Received: 23 August 2022 | Revised: 23 January 2023 | Accepted: 25 January 2023

DOI: 10.1111/fme.12617

Fisheries Management and Ecology WILEY

ARTICLE

Generalized additive models for categorical count data: An exploration of the decline of queen triggerfish *Balistes vetula* in the Bahamas and Turks and Caicos

John P. Urquhart | **Donald B. Olson** | **Elizabeth A. Babcock**

Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida, USA

Correspondence
John P. Urquhart, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami 33149, FL, USA.
Email: jackpurquhart@gmail.com

Present address
John P. Urquhart, School for Environment and Sustainability, University of Michigan, Ann Arbor, Michigan, USA

Abstract

Citizen science is growing in importance for ecosystem management and long-term monitoring. A large marine citizen-science project operated by the Reef Environmental Education Foundation (2021, Reef environmental education foundation volunteer fish survey project database, World Wide Web electronic publication) collected logarithmic categorical data for species abundance across a number of otherwise understudied reefs in The Bahamas and Turks and Caicos during 1994–2020. We used several statistical models to estimate the presence and abundance of trends from these data. Variously specified abundance and presence-absence models were fit to simulated count data, simulated categorized count data, and real-world categorical data for Queen Triggerfish (*Balistes vetula*). These models produced simple patterns of presence and abundance from simulated data with minimal bias that were reasonable predictions based on cross-validation. Based on model-based estimates of presence and abundance, the Queen Triggerfish population decreased significantly in The Bahamas and Turks and Caicos during 1994–2020. This simple method for imputing abundance from size-category counts at the level of individual diver observations, rather than aggregated across multiple observations, allows for higher resolution modeling of predictors of presence and abundance, with implications for other understudied reef-dwelling species.

KEYWORDS
abundance models, citizen science, categorical abundance, fisheries independent data, REEF, queen triggerfish

1 | INTRODUCTION

As marine conservation grows in importance, data are needed to monitor the efficacy of interventions and to observe disruptions to ecosystems to mitigate damage. However, data is expensive to collect and is likely focused on areas of perceived importance (Campbell et al., 2022). Citizen science presents an effective, low-cost way to measure species abundance at both regional and global scales.

Since the 1990s, the Reef Environmental Education Foundation (REEF, 2021) has provided citizen scientists with an opportunity to collect fish-count data while snorkeling and diving recreationally (REEF, 2021). Data collection has resulted in a dataset comprised of over 250,000 roving diver survey observations ("surveys") at over 15,000 locations globally (Rassweiler et al., 2020; REEF, 2021). The roving diver protocol used by REEF allows divers to freely move and record each species they observe, without restrictions on time

underwater or distance traveled. This data has proven valuable for determining the efficacy of marine protected areas, and for monitoring population trends (Gravem et al., 2020; Pattengill-Semmens & Semmens, 2003). However, several features of the dataset make fine-scale analysis difficult. Data collection is opportunistic, so the dataset contains many variables that are strongly correlated to the particularities of the sampling, such as depth of survey and bottom time, which makes hypothesis testing difficult. Data is also collected by many surveyors of differing levels of experience in identifying different species, which adds an additional layer of error to the dataset. These kinds of problems have been addressed in analyses of other large opportunistic datasets (e.g., eBird, Johnston et al., 2021). Current best practices for citizen science rely on statistical models (Johnston et al., 2021). Previous modeling has incorporated REEF presence-absence information into larger model-training sets, or combined information from multiple surveys and external count information to create abundance indices (Campbell et al., 2022; Grüss et al., 2018; Montecino-Latorre et al., 2016; Tolimieri et al., 2017).

Due to the abundance and diversity of reef fish, the abundance of each encountered species is difficult to estimate while diving. This led to the creation of a categorical count system that allows divers to mark whether they observed a single individual, a few individuals (2–10), many individuals (10–100), or abundant individuals (100+). This data also contains zeros, which imply that a species was not present or that the diver was unable to identify or did not see a species. While Single-Few-Many-Abundant (SFMA) categories are representative of actual abundance, they are condensed into a single categorical value. REEF divers record information in the form of a species checklist, allowing them to "check off" each species positively identified and mark its abundance category on a dive. For each checklist (i.e., survey), the date, time, and environmental variables are recorded, and REEF staff members internally mark the surveyor's experience level on each survey when the data are entered into the database. REEF also records a measure of surveyor experience. Surveyors move from novice to experienced by completing a specified number of surveys within a region and taking an identification quiz. Existing methods infer a mean abundance from REEF categories aggregated across multiple surveys, but do not account for information from confounding variables such as dive duration (Campbell et al., 2022; Tolimieri et al., 2017; Wolfe & Pattengill-Semmens, 2013a). At the survey level, REEF data can be interpreted as binary (presence-absence) or multinomial (by categories), although models based on these distributions omit much of the information on the abundance that is captured by SFMA categories.

One way to reduce errors produced by a lack of observer skill is to examine species that are easily identifiable. The Queen Triggerfish (*Balistes vetula*) is an easily identifiable species with purple markings along its body, ornate fins, and a distinct body shape (Figure 1). The Queen Triggerfish is also large enough to be easily observed by a roving diver, which alleviates some concern about surveyors being unable to identify a species and makes it a good candidate for a model organism in citizen science. While the Queen Triggerfish is of minor economic importance, it is a food fish locally, and few fisheries-based or fishery-independent data sources include this species. Artisanal fisheries have led to local reductions in population elsewhere, and the species is listed as near threatened by the IUCN red list because of depletion in part of its range (Liu et al., 2015; Sagarese et al., 2018). No data from The Bahamas and Turks and Caicos were used in evaluating Queen Triggerfish for listing (Figure 2).

FIGURE 1 An adult Queen Triggerfish (*Balistes vetula*) observed by the author on a reef slope in Belize on 2/19/2014.

Our objective was to determine if trends in the presence and abundance of Queen Triggerfish in The Bahamas, Turks, and Caicos could be estimated by accounting for variability in sampling details of REEF surveys (duration, experience, time of sampling). The method includes two steps, an imputation of the number of fish observed in each survey based on counts in each size category. and then a Generalized Additive Model (GAM) to predict mean abundance in each survey. Many predictor variables affect abundance nonlinearly, so linear regression is likely inappropriate. Thus, Generalized Additive Models were used because they allow for a non-linear relationship between predictor variables and the response variable, estimated through penalized smoothing functions. This family of models has already been applied to binomial data, but has not been applied to abundance information contained in REEF (Grüss et al., 2018). Like linear models, GAM predicts mean abundance for each observation as a function of predictor variables. This two-part estimation method was tested in simulation, and then applied to the Queen Triggerfish, to confirm that it worked in practice with covariates collected by REEF. Without the ability to account for survey covariates, changes in observed abundance could also be caused by changes in underlying survey conditions. These methods allow results to be more robust and are closer to best practices for other semi-structured citizen science initiatives (Johnston et al., 2021). Although the use of abundance information may not be necessary for all species because patterns in abundance and presence-absence are likely to be similar, abundance is more information-rich, and for some species, can be more informative (Joseph et al., 2006).

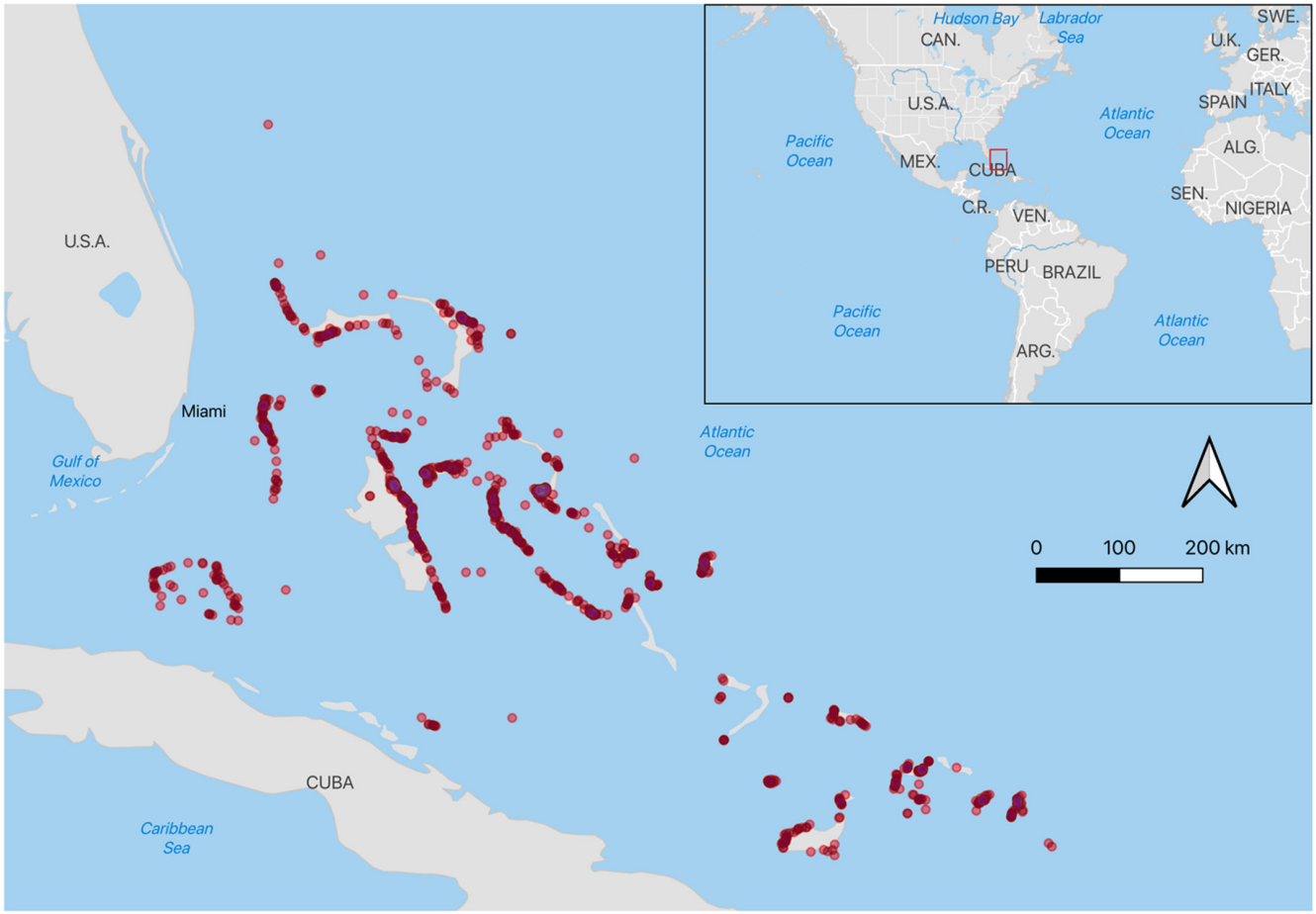

FIGURE 2 Locations of sites surveyed for reef fishes by divers in The Bahamas and Turks and Caicos during 1994–2020 (Reef Environmental Education Foundation, REEF).

2 | METHODS

2.1 | Imputation of categorical data

The number of observations in adjacent abundance categories was used to estimate the arithmetic mean of each category as if the sampled distribution was log-normal (Wolfe & Pattengill-Semmens, 2013a, 2013b). These averages were then used to estimate average abundance over a particular group of surveys. In REEF data, four abundance categories are non-zero: S = number of observations of lone organisms, F = number of observations of 2–10 organisms, M = number of observations of 11–100 organisms, and A = number of observations of >100 organisms. The mean abundance in each category was estimated as follows:

$$\text{AverageF} = (2*S + 4.16*F + 10*M)/(S+F+M) \quad (1a)$$

$$\text{AverageM} = (11*F + 33.8*F + 100*A)/(F+M+A) \quad (1b)$$

$$\text{AverageA} = (200*M + 348*A)/(M+A) \quad (1c)$$

$$\text{MeanAbundance} = (S + F*\text{AverageF} + M*\text{AverageM} + A*\text{AverageA})/(S+F+M+A) \quad (1d)$$

Where AverageF, AverageM, and AverageA are estimated average counts in the F, M, and A categories, and MeanAbundance is the average count per observation. Because S is the number of solitary fish observed (i.e., average = 1), S is used directly in equation 1d.

This method was derived by calibrating the average value for each category to actual count data that was taken concurrently with SFMA counts in California and creating a simulated confidence interval for the mean (Wolfe & Pattengill-Semmens, 2013a, 2013b). For our analysis, instead of aggregating scores to estimate mean abundance across multiple observations, values of each survey were imputed as the mean of its reported abundance category (i.e., AverageF, AverageM, or AverageA). These averages were calculated over the entire dataset so that each observation of the same SFMA category was given the same value. This method does not incorporate errors introduced by the process of imputing abundance from REEF categories. Substituting the mean for each category is a very simple method, but produces a y variable that is appropriately scaled to convey information about the relative abundance of observations.

Only 10 observations were of the abundant category of Queen Triggerfish in REEF data and were omitted due to missing information in important covariates. Therefore, models were trained with only sub-100 counts, and abundant counts were not considered. Averages were calculated over the entire dataset so that each observation of the same SFM category had the same value: AverageF = 3.234952 and AverageM = 12.518702.

2.2 | Simulated data

To examine potential bias in applying a negative binomial model to imputed count data from categories, a simulated dataset was created for model comparison in R (R Developement Core Team, 2020). Parameters were selected to mimic data for Queen Triggerfish used in this study (Figure 3). The Few and Many categories were slightly overrepresented in simulated data compared to actual data for Queen Triggerfish, and the abundant category was missing as in the actual data. Data were simulated by randomly generating a set of variables (x), with sample sizes ranging from 100 to 10,000, where x was taken from a uniform distribution from −3 to 3, and then generating count data (y) from a negative binomial distribution with mean $\mu = e^x$, and size = 10. Samples of 100, incremented in size by 100 (ranging from 100 to 10,000), were used to determine how model performance changed with the amount of training data. Each simulated data set resulted in abundance data that could be described by a negative binomial model with one independent x-variable of slope = 1 and intercept = 0 when modeled with a log link Generalized Linear Model (GLM). The negative binomial distribution is often applied to over-dispersed count data of organisms, such as citizen science data (Johnston et al., 2021). Simulated count data was then coerced into 0-SFM categories by replacing counts between 2 and 10 with F, and counts between 11 and 100 with M.

Negative binomial GLM was then applied with four versions of abundance data. In the first GLM, the y variable was the original uncategorized data, for comparison to models fitted to abundance imputed from SFM categories. In the second GLM, y was abundance data after being coerced into SFM categories and then converted back to abundance with imputed category means (Equation 1). In the third model, abundance (y) was imputed as minimum possible values for each observation category (i.e., $F = 2$, $M = 11$, $A = 101$). In the fourth GLM, an alternative exponential imputation method was used (Wolfe & Pattengill-Semmens, 2013a, 2013b), in which $y = 5.73$^(DEN-1)^1.28, where DEN was an index ($S = 1$, $F = 2$, $M = 3$, and $A = 4$). The mean method described in Equation 1 was considered to be the best (Wolfe & Pattengill-Semmens, 2013a, 2013b), whereas the other two methods were used to evaluate how much the imputation method influenced the imputed trend. All models were fit using the MGCV package in R, with the x variable assumed to have a linear relationship with the log link mean (Wood, 2017).

Categorized and uncategorized model predictions were compared to true uncategorized count data to measure how well-categorized data tracked true underlying abundance trends. True

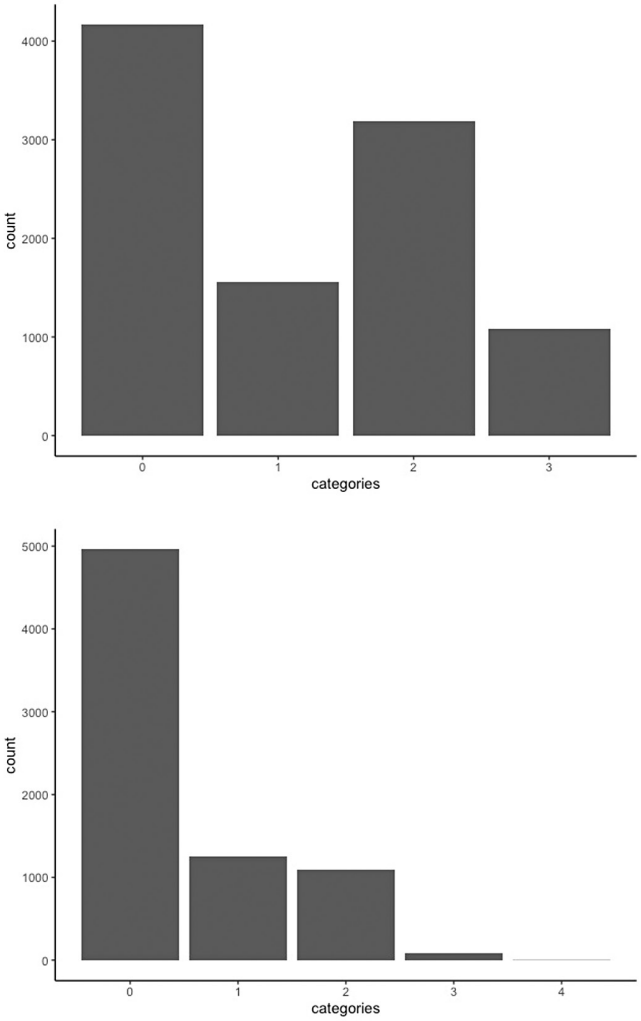

FIGURE 3 Counts of 10,000 simulated observations (Top) and actual data (Bottom) of zero (Category 0), one (Category 1), few (2–10), and many (11–100) Queen Triggerfish (*Balistes vetula*) observed by citizen-science divers in The Bahamas and Turks and Caicos during 1994–2020 (Reef Environmental Education Foundation, REEF).

and predicted data were compared with root mean square error (RMSE), R-squared (R^2), and average error:

$$R^2 = 1 - \sum_{i=1}^{N} \left(y_i - y_i'\right)^2 / \sum_{i=1}^{N} \left(y_i - \overset{!!}{y}\right)^2 \quad (2a)$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} \left(y_i - y_i'\right)^2 / N} \quad (2b)$$

$$\mathrm{AverageError} = \sum_{i=i}^{N} \left(y_i' - y_i\right) / N \quad (2c)$$

Where N is the number of values, y_i is the i^{th} recorded value, y_i' is the i^{th} predicted value, and $\overset{!!}{y}$ is the mean value of the variable.

Metrics were calculated with both y_i and y_i' as either uncategorized abundance counts or means imputed from categories to evaluate how much error was introduced by using an imputed mean rather than counts. Categorized model predictions were also compared to categorized counts using the fraction of predictions that were in the correct abundance category. Code Appendix S1 contains code used to simulate categorized values, fit models, and calculate metrics. After metrics were calculated for each simulated sample and model, the mean and standard deviation among samples were calculated for each metric.

2.3 | Fitting to REEF data

REEF data was used from 1994 through 2020 in The Bahamas and Turks and Caicos (Figure 2) (REEF). Data manipulation was in R using RStudio with the tidyverse, measurements, lubridate, and lunar packages (Birk, 2019; Grolemund & Wickham, 2011; Lazaridis, 2020; R Core Team, 2020; RStudio, 2022; Wickham et al., 2019). Most surveys were collected by novice REEF surveyors, so data were not excluded based on experience. This was justified by the use of an easily identified model species, and the inclusion of experience as a binary parameter in the model. Surveys for which depth, visibility, habitat, current, and start time were recorded as zero were excluded because zero reflected missing data for most variables. For example, the lowest depth category recorded by REEF (a snorkel) is recorded as a 1, not a 0. For Start Time, although dives can begin at 00:00, an anomalous number of dives began at that time compared to other times late in the evening, so we assumed they were data entry errors. Surface and bottom temperatures below 10°C were well below Bahamian climatology, so were excluded. Bottom times <10 min and longer than 100 min were excluded. This was done based on work done with eBird data which standardized survey efforts in a similar fashion (Johnston et al., 2021). This filter removed <3% of the available surveys and so effectively removed durations for which high precision estimation was not feasible. Average depth, current, and visibility were collected by REEF as ordered categories (Appendix S1). Dives deeper than 24 m were combined into one category due to the small number at that depth. They were not excluded because the ordinal parametric fit used to model categories was capable of fitting to a category for dives of a certain depth or greater. Habitats with <100 observations were excluded since the precise estimation of occurrence within those habitats was not possible. After data cleanup, all observations of abundant (>100) Queen triggerfish contained several data entry issues, were outliers (10 of 18,345 total observations), and were therefore excluded from the analysis. The original dataset of 18,345 observations was reduced to 7380 after observations with missing values were omitted (Table 1; Appendix S1).

Lunar phase in radians and decimal day of the year were calculated from the date recorded for each survey. Queen Triggerfish, like several other large reef fish, aggregate seasonally for breeding during the full moon (Bryan et al., 2019), the only known aggregation by this species. To accommodate this behavior, a Boolean variable (0,1) was used to represent whether or not a survey fell on a full moon during peak breeding activity from November to March. The full moon was defined as the quarter of the lunar cycle centered around the full moon (approximately 4 days before and after).

TABLE 1 Percentage of missing values for variables measured in association with observations of Queen Triggerfish (*Balistes vetula*) by citizen-science divers in The Bahamas and Turks and Caicos during 1994–2020 (Reef Environmental Education Foundation, REEF).

Variable (units)	% Missing
Date of observation (decimal year)	0.06%
Surface temperature during observation (°F)	42.2%
Bottom temperature during observation (°F)	31.0%
Duration of observation period (minutes)	2.70%
Start time of observation (decimal hour)	0.45%
Visibility during observation (categorical)	0.92%
Average Depth of observation (categorical)	0.52%
Current experienced during observation (categorical)	1.24%
Habitat surveyed (categorical)	2.69%
Latitude of observation site (decimal degrees)	10.2%
Longitude of observation site (decimal degrees)	10.2%

Variables were selected for inclusion in models using Akaike Information Criterion (AIC). AIC selection was used for selecting terms to reduce overfitting because models were fit to the same dataset (Hastie et al., 2009). The bidirectional selection was used, starting with backward selection, and then testing if adding any variables back to the final model improved the AIC score. The model with the lowest AIC was selected. After AIC selection, the selected model was fit to the entire valid dataset for selected variables. Variables included: Visibility, Depth, Bottom time, surface temperature, latitude, longitude, current, breeding vs non-breeding, moon phase, decimal day of the year, decimal year, surveyor experience (Expert or Novice), and start time. Cyclical variables were fit using a cyclic cubic spline, and other numeric variables were fit using thin-plate regressor splines. Latitude and longitude were incorporated into the model using a tensor product smooth, and the breeding variable was treated as an interaction variable on the latitude and longitude smooth because different locations were expected to exhibit different trends in abundance during spawning aggregations. REML was used as the parameter estimation method. The default basis dimension parameter K was used for smoothing functions, because exploratory fits with larger K did not change the statistical significance of the model terms, but increased run time from a few minutes to several hours. The gam.check function was used to inform k selection, and while several *k* values were highly significant, increasing *k* values past 100 did not improve values for gam.check. The overall form of models was as follows: $y \sim \text{experience} + s(\text{moon phase, bs} = \text{“cc”}) + \text{current} + \text{averaged depth} + \text{habitat} + s(\text{bottom}$

TABLE 2 Goodness of fit (R^2), root-mean-square error (RMSE), average error, final slope, and final intercept of negative binomial model metrics for simulated data (100–10,000 data points), from comparing model predictions to imputed categorical and original uncategorized data. Simulation parameters were set to imitate observations of Queen Triggerfish (*Balistes vetula*) in The Bahamas and Turks and Caicos during 1994–2020, based on marine citizen-science data (Reef Environmental Education Foundation, REEF). The final slope and intercept (with standard error) for final models were fit to the dataset of 10,000 points (expected to be 0 and 1, respectively).

Model input	R^2 relative to imputed data (SD)	RMSE relative to imputed data (SD)	Avg. error relative to imputed data (SD)	R^2 relative to uncat. Data (SD)	RMSE relative to uncat. Data (SD)	Avg. error relative to uncat. Data (SD)	Final slope (SE)	Final intercept (SE)
Imputed Mean	0.701 (0.014)	7.173 (0.577)	0.039 (0.012)	0.758 (0.014)	7.017 (0.501)	0.029 (0.014)	0.863 (0.007)	0.294 (0.012)
Imputed Minimum	0.646 (0.017)	3.588 (0.196)	−0.021 (0.003)	0.512 (0.015)	14.131 (0.858)	−1.430 (0.053)	0.868 (0.008)	−0.289 (0.016)
Imputed Exponential	0.590 (0.020)	172.306 (9.182)	−0.679 (0.095)	−2.816 (0.460)	110.765 (15.521)	5.081 (0.460)	1.203 (0.009)	0.520 (0.016)
Uncategorized abundance	NA	NA	NA	0.769 (0.012)	6.686 (0.481)	0.000 (0.011)	1.008 (0.007)	−0.004 (0.014)

temperature, bs = "tp") + *s*(decimal day of year, bs = "cc") + *s*(date, bs = "tp") + *s*(bottom time, bs = "tp") + *s*(start time, bs = "cc") + te(latitude, longitude, bs = "tp", by = vetula breeding dummy variable).

For comparison of different modeling techniques, three different interpretations of data were fit with GAMs using the MGCV package (Wood, 2017): binomial for presence/absence data, negative binomial for means imputed from abundance categories using Equation 1, and multinomial for abundance categories. The multinomial model was more limited than other models. Each category was fit with a different smooth function for the same explanatory variable, which significantly increased computation time to fit the model (i.e., the simplest multinomial model required a longer time to fit than the full AIC selected negative binomial model). Additionally, the complexity of the multinomial model was limited due to the division of data in each category being much smaller than the total dataset. In particular, fewer than 300 observations were in the "Many" category for this species, so a model with comparable complexity could not be fit for this category. Due to increased time to fit and increased complexity, a full AIC selection was not conducted for the multinomial model. The only parameters considered for this simpler model were a latitude-longitude tensor product and decimal date fit with thin-plate regressor splines. For comparison, a second negative binomial model was fit to compare to the multinomial model using the same variables. For negative binomial models, model fit was evaluated using scaled residuals calculated by the DHARMa residual library in R (Hartig, 2022). Simulated residuals, based on model assumptions, of actual data fitted to simulated data, were compared to test if the model was appropriately specified and if underlying assumptions were correct.

All models were compared using a 5-fold cross-validation. Testing subset R^2, RMSE, and average error were calculated for negative binomial and binomial models. For the binomial model, kappa, specificity, sensitivity, and AUC were calculated using the PresenceAbsence R package (Freeman & Moisen, 2008; Appendix S1). For the negative binomial model, predictions were coerced to SFM categories (zeros included) following the same protocol as divers and the percent of correctly predicted categories was calculated. For the multinomial model, the percentage of correctly predicted categories was used as the sole metric.

To determine population trends for Queen Triggerfish, binomial and negative binomial models were fit with the year (1994–2020) as a categorical variable for testing the significance of a potential temporal trend in presence and abundance. The anova.gam function in mgcv was used to determine whether the linear trend was significant, based on approximate p-values, which were sufficient for our analysis (Wood, 2017). Confidence intervals on relative abundance were calculated using 95% confidence intervals for each year from the GAM, applied to abundances imputed from categories. These confidence intervals were produced based on the assumption made by the GAM that it is fitting to negative binomial observations, and do not account for the error introduced through the imputation of categorical data.

3 | RESULTS

3.1 | Simulated data

Models fit to categorized data performed worse for all metrics than models fit to original data (Table 2). The exponential model had an overwhelmingly negative R^2 and large RMSE and was by far the worst-fitting model. The three best models converged to an asymptotic range of values after <5000 datapoints (Figure 4). Fitted metrics converged relatively quickly, so the additional variance in those metrics, relative to the uncategorized fit model, was primarily due to error of the imputation method rather than a variation in sample size (Table 2). Average error of mean value imputation was closest to zero for all categorical imputed models and resulted in the highest R^2. The R^2 of mean imputed model predictions relative to uncategorized data was also underestimated by R^2 of predictions relative to the value imputed into the model. The reverse was true for minimum value imputation. Minimum value imputation and mean value imputation both had a similar slope, although minimum value imputation was shifted down with a negative intercept. The mean value method was slightly overestimated, and the minimum value method severely underestimated the original uncategorized data.

3.2 | REEF data

The AIC selection process left the binomial model with 11 variables and the negative binomial model with 9 variables (Table 3). Lunar phase was excluded from both the binomial and negative binomial

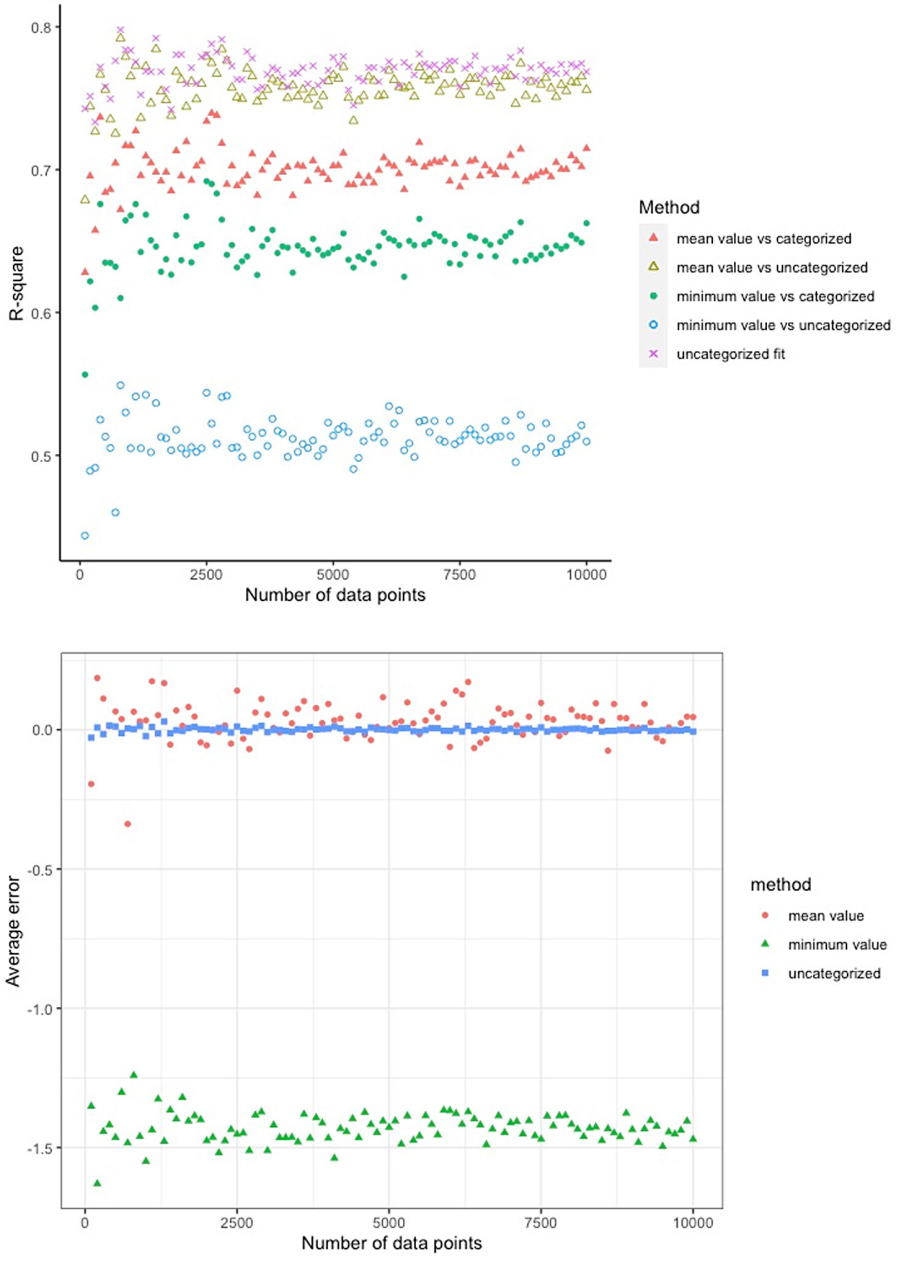

FIGURE 4 Goodness of fit (R^2) of predictions (Top) and average error of predicted values (Bottom) for the three best models relative to both uncategorized data and imputed data, based on citizen-science data (Reef Environmental Education Foundation, REEF) of Queen Triggerfish (*Balistes vetula*) in The Bahamas and Turks and Caicos during 1994–2020.

TABLE 3 Binomial and negative binomial ΔAIC scores for the best-fit model, with each variable removed (see Table 1 for variable definitions), for estimating the presence and abundance of Queen Triggerfish (*Balistes vetula*) in The Bahamas and Turks and Caicos during 1994–2020, based on marine citizen-science data (Reef Environmental Education Foundation, REEF).

Model terms omitted	Binomial ΔAIC	Negative binomial ΔAIC
Full model	0	0
Visibility	NA	15.32
Experience	49.39	15.38
Current	20.25	4.19
Bottom temperature	13.61	58.83
Decimal day of year	29.17	NA
Decimal year	125.48	188.69
Bottom time	88.78	62.92
Start time	103.96	50.14
Average depth	22.72	NA
Habitat	9.26	25.92
Latitude/longitude tensor	514.26	411.65
Tensor by = breeding variable	14.98	NA

Abbreviation: NA, term not included in the best-fit model.

models, Breeding season, decimal day of the year, and average depth were excluded from the negative binomial model, and visibility was excluded from the presence-absence.

Model predictive performance in cross-validation was poor for both negative binomial and binomial models (Table 3, Table 4). The multinomial model predicted the correct SFMA category 68.2% (±1.1% standard deviation) of the time, and the simplified negative binomial model predicted the correct category 44.7% (±1.1%) of the time. The binomial model could not predict FMA categories, but correctly predicted presence-absence 72.5% (±0.6%) of the time. The values for AUC, Kappa, Specificity, and Sensitivity for the binomial model are shown in the (Appendix S1).

Most features were consistent among splines in the different models, in particular for negative binomial and binomial models. Positive observations decreased with the decimal year (Figure 5) and increased linearly with bottom time. The spatial distribution of Queen Triggerfish did not differ among models, and the spatial distribution during the breeding season had a single high anomaly in the southwest for the binomial model, although the lack of inclusion of breeding season indicated this could be a spurious feature of the binomial model (Appendix S1). Queen Triggerfish were most often observed during the afternoon in 76 °F water. The DHARMa residual plots did not show a clear pattern in simulated residuals for all models fitted to are included in (Appendix S1).

Year was highly significant ($p < 0.01$) as an annual trend (linearly related to date) in both binomial and negative binomial models, which indicated the Queen Triggerfish population declined since 1994. The probability of observation and relative abundance both declined more than 50% since 1994 (Figure 5).

TABLE 4 Mean and standard deviation of goodness of fit (R^2), root-mean-squared error (RMSE), average error, and percent of categories predicted correctly for binomial and negative binomial cross-validation of ~15,000 surveys used to estimate the presence and abundance of Queen Triggerfish (*Balistes vetula*) in The Bahamas and Turks and Caicos during 1994–2020, based on marine citizen-science data (Reef Environmental Education Foundation, REEF). Binomial category prediction categories (1–0) differed from negative binomial prediction categories (0-SFM). Full prediction metrics for the binomial model are in Appendix S1.

Model	Binomial	Negative binomial
Mean R^2	0.155	0.084
SD R^2	0.024	0.171
Mean RMSE	0.185	2.230
SD RMSE	5.85×10^{-3}	0.315
Mean average error	-1.03×10^{-4}	0.037
SD average error	0.017	0.037
Mean % categories predicted correctly	72.54%	49.10%
SD % categories predicted correctly	0.56%	0.86%

4 | DISCUSSION

4.1 | Model interpretation

Various methods have been developed to extract information on the underlying abundance from REEF categories (Campbell et al., 2022; Tolimieri et al., 2017; Wolfe & Pattengill-Semmens, 2013a). By testing these methods in simulation, it can be determined how appropriate they may be in different contexts. The near-identical slope of the mean and minimum imputation methods suggests that for analyzing trends both methods may perform comparably well. In absolute terms, though, the minimum value consistently underestimated the true abundance and would be inferior in cases where the intercept of the model had a biologically relevant interpretation. Both the mean and minimum value imputation methods also showed bias in the slope in simulated data, which could lead to a failure to detect a significant trend. This limitation of REEF data is not unique to the imputation method, however, since the log categorical nature of the data collection allows for changes in abundance without changes in observed categories (Campbell et al., 2022). The exponential density score imputation did not work well for imputing individual categories, although other studies have used density scores to track population trends without translating them to abundance (Campbell et al., 2022). Depending on the study system, the use of abundance information or the use of pure presence-absence information can provide better model performance (Fukuda et al., 2012; Howard et al., 2014). In this study, the models that incorporated abundance information (the negative binomial and multinomial) performed worse than the presence-absence model

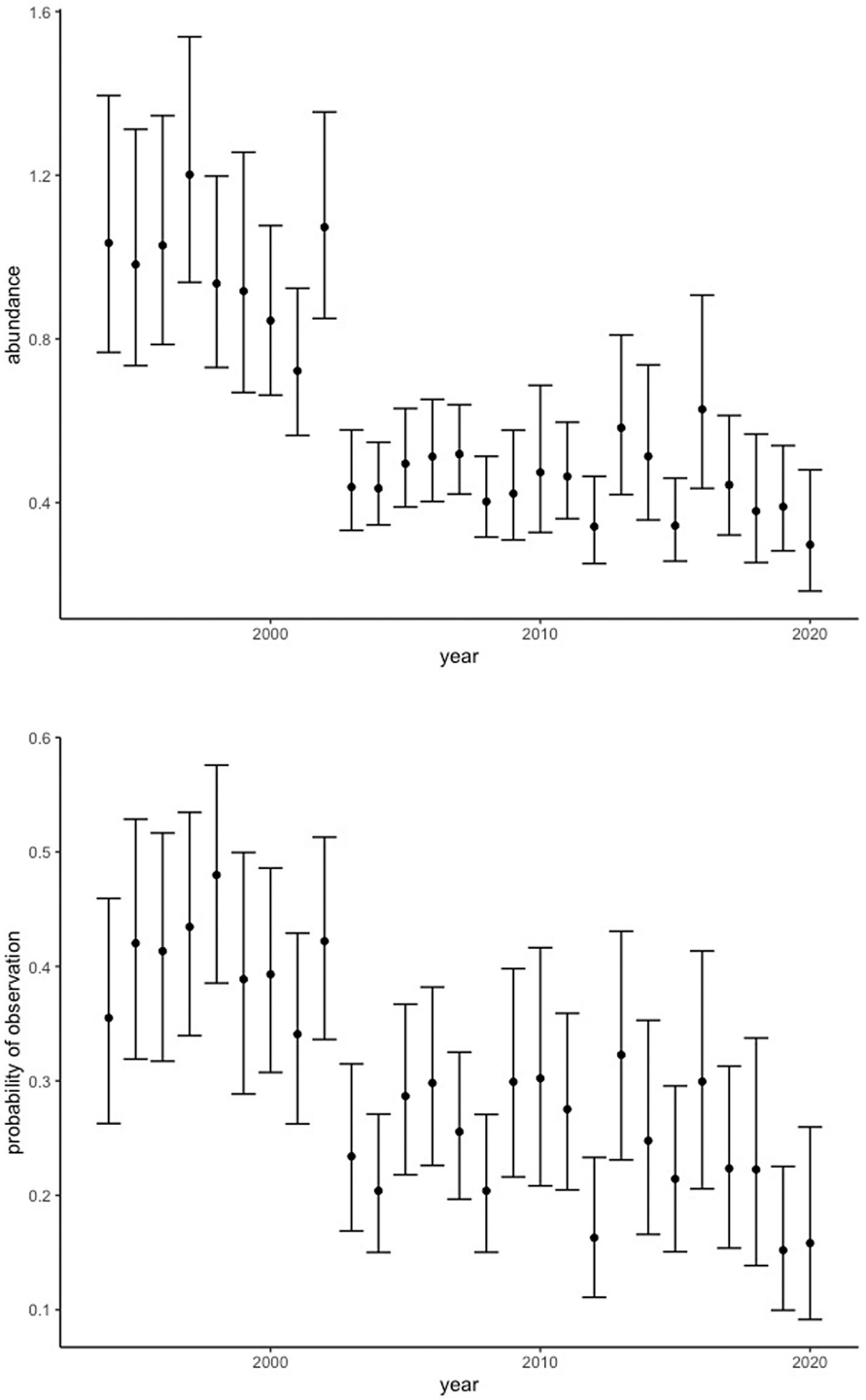

FIGURE 5 Relative abundance (Top) and probability of presence (Bottom), based on marine citizen-science data (Reef Environmental Education Foundation, REEF) (+95% confidence intervals), of Queen Triggerfish (*Balistes vetula*) in The Bahamas and Turks and Caicos during 1994–2020.

(binomial). This likely depends on the species and study system and more comprehensive research is required to determine the relative efficacy of different methods applied to this data.

Accounting for variation from sources other than changes in abundance, as in our study, is important for the estimation of abundance indices and CPUE from fisheries data (Maunder & Punt, 2004). For other citizen-science data, similar work has been done to account for observer effort, which can confound results if not considered (Johnston et al., 2021). Differences between abundance and presence-absence models can also inform decisions. Exclusion of the breeding variable in the abundance model suggests that either breeding behavior does not affect local abundances of Queen Triggerfish, or that there is not enough data to accurately predict over that range. The second seems more likely since this species forms seasonal breeding aggregations (Bryan et al., 2019). Exclusion of visibility from the presence-absence model, but inclusion of visibility in the abundance model, suggests that visibility hinders the counting of many individuals, but not the detection of the species when present. The 50% decline in the presence and abundance of Queen Triggerfish in The Bahamas and Turks and Caicos over the

last 20 years we detected is immediately relevant for species conservation and suggests that further investigation and management intervention may be warranted. Causes of this decline likely include fishing pressure, changes in diet, and habitat loss (Bryan et al., 2019; Hernández et al., 2019; Reinthal et al., 1984; Sadovy De Mitcheson et al., 2008). Current classification of Queen Triggerfish by the IUCN Red List is based on surveys of population trends from fisheries data within the Caribbean, so surveys of population trends are needed in The Bahamas and Turks and Caicos (Liu et al., 2015).

For many coastal fisheries, limited data on catch or abundance limits the ability of traditional models to detect population trends (Sagarese et al., 2018). REEF data can provide broad-scale, inexpensive data that is independent of fisheries catch for coastal species (Campbell et al., 2022). By translating categorical abundance data into values that can be modeled at the survey level, we developed and tested a simple and effective way to account for variation in survey characteristics while predicting overall trends in abundance of Queen Triggerfish in The Bahamas and Turks and Caicos. We found that trends in abundance and presence of Queen Triggerfish differed little, but the same may not be true for other species (Campbell et al., 2022). These methods could be quickly applied to other coastal fisheries and could be used to further incorporate REEF data into the IUCN red list process beyond presence-absence models already in use (Gravem et al., 2020).

4.2 | Potential future research

Surveyor experience, species identification, data screening, and missing data all likely affected our findings. The binary surveyor experience level used by REEF was a significant factor for the models used in our study. We used the Queen Triggerfish as a model organism due to its ease of identification and observation, whereas other species are likely to be more difficult to identify or observe. Incorporating information about the relative detectability of species might be necessary for multispecies studies (Ashley et al., 2022). Further exploration is needed of the interplay between surveyor experience and species identification. For example, other surveyor experience metrics could improve model performance by creating a continuous experience metric (Kelling et al., 2015). Surveyor experience also likely affects missing data. We omitted data if records in multiple fields were missing, which could bias results if data is not randomly missing. Future studies could account for missing data either through multiple imputations or data augmentation (Nakagawa & Freckleton, 2008). Such analyses were outside the scope of our study, but given that roughly half of the data available for our study was omitted due to problems with missing data, the need for better missing data handling is an important area for potential future research.

We were only able to examine variables collected by REEF surveyors, while broader explorations of multiple long-term monitoring datasets have successfully integrated REEF observations (Grüss et al., 2018). Similar to presence-absence modeling that integrated REEF observations, integration with actual counts using similar methods is promising. The precision of REEF abundance category observations was lower than actual counts, but the benefits of its inclusion could overcome quality concerns. Similarly, many of the explanatory variables collected by REEF could be supplemented by other sources, such as information from coral reef mapping. Divers often swim through multiple habitats while on a dive and many divers do not enter habitat information into their survey data. The use of external sources would increase the amount of usable data for modeling and thereby improve performance. In addition to external environmental data, actual count data from the same region could be compared to model-based estimates of presence and abundance. The conversion function we used in our analysis was constructed by comparing actual counts to REEF categories, but the models we used have yet to be validated in this way (Wolfe & Pattengill-Semmens, 2013a, 2013b).

More sophisticated imputation methods could be used to infer counts from SFMA categories. Substituting the inferred mean is a common and simple imputation method, and performed well in our study. Future work could use a more complex model to infer means, perhaps by separating data by years or habitat types, for which distributions of abundance categories may differ. Methods such as multiple imputation or Bayesian methods could be used to assign each imputed count to a theoretical distribution of values based on a truncated lognormal distribution within a range of possible values (e.g., 2 to 10 for *F*) (Nakagawa & Freckleton, 2008). Such methods might more accurately capture the uncertainty in estimated relationships between variables because they would include imputation error in the *y* variable. However, such methods would lack the simplicity and ease of use of our method.

5 | CONCLUSIONS

Models explored herein suggested a biologically relevant decline in Queen Triggerfish abundance and were able to account for several variables collected within the REEF dataset. Similar methods may be useful for expanding the analysis of REEF abundance data to other species in other areas while accounting for variance introduced by surveyor behavior and skill. Future research should compare model-based relative abundance estimates to real-world counts, incorporate other environmental variables, and explore other methods for incorporating surveyor experience. Our model-based estimates of abundance based on REEF data imply that Queen Triggerfish are declining in The Bahamas and Turks and Caicos and may require management interventions to reverse the decline.

ACKNOWLEDGMENTS

Thanks to Dr. Kathleen Sullivan Sealey for assistance in accessing REEF data and review of the written work. Thanks Dr. Michael Schmale for introducing the potential of the REEF dataset to create abundance indices. Finally, thanks to all of the recreational SCUBA divers who participated in gathering this data.

FUNDING INFORMATION

No authors received funding for this work.

CONFLICT OF INTEREST STATEMENT

Authors declare no conflict of interest with this work.

DATA AVAILABILITY STATEMENT

Data are available from reef.org, REEF requests that the data are not distributed to other parties without prior notification to REEF's Co-Executive Director of Science and Engagement and that the geographic location data are not shared or published.

REFERENCES

Ashley, E.A., Pattengill-Semmens, C.V., Orr, J.W., Nichols, J.D. & Gaydos, J.K. (2022) Documenting fishes in an inland sea with citizen scientist diver surveys: using taxonomic expertise to inform the observation potential of fish species. *Environmental Monitoring and Assessment*, *194*, 227. Available from: https://doi.org/10.1007/s10661-022-09857-1

Birk, M. A. (2019). measurements: Tools for units of measurement. R package version 1.4.0. URL: https://CRAN.R-project.org/package=measurements

Bryan, D.R., Feeley, M.W., Nemeth, R.S., Pollock, C. & Ault, J.S. (2019) Home range and spawning migration patterns of queen triggerfish Balistes vetula in St. Croix, US Virgin Islands. *Marine Ecology Progress Series*, *616*, 123–139.

Campbell, J., Yakimishyn, J., Haggarty, D., Juanes, F. & Dudas, S. (2022) Citizen science surveys provide novel nearshore data. *Fisheries*, *48*, 8–19. https://doi.org/10.1002/fsh.10831

Freeman, E.A. & Moisen, G. (2008) PresenceAbsence: an R package for presence-absence model analysis. *Journal of Statistical Software*, *23*(11), 1–31.

Fukuda, S., Mouton, A.M. & De Baets, B. (2012) Abundance versus presence/absence data for modelling fish habitat preference with a genetic Takagi–Sugeno fuzzy system. *Environmental Monitoring and Assessment*, *184*(10), 6159–6171.

Grüss, A., Chagaris, D.D., Babcock, E.A. & Tarnecki, J.H. (2018) Assisting ecosystem-based fisheries management efforts using a comprehensive survey database, a large environmental database, and generalized additive models. *Marine and Coastal Fisheries*, *10*(1), 40–70. Available from: https://doi.org/10.1002/mcf2.10002

Gravem, S.A., Heady, W.N., Saccomanno, V.R., Alvstad, K.F., Gehman, A.L.M., Frierson, T.N. et al. (2020) *Pycnopodia helianthoides*. The IUCN red list of threatened species 2020 e.T178290276A178341498.

Grolemund, G. & Wickham, H. (2011) Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*, 25.

Hartig, F. (2022) Residual diagnostics for hierarchical (Multi-Level / Mixed) regression models. Version 0.4.6. URL: https://florianhartig.github.io/DHARMa/

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Berlin: Springer Science & Business Media.

Hernández, R., Jesús, M., Alvarado, N.P., Vélez, K.C., Nemeth, R., Appeldoorn, R. et al. (2019) Queen triggerfish reproductive biology in U.S. Caribbean waters. *Transactions of the American Fisheries Society*, *148*(1), 134–147. Available from: https://doi.org/10.1002/tafs.10124

Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D. & Willis, S.G. (2014) Improving species distribution models: the value of data on abundance. *Methods in Ecology and Evolution*, *5*(6), 506–513.

Johnston, A., Hochachka, W.M., Strimas-Mackey, M.E., Ruiz Gutierrez, V., Robinson, O.J., Miller, E.T. et al. (2021) Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265–1277.

Joseph, L.N., Field, S.A., Wilcox, C. & Possingham, H.P. (2006) Presence–absence versus abundance data for monitoring threatened species. *Conservation Biology*, *20*, 1679–1687. Available from: https://doi.org/10.1111/j.1523-1739.2006.00529.x

Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J. et al. (2015) Can observation skills of citizen scientists Be estimated using species accumulation curves? *PLoS One*, *10*(10), e0139600. Available from: https://doi.org/10.1371/journal.pone.0139600 lunar: Lunar Phase & Distance, Seasons and Other Environmental Factors.

Lazaridis, E. (2022). lunar: Lunar phase & distance, seasons and other environmental factors (version 0.2-01). Available from CRAN

Liu, J., Zapfe, G., Shao, K.-T., Leis, J.L., Matsuura, K., Hardy, G. et al. (2015) *Balistes vetula*. *The IUCN Red List of Threatened Species 2015*. e.T2539A97664057.

Maunder, M.N. & Punt, A.E. (2004) Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*, *70*(2–3), 141–159.

Montecino-Latorre, D., Eisenlord, M.E., Turner, M., Yoshioka, R., Harvell, C.D., Pattengill-Semmens, C.V. et al. (2016) Devastating transboundary impacts of sea star wasting disease on subtidal asteroids. *PLoS One*, *11*(10), e0163190.

Nakagawa, S. & Freckleton, R.P. (2008) Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592–596. Available from: https://doi.org/10.1016/j.tree.2008.06.014

Pattengill-Semmens, C.V. & Semmens, B.X. (2003) Conservation and management applications of the reef volunteer fish monitoring program. In: *Coastal monitoring through partnerships*. London: Springer Nature, pp. 43–50.

R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/

Rassweiler, A., Dubel, A.K., Hernan, G., Kushner, D.J., Caselle, J.E., Sprague, J.L. et al. (2020) Roving divers surveying fish in fixed areas capture similar patterns in biogeography but different estimates of density when compared with belt transects. *Frontiers in Marine Science*, *7*, 272.

REEF. (2021) *Reef environmental education foundation volunteer fish survey project database*. World Wide Web electronic publication. Accessed February 26, 2021. URL: www.REEF.org

Reinthal, P.N., Kensley, B. & Lewis, S.M. (1984) Dietary shifts in the queen triggerfish, Balistes vetula, in the absence of its primary food item. *Diadema Antillarum*. *Marine Ecology*, *5*(2), 191–195.

RStudio. (2022) *Integrated development for RStudio*. PBC, Boston.

Sadovy De Mitcheson, Y., Cornish, A., Domeier, M., Colin, P.L., Russell, M. & Lindeman, K.C. (2008) A global baseline for spawning aggregations of reef fishes. *Conservation Biology*, *22*(5), 1233–1244. Available from: https://doi.org/10.1111/j.1523-1739.2008.01020.x

Sagarese, S., Rios, A., Cass-Calay, S., Cummings, N., Bryan, M., Stevens, M. et al. (2018) Working towards a framework for stock evaluations in data-limited fisheries. *North American Journal of Fisheries Management*., *38*, 507–537. Available from: https://doi.org/10.1002/nafm.10047

Tolimieri, N., Holmes, E.E., Williams, G.D., Pacunski, R. & Lowry, D. (2017) Population assessment using multivariate time-series analysis: a case study of rockfishes in Puget Sound. *Ecology and Evolution*, *7*(8), 2846–2860.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.'.A., François, R. et al. (2019) Welcome to the tidyverse. *Journal of Open Source Software*, *4*, 1688. Available from: https://doi.org/10.21105/joss.01686

Wolfe, J.R. & Pattengill-Semmens, C.H.R.I.S.T.Y.V. (2013b) Fish population fluctuation estimates based on fifteen years of reef volunteer diver data for the Monterey peninsula, California. *California Cooperative Oceanic Fisheries Investigations Report*, *54*, 141–154.

Wolfe, J.R. & Pattengill-Semmens, C. (2013a) Estimating fish populations from REEF citizen science volunteer diver order-of-magnitude surveys. *California Cooperative Oceanic Fisheries Investigations Reports*, *54*, 127–140.

Wood, S.N. (2017) *Generalized additive models: an introduction with R*, 2nd edition. Boca Raton: CRC Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Urquhart, J P., Olson, D B. & Babcock, E A. (2023). Generalized additive models for categorical count data: An exploration of the decline of queen triggerfish *Balistes vetula* in the Bahamas and Turks and Caicos. *Fisheries Management and Ecology*, *00*, 1–12. https://doi.org/10.1111/fme.12617