fisheries and conservation. Fish Fish 18:128–144. https://doi.org/10.1111/faf.12132
Ferraro MB, Giordani P (2015) A toolbox for fuzzy clustering using the R programming language. Fuzzy Sets Syst 279:1–16. https://doi.org/10.1016/j.fss.2015.05.001
Fulton EA, Blanchard JL, Melbourne-Thomas J et al (2019) Where the ecological gaps remain, a modelers' perspective. Front Ecol Evol 7:424. https://doi.org/10.3389/fevo.2019.00424
Graham RT, Castellanos DW (2005) Courtship and spawning behaviors of carangid species in Belize. Fish Bull 103:426–432
Graham RT, Carcamo R, Rhodes KL et al (2008) Historical and contemporary evidence of a mutton snapper (Lutjanus analis Cuvier, 1828) spawning aggregation fishery in decline. Coral Reefs 27:311–319. https://doi.org/10.1007/s00338-007-0329-4
Green J, Willis K, Hughes E et al (2007) Generating best evidence from qualitative research: the role of data analysis. Aust N Z J Public Health 31:545–550. https://doi.org/10.1111/j.1753-6405.2007.00141.x
Guilford T, Meade J, Willis J et al (2009) Migration and stopover in a small pelagic seabird, the Manx shearwater *Puffinus puffinus*: insights from machine learning. Proc R Soc B Biol Sci 276:1215–1223. https://doi.org/10.1098/rspb.2008.1577
Halsey LG (2019) The reign of the *p* -value is over: what alternative analyses could we employ to fill the power vacuum? Biol Lett 15:20190174. https://doi.org/10.1098/rsbl.2019.0174
Hamilton R, De Mitcheson YS, Aguilar-Perera A (2012) The role of local ecological knowledge in the conservation and management of reef fish spawning aggregations. In: Sadovy de Mitcheson Y, Colin P (eds) Reef fish spawning aggregations: biology, research and management. Springer, Dordrecht, pp 331–369
Hanski I, Gaggiotti O (2004) Ecology, genetics and evolution of metapopulations. Academic Press, Cambridge
Holder PE, Griffin LP, Adams AJ et al (2020) Stress, predators, and survival: exploring permit (*Trachinotus falcatus*) catch-and-release fishing mortality in the Florida Keys. J Exp Mar Bio Ecol 524:151289
Hoppner F, Klawonn F, Kruse R, Runkler T (2000) Fuzzy cluster analysis: methods for classification, data analysis and image recognition. Wiley, West Sussex
Hothorn T, Zeileis A (2015) Partykit: A modular toolkit for recursive partytioning in R. J Mach Learn Res 16(1):3905–3909
Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15:651–674. https://doi.org/10.1198/106186006X133933
Hussey NE, Kessel ST, Aarestrup K et al (2015) Aquatic animal telemetry: a panoramic window into the underwater world. Science 348:1255642. https://doi.org/10.1126/science.1255642
Johnson DH (1981) The use and misuse of statistics in wildlife habitat studies. Use Multivar Stat Stud Wildl Habitat Gen Tech Rep Stn. https://doi.org/10.5962/bhl.title.99662
Kobara S, Heyman WD (2008) Geomorphometric patterns of Nassau grouper (*Epinephelus striatus*) spawning aggregation sites in the Cayman Islands. Mar Geod 31:231–245. https://doi.org/10.1080/01490410802466397
Kobara S, Heyman WD (2010) Sea bottom geomorphology of multispecies spawning aggregation sites in Belize. Mar Ecol Prog Ser 405:243–254. https://doi.org/10.3354/meps08512
Kuhn M, Wing J, Weston S et al (2019) caret: Classification and regression training. R package version 6.0-84. https://CRAN.Rproject.org/package=caret
Leis JM, Carson-Ewart BM, Hay AC, Cato DH (2003) Coral-reef sounds enable nocturnal navigation by some reef-fish larvae in some places and at some times. J Fish Biol 63:724–737. https://doi.org/10.1046/j.1095-8649.2003.00182.x
Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22. https://doi.org/10.1177/154405910408300516
Lowerre-Barbieri SK, Walters S, Bickford J et al (2013) Site fidelity and reproductive timing at a spotted seatrout spawning aggregation site: individual versus population scale behavior. Mar Ecol Prog Ser 481:181–197. https://doi.org/10.3354/meps10224
Lowerre-Barbieri SK, Walters Burnsed SL, Bickford JW (2016) Assessing reproductive behavior important to fisheries management: a case study with red drum, *Sciaenops ocellatus*. Ecol Appl 26:979–995. https://doi.org/10.1890/15-0497
Lowerre-Barbieri S, DeCelles G, Pepin P et al (2017) Reproductive resilience: a paradigm shift in understanding spawner–recruit systems in exploited marine fish. Fish Fish 18:285–312. https://doi.org/10.1111/faf.12180
Ludwig D, Hilborn R, Walters C (1993) Uncertainty, resource exploitation, and conservation: lessons from history. Science. https://doi.org/10.1126/science.260.5104.17
Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18:189–197. https://doi.org/10.1016/S0169-5347(03)00008-9
Menzies CR (2006) Traditional ecological knowledge and natural resource management. U of Nebraska Press, Lincoln
Michener R, Lajtha K (2008) Stable isotopes in ecology and environmental science, 2nd edn. Wiley, Hoboken
Michener WK, Jones MB (2012) Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol Evol 27:85–93. https://doi.org/10.1016/j.tree.2011.11.016
Molnar C (2019) Interpretable machine learning. a guide for making black box models explainable. Lean Publishing
Mourier J, Maynard J, Parravicini V et al (2016) Extreme inverted trophic pyramid of reef sharks supported by spawning groupers. Curr Biol 26:2011–2016. https://doi.org/10.1016/j.cub.2016.05.058
Nguyen VM, Young N, Cooke SJ (2017) A roadmap for knowledge exchange and mobilization research in conservation and natural resource management. Conserv Biol 31:789–798. https://doi.org/10.1111/cobi.12857
Nguyen VM, Young N, Cooke S (2018) Applying a knowledge-action framework for navigating barriers to incorporating telemetry science into fisheries management and conservation: a qualitative study. Can J Fish Aquat Sci 75:1733–1743
Nowlin WH, Vanni MJ, Yang LH (2008) Comparing resource pulses in aquatic and terrestrial ecosystems. Ecology 89(3):647–659
Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83:171–193. https://doi.org/10.1086/587826

Oppel S, Strobl C, Huettmann F (2009) Alternative methods to quantify variable importance in ecology. Ludwig-Maximilians-Universität, München, pp 1–7

Paris CB, Atema J, Irisson JO et al (2013) Reef odor: a wake up call for navigation in reef fish larvae. PLoS ONE 8:e72808. https://doi.org/10.1371/journal.pone.0072808

Peters DPC, Havstad KM, Cushing J et al (2014) Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. Ecosphere 5:1–15. https://doi.org/10.1890/ES13-00359.1

Piironen J, Paasiniemi M, Vehtari A (2020) Projective inference in high-dimensional problems: prediction and feature selection. Electron J Stat. https://doi.org/10.1214/20-ejs1711

Pitcher TJ (2001) Fish schooling. In: Steele JH, Thorpe SA, Turekian KK (eds) Encyclopedia of ocean sciences: marine biology, pp 337–349

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.Rproject.org/

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. https://doi.org/10.1145/2939672.2939778

RStudio Team (2016) RStudio: Integrated Development for R. RStudio Inc., Boston, MA. http://www.rstudio.com/

Sadovy Y, Domeier M (2005) Are aggregation-fisheries sustainable? Reef fish fisheries as a case study. Coral Reefs 24:254–262. https://doi.org/10.1007/s00338-005-0474-6

Sala E, Ballesteros E, Starr RM (2001) Rapid decline of Nassau grouper spawning aggregations in belize: fishery management and conservation needs. Fisheries 26:23–30. https://doi.org/10.1577/1548-8446(2001)026<0023:rdongs>2.0.co;2

Salski A (2007) Fuzzy clustering of fuzzy ecological data. Ecol Inform 2:262–269. https://doi.org/10.1016/j.ecoinf.2007.07.002

Sancho G, Petersen CW, Lobel PS (2000) Predator–prey relations at a spawning aggregation site of coral reef fishes. Mar Ecol Prog Ser. https://doi.org/10.3354/meps203275

Santos RO, Rehage JS, Kroloff EKN et al (2018) Combining data sources to elucidate spatial patterns in recreational catch and effort: fisheries-dependent data and local ecological knowledge applied to the South Florida bonefish fishery. Environ Biol Fishes 201:299–317

Silvano RAM, MacCord PFL, Lima RV, Begossi A (2006) When does this fish spawn? Fishermen's local knowledge of migration and reproduction of Brazilian coastal fishes. Environ Biol Fishes 76:371–386. https://doi.org/10.1007/s10641-006-9043-2

Simpfendorfer CA, Huveneers C, Steckenreuter A et al (2015) Ghosts in the data: false detections in VEMCO pulse position modulation acoustic telemetry monitoring equipment. Anim Biotelemetry 3:55. https://doi.org/10.1186/s40317-015-0094-z

Soria M, Dagorn L, Potin G, Fréon P (2009) First field-based experiment supporting the meeting point hypothesis for schooling in pelagic fish. Anim Behav 78:1441–1446. https://doi.org/10.1016/j.anbehav.2009.09.025

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform 8:25. https://doi.org/10.1186/1471-2105-8-25

Strobl C, Boulesteix A-L, Kneib T et al (2008) Conditional variable importance for random forests. BMC Bioinform 9:307. https://doi.org/10.1186/1471-2105-9-307

Strobl C, Hothorn T, Zeileis A (2009) Party on! A new, conditional variable importance measure available in the party package. Technical Report Number 050, Department of Statistics, University of Munich

Waterhouse L, Heppell SA, Pattengill-Semmens CV et al (2020) Recovery of critically endangered Nassau grouper (*Epinephelus striatus*) in the Cayman Islands following targeted conservation actions. Proc Natl Acad Sci USA 117:1587–1595. https://doi.org/10.1073/pnas.1917132117

West JB, Bowen GJ, Cerling TE, Ehleringer JR (2006) Stable isotopes as one of nature's ecological recorders. Trends Ecol Evol 21:408–414. https://doi.org/10.1016/j.tree.2006.04.002

Young N, Gingras I, Nguyen VM et al (2013) Mobilizing new science into management practice: the challenge of biotelemetry for fisheries management, a case study of Canada's Fraser River. J Int Wildl Law Policy 16:331–351. https://doi.org/10.1080/13880292.2013.805074

Zeller DC (1998) Spawning aggregations: Patterns of movement of the coral trout *Plectropomus leopardus* (Serranidae) as determined by ultrasonic telemetry. Mar Ecol Prog Ser 162:253–263. https://doi.org/10.3354/meps162253

Zeng X, Adams A, Roffer M, He R (2018) Potential connectivity among spatially distinct management zones for Bonefish (*Albula vulpes*) via larval dispersal. Environ Biol Fishes 102:233–252. https://doi.org/10.1007/s10641-018-0826-z

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 1:3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x