Two-Stage Boosted Regression Tree Model to Characterize Southern Flounder Distribution in Texas Estuaries at Varying Population Sizes

Authors: John T. Froeschke, and Bridgette F. Froeschke

Source: Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem Science, 8(8) : 222-231

Published By: American Fisheries Society

URL: https://doi.org/10.1080/19425120.2015.1079577
Two-Stage Boosted Regression Tree Model to Characterize Southern Flounder Distribution in Texas Estuaries at Varying Population Sizes

John T. Froeschke*
Gulf of Mexico Fishery Management Council, 2203 North Lois Drive, Suite 1100, Tampa, Florida 33511, USA

Bridgette F. Froeschke
Biology Department, College of Natural and Health Sciences, University of Tampa, 401 West Kennedy Boulevard, Box U, Tampa, Florida 33606, USA

Abstract

Linking trends in fish population abundance to environmental characteristics is often difficult because fish use a variety of habitats throughout their ontogeny and may exhibit large interannual fluctuations in abundance. We developed a two-stage boosted regression tree model to investigate spatiotemporal patterns of Southern Flounder Paralichthys lethostigma abundance and distribution in Texas estuaries. We used a 36-year fishery-independent data set (1977–2012) to correlate distribution with environmental conditions and seasonal or long-term changes in abundance. Adult Southern Flounder were sampled with gill nets using a random-stratified design. Predictions of abundance were made to grids of environmental data to identify “hot spots” as well as seasonal or decadal shifts in distribution. Models were fit using cross validation, and variance was estimated using nonparametric bootstrapping. Depth, temperature, distance to a tidal inlet, and salinity were the primary environmental determinants of Southern Flounder distribution and abundance. Because distribution and response to environmental conditions can depend on population abundance, we also developed a standardized index of annual abundance using the same two-stage boosted regression tree model. The index identifies a long-term decline in abundance punctuated by recent recovery (possibly in response to management actions). Mapped results identify a coastwide decline in abundance between 1980–1984 and 2005–2009 but the magnitude varied substantially, suggesting disproportionate changes in abundance across the study area.

Characterizing spatial and temporal patterns of abundance is a cornerstone of the management of exploited species (Hildalgo et al. 2011; Campbell 2015; Thorson et al. 2015). Historically, resource management has focused on characterizing temporal patterns of abundance (Lo et al. 1992; Campbell 2015) using time series analyses or regression models that may include environmental variables to improve estimates of changes in stock abundance (Maunder...
and Punt 2004; Froeschke et al. 2013a). Growing interest in ecosystem-based fishery management (Pikitch et al. 2004) has elevated the need for spatially explicit descriptions of species distributions in relation to environmental factors and fish population size (Essington and Punt 2011; Drexler and Ainsworth 2013; Collie et al. 2016). To properly interpret these data, it is appropriate to consider the relative abundance of a species during the period from which the survey data used to build the model are collected (Jensen et al. 2005). This seems less common in practice as this information is often unknown, but recent research has demonstrated that species–environment relationships can change with abundance and that abundance may be more closely linked to environmental conditions at low population sizes (Hildalgo et al. 2011).

In the Gulf of Mexico, Southern Flounder Paralichthys lethostigma are harvested both recreationally and commercially and support an economically important fishery, but declines in abundance (Froeschke et al. 2011) have led to reduced recreational and commercial catches. Southern Flounder populations declined from the 1970s to 2008 in Texas (Froeschke et al. 2011) before rebounding in recent years, coinciding with more restrictive fishery regulations. Texas recently reduced daily possession limits (March 2009) from 10 to 5 fish for every month but November. Peak spawning occurs in November, and adults migrate offshore to spawn and are more vulnerable to fishing pressure during migration. Thus, anglers are limited to a 2-fish daily bag limit during that month and gigging (harvesting using a multipronged spear) is prohibited in both commercial and recreational fishing.

Despite dramatic declines in the abundance of Southern Flounder in Texas, declines in recruitment have been relatively minor (Froeschke et al. 2011). The decline in adult Southern Flounder abundance has likely resulted from increased adult or subadult mortality (Froeschke et al. 2011). However, very little is known about the spatial and temporal environmental requirements for adult Southern Flounder and how their distribution changes with abundance. This has implications for resource managers and harvesters, as some areas may be disproportionately affected by changes in abundance. We developed statistical models to characterize species–environment relationships for Southern Flounder using boosted regression trees (BRTs). This is a model-averaging (ensemble) method that allows for both interactive and nonlinear relationships and often has a predictive performance superior to that of other common methods (e.g., generalized linear models; Elith et al. 2008). The objectives of this study were (1) to characterize species–habitat relationships for Southern Flounder, (2) to depict their spatial distribution patterns, and (3) to evaluate changes in their distribution relative to changes in their abundance during the survey period.

METHODS

Field collections.—Fishery-independent Southern Flounder catch data were provided by the Texas Parks and Wildlife Department from its estuarine gill-net survey initiated in 1975. For this analysis, data were included from 1977 through 2012. Data were collected using a stratified cluster sampling design in which each estuary was a stratum with a fixed effort (n = 45 per estuary per season [20 per estuary per season in Sabine Lake]). Sampling occurred each spring (April, May, and June) and fall (September, October, and November); specific sample sites were selected independently and without replacement each season (Martinez-Andrade et al. 2009). Gill nets were 183 m in length and set perpendicularly to the shore. Each net contained four panels with stretched mesh sizes of 76, 102, 127, and 152 mm. The gill nets were deployed approximately 1 h prior to sunset, allowed to soak overnight, and retrieved the following morning.

Southern Flounder are strongly influenced by abiotic environmental conditions, including temperature (Froeschke et al. 2013b, 2013c) and salinity (Froeschke et al. 2013c). Salinity, temperature, dissolved oxygen, and depth were measured concomitantly with the gill-net sets and were treated as explanatory variables in the model (described below). Sampling month and year were also included as variables. Soak time was used as an additional explanatory variable to control for effort-related effects on catch (Maunder and Punt 2004). The distance from each sample location to the nearest tidal inlet was calculated using the cost-distance function in ArcGIS (Whaley et al. 2007; Froeschke et al. 2010). Two cost-distance surfaces were created, one without Packery Channel inlet (which did not open until June 2005; Figure 1) and one that included Packery Channel inlet (which applied to all sampling events after June 2005).

Modeling approach.—Preliminary analyses indicated that Southern Flounder catch data were “zero inflated” with a long tail (i.e., there were a few samples with very large catches), which is typical of many fishery data sets (Martin et al. 2005; Arab et al. 2008). A delta-lognormal approach (Lo et al. 1992; Ingram et al. 2010) was used to develop a standardized index of abundance across time and space. This method combines two submodels: one including an analysis of the probability of capture (occurrence) and one including an analysis of the log-transformed catch on positive sets (abundance). The delta approach uses these submodels to construct a single, standardized CPUE index (Lo et al. 1992).

The standardized CPUE (\(L_i\)) is defined as

\[ L_i = c_i p_i, \]
where $c_i$ is the estimate of $\ln$(catch) for each positive sample $i$ and $p_i$ is the estimated probability of occurrence for sample $i$. Both $c_i$ and $p_i$ were estimated using boosted regression trees (Elith et al. 2008). The Southern Flounder catch data were assumed to have a lognormal distribution for abundance and a binomial distribution for occurrence and were modeled using the following equations:

$$\ln(c) = X\beta + \varepsilon$$  

and

$$p = e^{X\beta + \varepsilon} / (1 + e^{X\beta + \varepsilon})$$

where $c$ is a vector of positive catch data, $p$ is a vector of presence/absence data, $X$ is the design matrix for the main effects, $\beta$ is a vector of parameters for the main effects, and $\varepsilon$ is a vector of independent, normally distributed errors with mean 0 and variance $\sigma^2$ (Ingram et al. 2010). A standardized delta index for each calendar year was calculated as the annual mean delta value for all samples within the year. Confidence limits were estimated using nonparametric bootstrapping with replacement ($n = 500$; Efron and Tibshirani 1993). Analyses were completed in R (version x64 3.1.; R Development Core Team) using the “gbm” library and functions from Elith et al. (2008) and J. T. Froeschke (unpublished).

The relative importance of predictors was estimated by averaging the number of times a variable was selected for splitting and the squared improvement resulting from these splits (Friedman 2001; Friedman and Meulman 2003). Response values ranged from 0 to 100, where larger numbers indicate a greater influence on the dependent variable. A tree complexity parameter controlled the number of nodes and allowed the interactive effects to be modeled in an automated manner with minimal impacts on overfitting (Leathwick et al. 2006). A learning rate parameter was used to control the relative contribution of individual trees to the ensemble model and to balance model performance with computation time.

FIGURE 1. Gill-net sample locations ($n = 24,756$) from 1977 to 2012.
Model fitting and selection.—To estimate the probability of occurrence of Southern Flounder in a given sample based on the covariates, a BRT was fit to these data using a tree complexity of 5 and a learning rate of 0.001 with a binomial error distribution. Trees were added sequentially to minimize cross-validated residual deviance without overfitting. The positive samples were log transformed and a BRT was fit to these data using a tree complexity of 5, a learning rate of 0.001, and a Gaussian error distribution. A model simplification routine (“gbm.simplify”) similar to a backward selection routine in regression (Elith et al. 2008) was used. Variables were ranked in order of decreasing importance and were removed until significant increases in residual deviance were found between the full and reduced models. This routine was employed separately for both the occurrence and abundance models and permitted different explanatory variables to be included in each submodel. A delta value for each sample was computed as the product of the probability of occurrence and the unlogged value from the abundance submodel.

Spatial grids.—Environmental variables were measured at each gill-net set (n = 24,756) and subsequently interpolated into raster grids using ordinary kriging (Saveliev et al. 2007; Elith et al. 2008; Froeschke and Froeschke 2011) via the “autoKrige” function in the automap package in R (Hiemstra et al. 2009). Environmental grids were developed for each month (April, May, June, September, October, and November) and year combination between 1980 and 2012 to permit spatially specific predictions during specific months and/or time periods. Depth was interpolated into a single layer (all sample months and years) based on measured depths during sampling using inverse-distance weighting, which is less computationally intensive than ordinary kriging yet performs well with large, closely spaced samples.

The fitted delta-BRT model was used to predict relative abundance to the mapped grids of environmental conditions across the study area for each month–year combination. Predictions were averaged across months to examine monthly distribution patterns. Predictions were made to environmental grids in two periods: (1) 1980–1984 (high abundance) and (2) 2005–2009 (low abundance). Finally, the percent change of the fitted grids between time periods was calculated (i.e., ([low abundance – high abundance]/ high abundance) × 100) to identify spatially explicit percent changes of predicted relative abundance throughout the study area.

Time series.—Long-term changes in relative abundance were determined by multiplicatively combining the occurrence and abundance submodels. Mean annual CPUE was then calculated as the mean of all samples within a given year. The variance of annual CPUE was calculated using 500 nonparametric bootstrap iterations (with resampling) based on randomized subsets of 70% of these sample data.

RESULTS
Southern Flounder Distribution and Habitat Modeling
Southern Flounder were captured in 9,486 of 24,756 samples (39.8%). Abundance per sample ranged from 0 to 31 (Figure 2), with the vast majority of samples capturing less than 5 individuals. Southern Flounder were captured each month during the survey. However, seasonality was observed, as both the frequency of occurrence and catch increased each month from April to November, likely associated with fall spawning activities. Distribution and abundance patterns were affected by all eight variables included in the model (Figure 3). In the frequency-of-occurrence submodel, the most important predictors were depth (occurrence decreases with depth), salinity (which produced a bell-shaped curve peaking at ~25 psu), year sampled, and distance to an inlet. Time was included in the model to account for intersample variation in set time, and although it was the least important variable it was still nonnegligible, accounting for 4.8% of the explained variance.

In the abundance submodel, only five of the eight variables considered contributed significantly. Distance to an inlet, month sampled, and depth were the most important contributors to the explained deviance (Figure 4). In relation to distance to an inlet, abundance had a bimodal distribution, as sample locations at intermediate distances from the tidal inlets yielded the fewest animals in positive samples. As in the occurrence model, fitted abundance increased every month from April to November.

Time Series
The resulting time series for CPUE shows an approximately linear decline from 1977 to 2008. From 2009 to 2012, however, CPUE increased rapidly and by 2011 attained levels not observed since the early 1990s (Figure 5).

Spatial Patterns
Predicted abundance increased each month, with the highest intensity near tidal inlets in the fall (October and November). Abundance also increased modestly from the north to south in a pattern that is consistent across the months sampled (Figure 6). As spatial distribution patterns can be affected by population abundance, the predicted distribution was also derived for periods of relatively high (1980–1984) and low (2005–2009) abundance. During the period of high abundance, Southern Flounder were more evenly distributed across the bays, with typical seasonal peaks in October and November. Coastwide reductions in CPUE were observed during the study, reaching their lowest levels during the period of low abundance. To examine spatially explicit changes in CPUE between time periods, the percent change between 1980–1984 and 2005–2009 was calculated for each cell (Figure 7). The greatest declines were observed in the...
southernmost estuaries (upper and lower Laguna Madre). Notable declines were also evident in Baffin Bay during the October–November spawning period. South Texas estuaries support the greatest proportion of age-2 and older Southern Flounder, and this was the most variable region when periods of historically high and low abundance were compared.

**DISCUSSION**

The distribution and abundance of age-2 and older Southern Flounder were influenced by temporal, physical, and spatial processes, and correlative relationships were identified for both the probability of occurrence and abundance using a two-stage boosted regression tree model. The probability of occurrence was most closely related to depth, salinity, and year sampled. It was greatest at locations near tidal inlets, at shallow depths, and during the fall months (October and November), and this pattern was consistent throughout the study period. Many estuarine species (including Southern Flounder) spawn offshore and juveniles recruit into estuaries via tidal inlets (Nañez-James et al. 2009). As a result, the abundance of such species is often greatest near inlets during the fall spawning season (Whaley et al. 2007; Froeschke et al. 2010, 2013b; Froeschke and Froeschke 2011). Our results also indicate greater abundance in the areas furthest from the inlets. We hypothesize that this reflects the abundance of Southern Flounder in the southern estuaries that are isolated from the tidal inlets. These areas are shallow, productive areas with large seagrass meadows and are distant from major human population areas (Mckee 2008). Overall, the current study considered inlets and depth ranges across a variety of habitat types, and it suggests that inlet proximity and depth are important features of habitat quality across biotic habitat types.

Salinity is probably the most important environmental variable influencing the distribution of macrofaunal organisms in northern Gulf of Mexico estuaries (Rakocinski et al. 1997). In this analysis, salinity was the most important environmental predictor of occurrence and the only water quality variable that was retained in the abundance submodel. In both submodels, the response to salinity was bell-shaped, peaking at ~25 psu with a range of 10–40 psu; this suggests that habitat affinity is reduced at elevated salinities. This pattern is consistent across many estuarine species and indicates some potential ramifications of reduced freshwater inflow into these bay systems as historic inflows are increasingly diverted for human usage (Montagna et al. 1992). In addition to providing salinity balance, freshwater inflows perform many other functions in estuaries, including the transport of sediment, nutrients, and allochthonous organic matter. They may also affect the timing and extent of migration of estuarine species (Longley 1994). For this reason, examination of salinity patterns across time and space can provide a valuable indicator of estuarine habitat condition and function. The
FIGURE 3. Smoothed partial-effects plots for the model of Southern Flounder occurrence. A boosted regression tree (BRT) model was used to relate the probability of occurrence to various environmental, spatial, and temporal variables. The solid lines represent the mean fits, the gray polygons the 95% confidence intervals of the partial effects determined using a generalized additive model. The y-axes are on the logit scale with mean zero; the x-axes have the following units: mean depth (m), salinity (psu), year, inlet distance (cost-distance units), month, temperature (°C), dissolved oxygen (DO; mg O₂/L), and soak time (h; included to account for minor variations in soak time among samples). The percentages indicate the proportions of explained deviance attributable to each predictor variable.
opening of the Packery Channel inlet in 2005 (Reese et al. 2008) at the northern end of the Laguna Madre and the recent dredging of the inlet near Port Mansfield may mitigate some of the concerns about increasing salinity; however, this also complicates the relationship between freshwater inflow and salinity. A full-scale analysis of spatiotemporal salinity patterns was beyond the scope of this study, but it remains an area in need of further investigation to improve
our understanding of species–environment relationships in estuaries.

The mapped distribution patterns from the combined delta-BRT model depict strong seasonal variation in abundance, peaking in the fall months (October and November). This pattern is most likely related to fall spawning behavior (Stunz 2000) and was consistent throughout the study period, albeit with changes in intensity related to varying stock abundance or environmental conditions. Spatially explicit distribution information is an important tool for the management of exploited species, but it is infrequently available for species at a scale of use to management (Leathwick et al. 2006). Fortunately, improved analytical tools and spatially explicit data make such information available for many species and regions, and this can allow better integration of the role of humans in the environment and an exploration of the effects of management actions on the resource throughout the management area (Colloca et al. 2003).

A unique feature of this study is that sampling occurred over a long time period (>35 years) and encompassed periods of both historically high and low abundance. Hildalgo et al. (2011) demonstrated that stock depletion can enhance the impact of environmental forcing on fish populations, as both the catch per unit effort of European hake Merluccius merluccius and its correlation with environmental variables increased

![Figure 5](https://bioone.org/journals/Marine-and-Coastal-Fisheries:-Dynamics,-Management,-and-Ecosystem-Science) Mean annual standardized CPUE (i.e., the predicted number of fish per net set) of Southern Flounder in Texas coastal waters, 1977–2012. The solid line shows the mean annual fit from the delta-BRT model; the gray lines reflect individuals bootstrap iterations (n = 500).

![Figure 6](https://bioone.org/journals/Marine-and-Coastal-Fisheries:-Dynamics,-Management,-and-Ecosystem-Science) Monthly predictions of Southern Flounder CPUE from the fitted delta-BRT model averaged across the years 1980–2012.
during periods of low population size. A similar result is plausible with Southern Flounder and would result in increased risk of decline during periods (or regions) with unfavorable environmental conditions even if the declines were driven by fishing exploitation rather than environmental forcing. From a fishery perspective this could affect our ability to determine appropriate harvest levels, and our results suggest that the impacts would be unevenly distributed across the study area. The historical decline of Southern Flounder in Texas is thought to be driven by exploitation (Stunz 2000; Froeschke et al. 2011). The more restrictive harvest regulations implemented in recent years (https://tpwd.texas.gov) coincide with increasing Southern Flounder CPUE through the terminal year of the study period.

As with all statistical modeling approaches, there are limitations to this methodology because data mining techniques can only identify relationships (correlations) if they actually exist (Brodley et al. 1999) and the residual deviance in the BRT submodels suggests that not all of the variables affecting distribution and abundance were included. While some of these factors can be examined in future studies (e.g., the role of predators and food web dynamics), other, more complex patterns (e.g., ocean circulation and long-term climate patterns) may continue to challenge this and other approaches trying to link animal distribution with the environment. In summary, fisheries management practices in many regions are aiming toward ecosystem-based approaches that consider aspects (including spatial distribution and abundance patterns) that enable identification of critical habitat or predict the effects of place-based perturbations (Pikitch et al. 2004). This study demonstrates that the relative stock size during the sampling period greatly affects spatial distribution patterns and can affect our identification of the regions and environmental conditions that support fish populations.

ACKNOWLEDGMENTS

Gill-net data for adult Southern Flounder were courtesy of Mark Fisher of the Texas Parks and Wildlife Department, Coastal Fisheries Division.
REFERENCES


Ingram, G. W., W. J. Richards, J. T. Lamkin, and B. Muhlting. 2010. Annual indices of Atlantic Bluefin Tuna (Thunnus thynnus) larvae in the Gulf of Mexico developed using delta-lognormal and multivariate models. Aquatic Living Resources 23:35–47.


