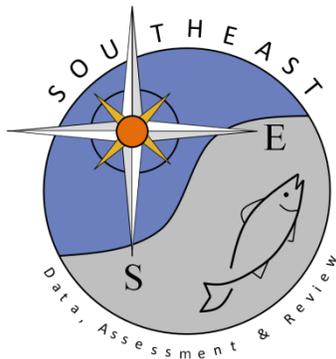


Atlantic Sharpnose Shark: Standardized index of relative abundance using boosted regression trees and generalized linear models

John Froeschke and Marcus Drymon

SEDAR34-WP-12

10 June 2013



This information is distributed solely for the purpose of pre-dissemination peer review. It does not represent and should not be construed to represent any agency determination or policy.

Please cite this document as:

Froeschke, J. and M. Drymon. 2013. Atlantic Sharpnose Shark: Standardized index of relative abundance using boosted regression trees and generalized linear models. SEDAR34-WP-12. SEDAR, North Charleston, SC. 31 pp.

Atlantic Sharpnose Shark: Standardized index of
abundance using boosted regression trees and
generalized linear models

John Froeschke and Marcus Drymon

June 10, 2013

List of Tables

1	Summary of variables selected for binomial GLM model. Model selection used an exhaustive search algorithm and models were selected using Akaike’s information criterion (AIC). The 20 best models are illustrated and the model with the lowest AIC value that included year as a covariate was selected as ”best”.	20
2	Summary of variables selected for log-normal GLM model. Model selection used an exhaustive search algorithm and models were selected using Akaike’s information criterion (AIC). The 20 best models are illustrated and the model with the lowest AIC value that included year as a covariate was selected as ”best”	21
3	Abundance index statistics for Atlantic sharpnose shark sampled during the Dauphin Island Sea Lab (DISL) bottom longline survey, 2006-2011. Shown are nominal CPUE (catch/100 hooks/hour), total sets per year, proportion positive sets (PPS), standardized delta-glm index, and standardized delta-glm index.	28
4	Abundance index statistics for Atlantic sharpnose shark sampled during the Dauphin Island Sea Lab (DISL) bottom longline survey, 2006-2011. Shown are nominal CPUE (catch/100 hooks/hour), total sets per year, proportion positive sets (PPS), standardized delta-brt index, and standardized delta-glm index.	28

List of Figures

1	Atlantic sharpnose shark nominal catch-per unit-effort (CPUE) from 2006 to 2011.	5
2	Histogram of Atlantic Sharpnose Sharks CPUE from 2006 to 2011. .	6
3	Size frequency of 1196 captured Atlantic sharpnose shark from 2006 to 2011.	11
4	Map and plot of residuals from binomial portion of fitted BRT model.	12
5	A receiver operating characteristic (ROC) curve illustrating the performance of the BRT sub-model. Model performance can be estimated using the area under the curve (AUC) where AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming ’positive’ ranks higher than ’negative’). Model values approaching 1 indicate superior model performance.	13
6	Mean contributions (percent) of variables predicting presence or absence Atlantic sharpnose shark from fitted BRT model.	14

7	Map and plot of residuals from log-normal portion of fitted BRT model.	15
8	Diagnostic QQ plot used to assess model fit. Inspection of model residuals suggest an imperfect but acceptable fit to these data. . . .	16
9	Mean contributions (percent) of variables predicting log-normal abundance of Atlantic sharpnose shark from fitted BRT model in samples where at least one shark was captured.	17
10	Spatially-explicit delta-index fitted values at each sampling location using a boosted regression tree (BRT) model.	18
11	Comparison of annual (raw) and standardized (delta) CPUE estimates for Atlantic sharpnose shark using boosted regression trees (BRT).	19
12	Map and plot of residuals from binomial portion of fitted BRT model.	22
13	A receiver operating characteristic (ROC) curve illustrating the performance of the GLM sub-model. Model performance can be estimated using the area under the curve (AUC) where AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). Model values approaching 1 indicate superior model performance.	23
14	Map and plot of residuals from log-normal portion of fitted GLM model.	24
15	Diagnostic QQ plot used to assess model fit. Inspection of model residuals suggest an imperfect but acceptable fit to these data. . . .	25
16	Comparison of annual (raw) and standardized (delta) CPUE estimates for Atlantic sharpnose shark using generalized linear models (GLMs).	26
17	Comparison of mean annual delta index values using generalized linear models (GLMs) and boosted regression trees (BRTs).	27

1 Introduction

In May 2006, the Dauphin Island Sea Lab (DISL), in conjunction with the National Marine Fisheries Service Mississippi Labs (NMFS MS Labs), initiated a monthly nearshore longline survey in Alabama coastal waters. This survey was initially designed to complement the annual NMFS bottom longline survey (SEFSC BLL), while sampling throughout the year and in waters inaccessible to large NMFS vessels. Atlantic sharpnose sharks *Rhizoprionodon terraenovae* are the most common component of the catch. Between May 2006 and December 2011, 1,196 Atlantic sharpnose sharks were captured during 446 bottom longline sets. Nominal and standardized catch per unit effort (CPUE, sharks/100 hooks/hour) and length frequency distributions are presented below.

2 Materials and Methods

2.1 Field Collections

Fisheries independent data were collected from 2006 to 2011 off the Alabama coast (Figure 1). At each station, a single bottom-longline was set and soaked for one hour. The main line consisted of 1.85 km (1 nm) of 4 mm monofilament (545 kg test) sampled with 100 gangions. Each gangion was made of 3.66 m of 3 mm (320 kg test) monofilament. Gangions consisted of a longline snap and a 15/0 circle hook, baited with Atlantic mackerel *Scomber scombrus*. The longline was anchored to the bottom with weights at the start, middle and end of the mainline, and identified with buoys at each end. All sharks that could be safely boated were removed from the mainline, unhooked and identified to species. Biotic variables collected included sex, length (precaudal, fork, natural and stretch total), weight and maturity (when possible). Maturity in males was assessed following Clark and Von Schmidt (1965). Sharks were tagged in the primary dorsal fin with a plastic rototag. Abiotic variables collected included depth as well as surface and bottom values for temperature, salinity and dissolved oxygen using a Seabird SBE911 plus, or an SBE 25 CTD (2006-2009), and recently with a Hydrolab MS5 multiprobe.

2.2 Modeling approach

Standardized catch per unit effort of Atlantic sharpnose sharks was calculated to examine changes in catch-per-unit-effort (CPUE) over time (Maunder and Punt 2004). Preliminary analyses indicated that Atlantic Sharpnose CPUE were "zero-inflated" with a long tail (i.e., few samples with larger than expected catch) which is typical of many fishery independent data sets (Figure 2).

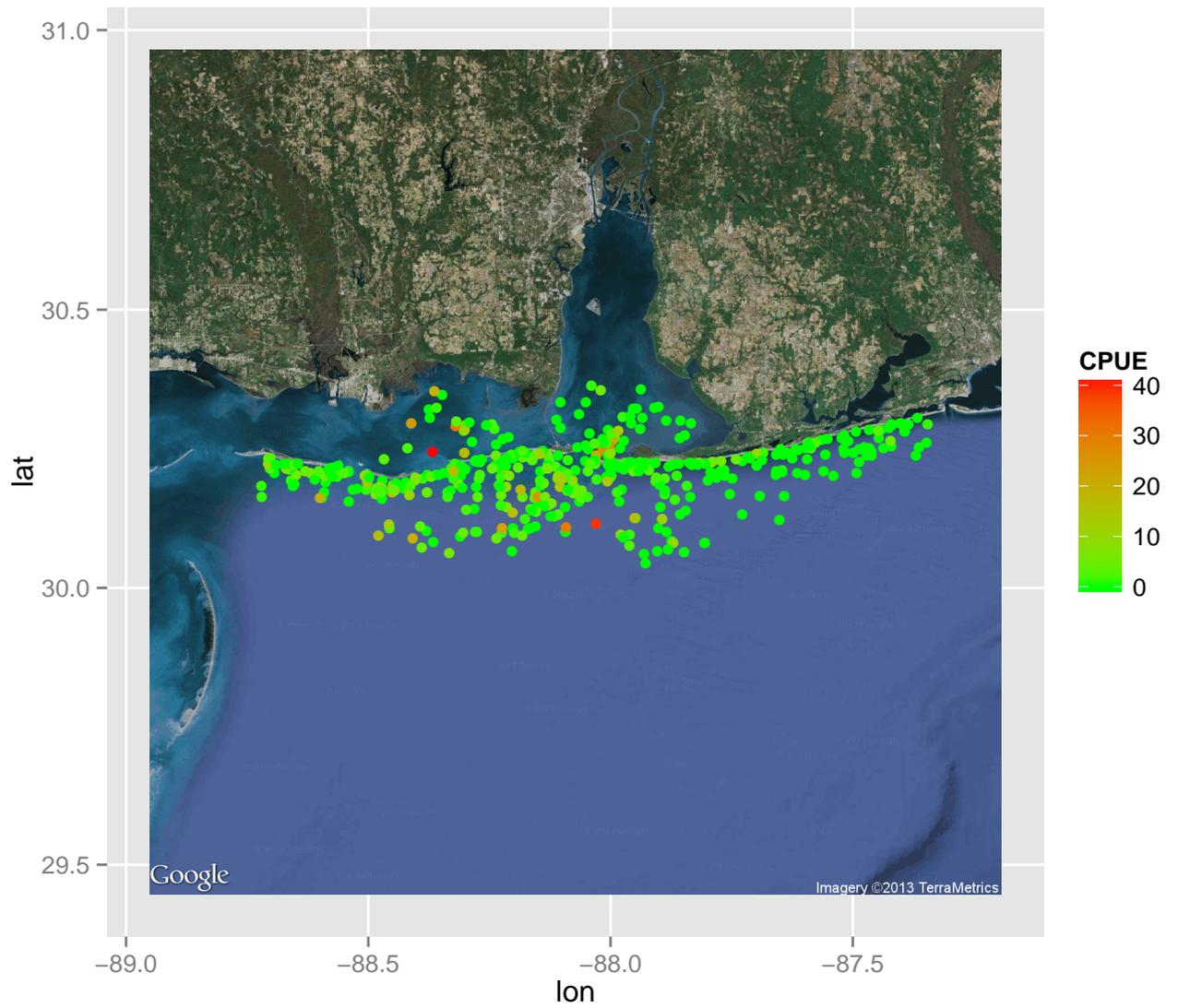


Figure 1: Atlantic sharpnose shark nominal catch-per unit-effort (CPUE) from 2006 to 2011.

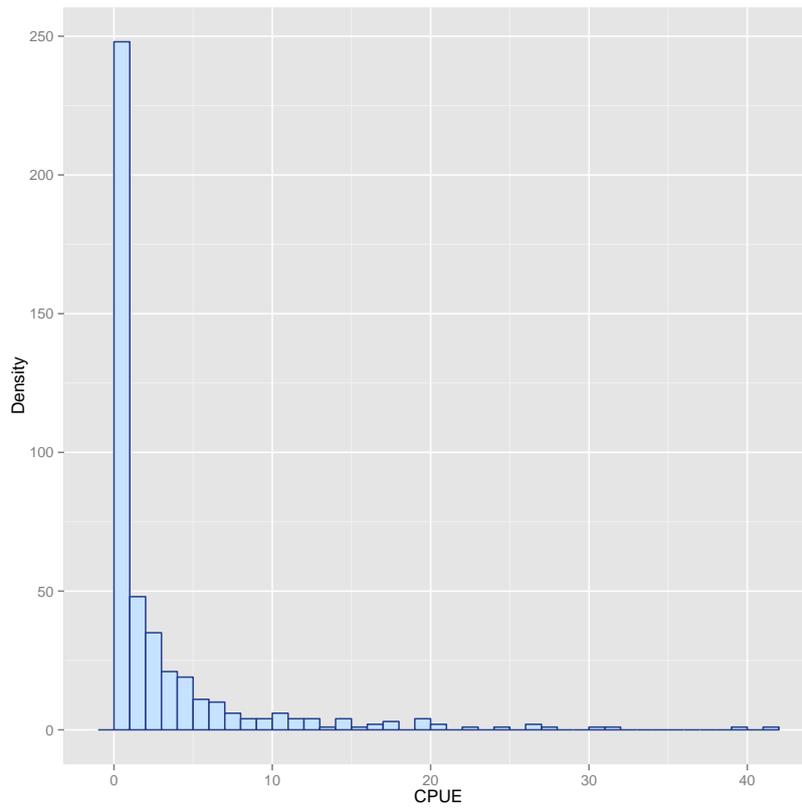


Figure 2: Histogram of Atlantic Sharpnose Sharks CPUE from 2006 to 2011.

A delta-lognormal approach (Lo et al. 1992) was used to develop a standardized index of abundance. In general, this approach models separately the proportion of positive sets (PPS, sets that captured Atlantic sharpnose sharks/total sets made) and the catch rates on positive sets and combines these indices to construct a single standardized CPUE index (Cass-Calay and Schmidt 2009).

Specifically,

$$l_y = c_y p_y \quad (1)$$

where c_y is the estimate of mean CPUE for positive catches only for year y , and p_y is the estimate of mean probability of occurrence during year y . Both c_y and p_y were estimated using generalized linear models and boosted regression trees. Two modeling approaches were used to compare the performance of a relatively new technique, *boosted regression trees (BRT)* Elith et al. (2008) with the current standard practice delta approach using generalized linear models. For both modeling approaches, data used to estimate abundance for positive catches c and probability of occurrence p were assumed to have a lognormal distribution and a binomial distribution, respectively, and modeled using the following equations:

$$\ln(c) = \beta X + \varepsilon \quad (2)$$

and

$$p = \frac{e^{X\beta + \varepsilon}}{1 + e^{X\beta + \varepsilon}} \quad (3)$$

respectively, where c is a vector of the positive catch data, p is a vector of the presence/absence data, X is the design matrix for main effects, β is the parameter vector for main effects, and ε is a vector of independent normally distributed errors with expectation zero and variance σ^2 (Ingram et al 2010).

For both modeling approaches, a standardized delta index for each year was calculated as the annual mean delta value for all samples within a calendar year. Confidence limits for each approach was estimated using non-parametric bootstrapping with replacement (n=200) (Efron and Tibshirani 1993). Analyses were carried out in R (version x64 3.0.0, R Development Core Team, 2013). The boosted regression tree models were fit using the "gbm" library supplemented with functions from Elith et al. (2008) and J.T. Froeschke. The initial GLM sub-models were fit using the exhaustive search algorithm implemented in the "bestglm" library in R.

2.2.1 Boosted regression trees

Boosted regression trees use a model averaging (ensemble) method that allows for both explanation and prediction (Elith et al. 2008). However, BRTs have only recently been applied to ecological questions (Friedman 2001, Leathwick et

al. 2006, Elith et al. 2008, Froeschke et al. 2010, Froeschke and Froeschke 2011). Given the ability of BRTs to model interactions and automatically select important variables, as well as robustness to outliers and missing data, BRT models are growing in popularity. Each individual model consists of a simple regression tree based on a series of binary splits constructed from the predictor variables (Hastie et al. 2001), accommodates continuous or categorical predictors, missing values, and is not affected by transformation or outliers. This technique can also fit complex non-linear relationships and often has superior predictive performance to other techniques such as generalized linear and additive models that are often used to develop standardized indices of abundance (Lo et al. 1992). Relative importance of explanatory variables can be estimated by averaging the number of times a variable is selected for splitting and the squared improvement resulting from these splits (Friedman 2001, Friedman and Muelman 2003). Values are scaled to 100 and, higher numbers indicate a stronger influence on the response variable. The ability to model interactions is controlled by a tree complexity (tc) parameter where the value specifies the number of nodes on each tree and subsequently the ability to model interactions (Leathwick et al. 2006).

2.2.2 BRT Model fitting and variable selection

A delta-model was developed using BRT (hereafter: delta-brt). The objective was to estimate the probability of catching one or more Atlantic Sharpnose Sharks in a sample as well as the expected number in samples where one or more animals were caught. To estimate the probability of occurrence, a binomial BRT was fit to these data using a tree complexity of 5 with a learning rate = 0.001 to minimize residual deviance without overfitting. The positive samples were log transformed and a BRT was fit to these data using a tree complexity of 5 with a learning rate = 0.001 with a Gaussian error distribution. A delta-value for each positive sample was computed as the product of the probability of occurrence times un-logged CPUE value from the positive BRT model using functions from J.T. Froeschke.

After the full model was fit, a model simplification routine ("gbm.simplify") was used that is similar to a backward selection routine in regression (Elith et al. 2008). Variables were removed until significant increases in residual deviance were found between the full and reduced models. This routine was employed for both the binomial and log-normal models and permits different explanatory variables to be included in each sub-model.

2.2.3 GLM Model fitting and variable selection

As a basis for comparison, a standardized index of abundance was also developed using generalized linear models following the protocol and Lo et al (1992). This

method combines two generalized linear models; an analysis of the probability of capture and a second analysis of the number of individuals captured in positive samples. The delta-glm is the product of fitted values from the binomial and negative binomial model for each observation. Model selection for each sub-model was selected from using an exhaustive search algorithm that explores all possible combinations of main-effects models and candidate models were selected based on lowest Akaike's information criterion (AIC) values. Using this approach, the model with the most support may not include "Year" as a predictor variable, although this variable is typically retained in the delta-glm modeling approach. Therefore, the model with the lowest AIC value including Year as a covariate was selected as the "best" model and does not require that same model formulation for each submodel.

2.3 Model validation

Model validation process was identical for both the delta-brt and delta-glm models. Residuals were extracted and geo-referenced to identify spatial patterns (i.e., locations or regions where model fits may be poor) and plotted in histograms to examine overall model fits. This approach was conducted separately for both sub-models (binomial and log-normal).

2.3.1 Binomial models residuals

For the binomial sub-models a receiver-operator-characteristic curve (ROC) was fit as an indicator of model performance. Values for ROC estimate the degree to which fitted values discriminate between observed presences and absences and can be interpreted as the probability that a presence for a species drawn at random will have a higher fitted probability than an absence drawn at random (Parisen and Moritz 2009). The area under the ROC curve can be integrated and interpreted as an Area Under the Curve (AUC) value that has a range from 0.5 to 1. Using this metric, a value of one indicates perfect discrimination of probabilities between presence and absence samples and a value of 0.5 indicates that model discrimination is no better than random. While models with AUC values greater than 0.6 are considered useful (Parisien and Moritz 2009), values greater than 0.8 are considered very good, and above greater than 0.9 excellent (Lane et al. 2009).

2.3.2 Model validation: Log-normal model

For the log-normal model, residuals were extracted and plotted in the same manner as the binomial model, however, only positive catches were included. A QQ-plot was also used to examine model fits.

3 Results

3.1 Length

Size of Atlantic sharpnose sharks sampled ranged from 483 to 963 mm FL from 2006 to 2011 (Figure 3).

3.2 BRT

An initial full model was fit where CPUE was modeled as a function of year sampled, month, longitude, sample block, and depth. Model selection of the binomial sub-model indicated that all variables should be retained and inspection of residuals indicated an acceptable model fit (Figure 4). Evaluation of the binomial sub-model suggested good predictive performance $AUC = 0.89$ (Figure 5). One benefit of the BRT approach is the ability to estimate the relative influence of predictor variables on the response. For the binomial sub-model, longitude of sampling location was the most important predictor of occurrence as frequency of occurrence increased from east to west (Figure 6).

Model selection for log-normal BRT sub-model also indicated that all predictors considered should be retained in the model and inspection of residuals indicated an acceptable model fit (Figure 7). To provide further detail about the model fit, a QQplot was also produced (Figure 8). Examination of the relative influence of predictors suggest that the spatial variables were the most important predictors of abundance (Figure 9).

3.2.1 delta-brt index

The delta-brt model (combining binomial and negative binomial models) is the product of fitted values from the binomial and negative binomial model for each observation (Figure 10). Annual fits were computed as the mean of delta fits for each sample year and compared to raw estimates of CPUE (Figure 11). Comparatively, the delta index has less annual variability than the raw CPUE and the estimates and the abrupt increase in 2010 are moderated (Table 3).

3.3 GLM

An initial full model was fit where CPUE was also modeled as a function of year sampled, month, longitude, sample block, and depth. Model selection of the binomial sub-model indicated that Year, Month, and Block variables should be retained (Table 1) and inspection of residuals indicated an acceptable model fit (Figure 12). Evaluation of the binomial sub-model suggested reasonable predictive performance

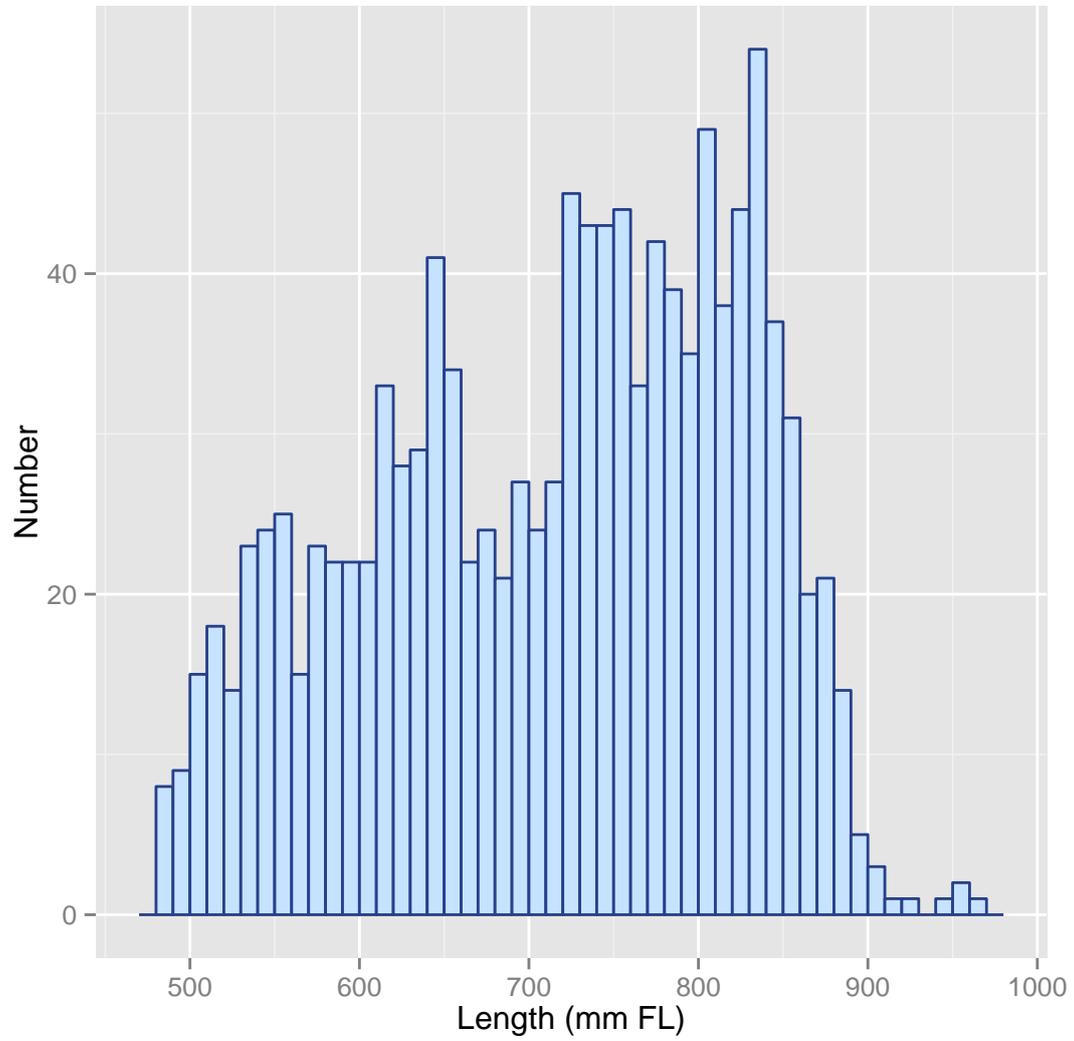


Figure 3: Size frequency of 1196 captured Atlantic sharpnose shark from 2006 to 2011.

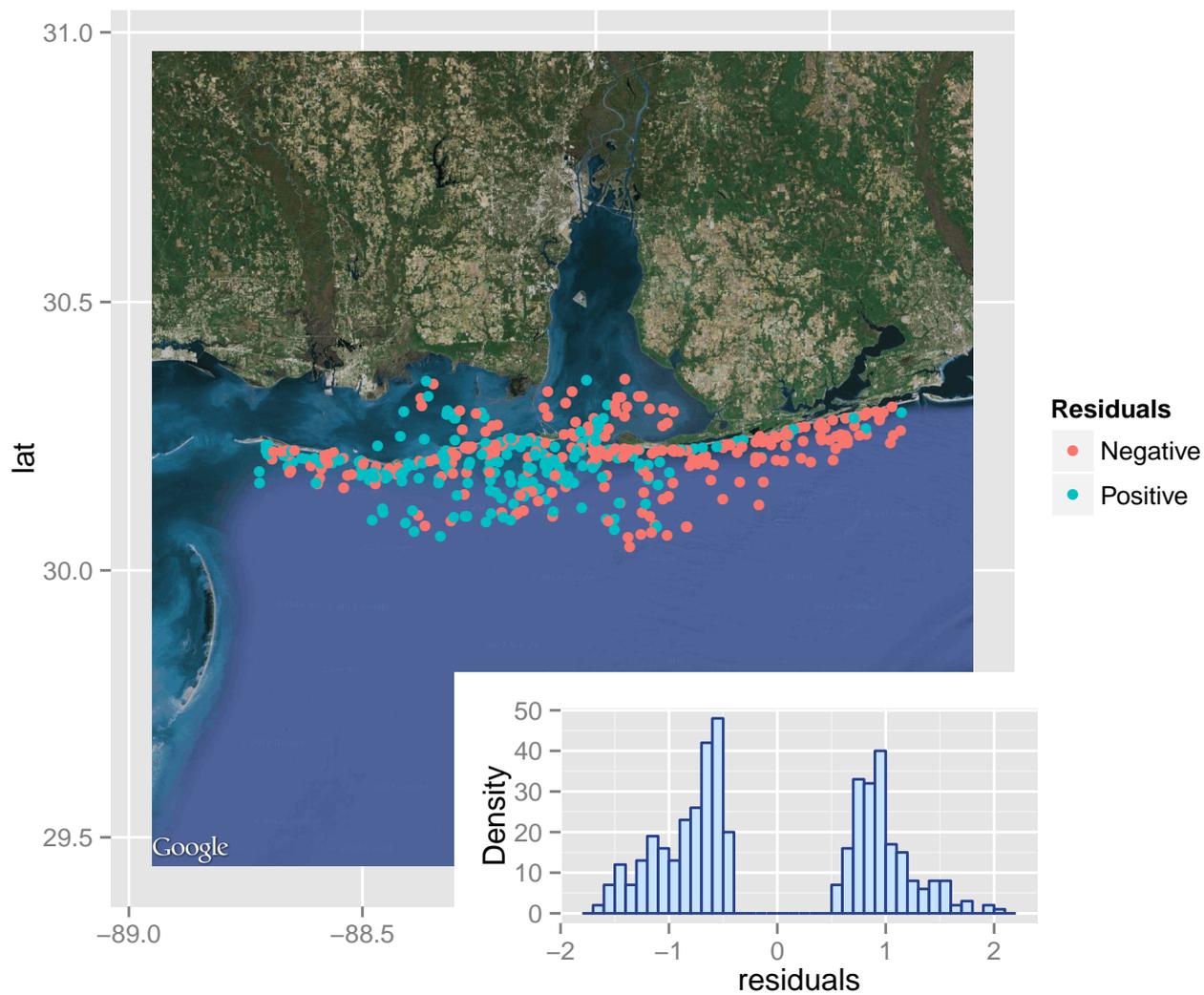


Figure 4: Map and plot of residuals from binomial portion of fitted BRT model.

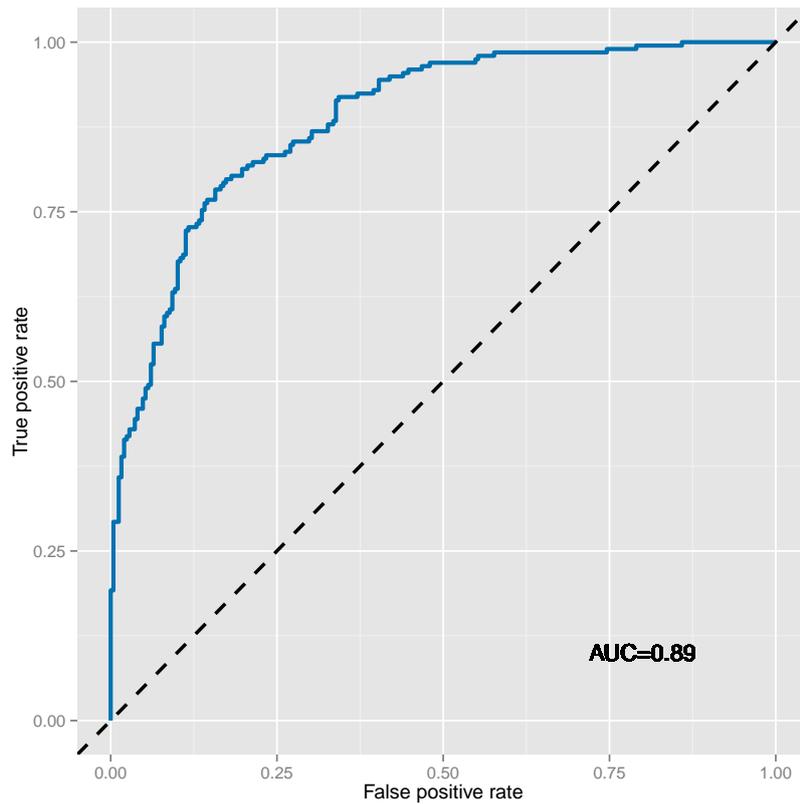


Figure 5: A receiver operating characteristic (ROC) curve illustrating the performance of the BRT sub-model. Model performance can be estimated using the area under the curve (AUC) where AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). Model values approaching 1 indicate superior model performance.

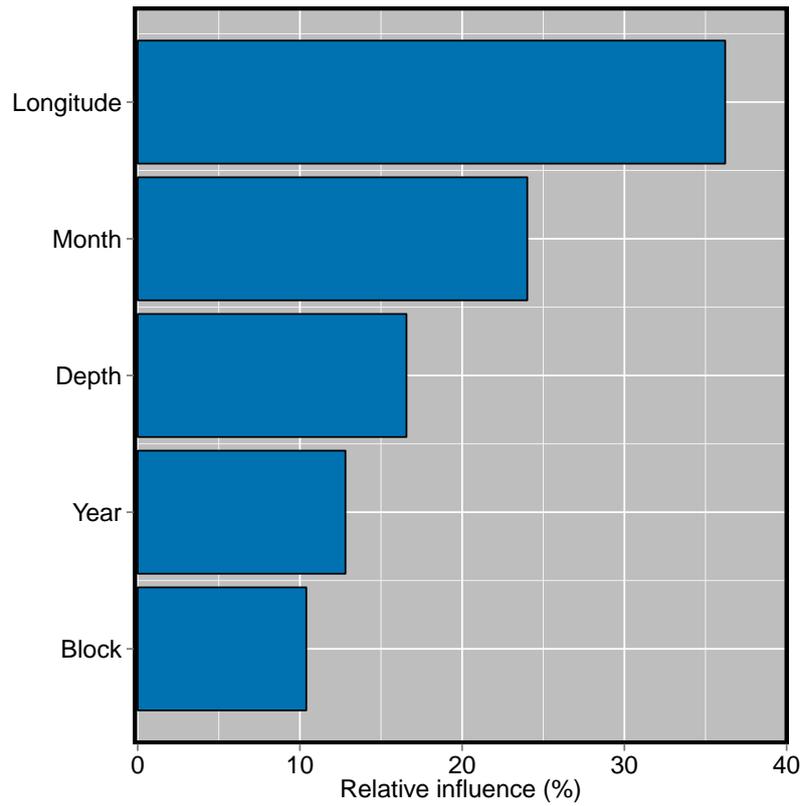


Figure 6: Mean contributions (percent) of variables predicting presence or absence Atlantic sharpnose shark from fitted BRT model.

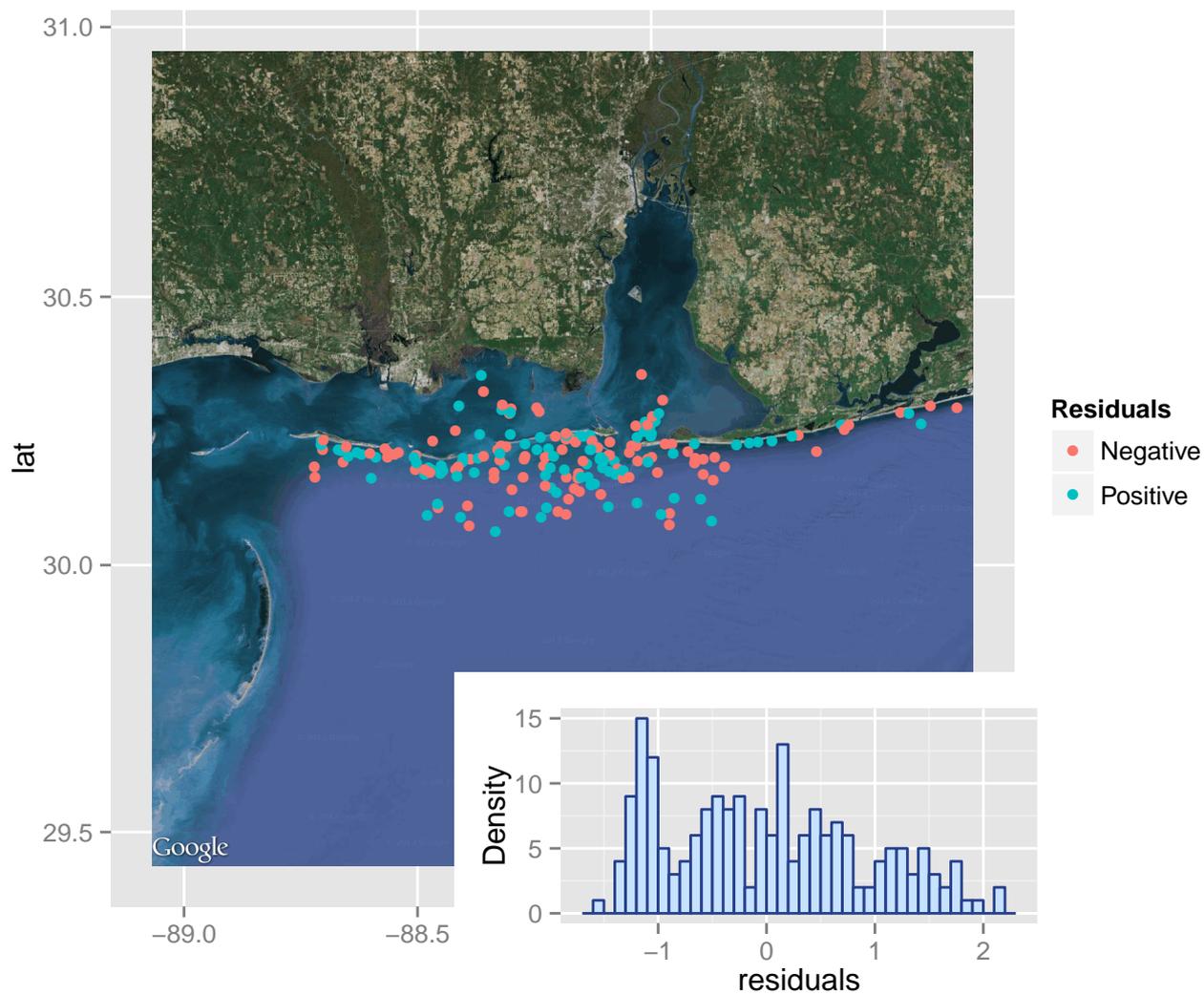


Figure 7: Map and plot of residuals from log-normal portion of fitted BRT model.

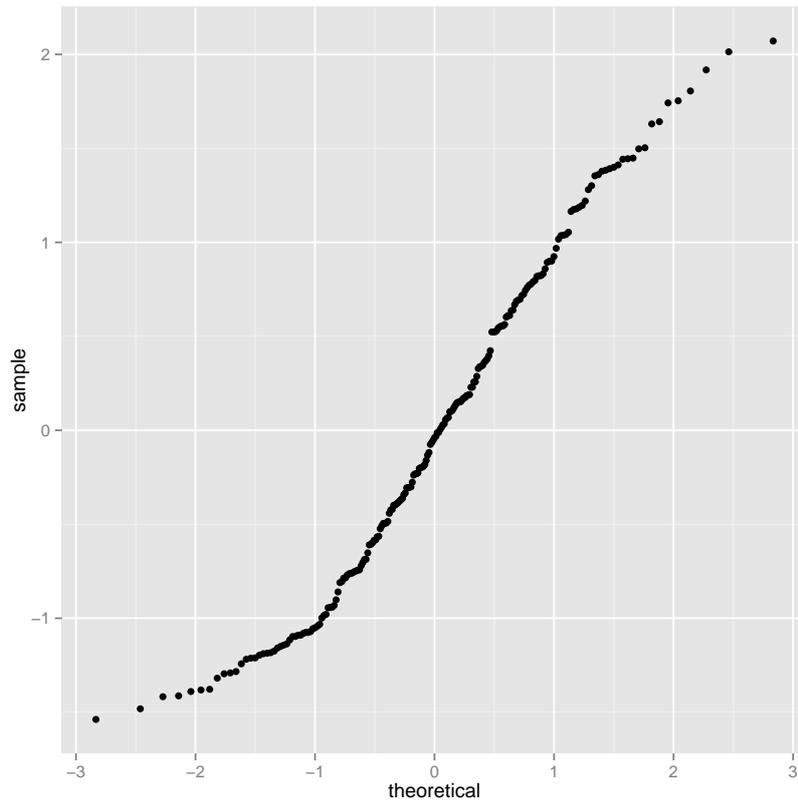


Figure 8: Diagnostic QQ plot used to assess model fit. Inspection of model residuals suggest an imperfect but acceptable fit to these data.

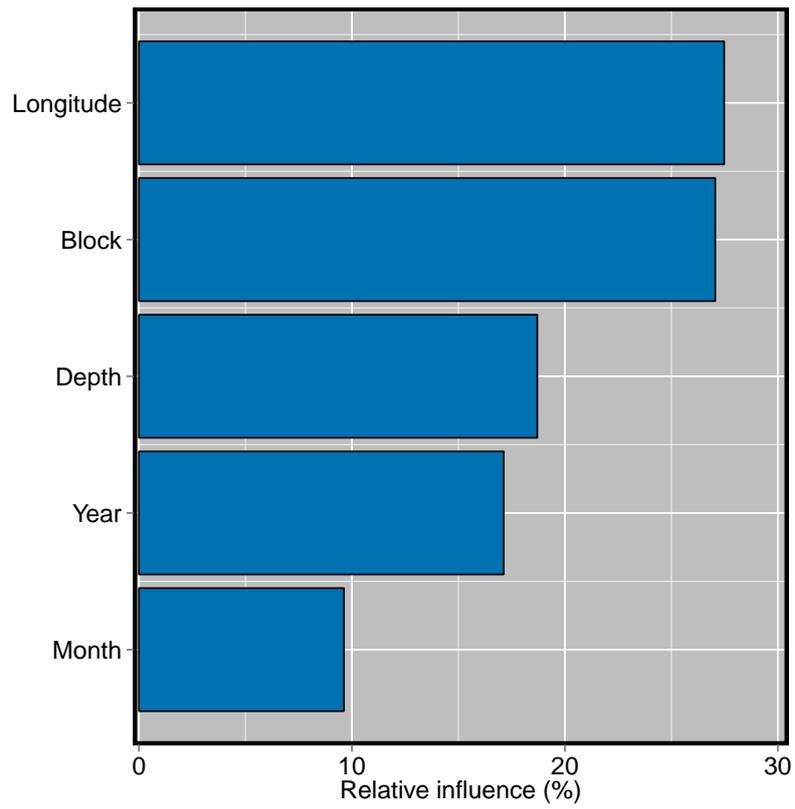


Figure 9: Mean contributions (percent) of variables predicting log-normal abundance of Atlantic sharpnose shark from fitted BRT model in samples where at least one shark was captured.

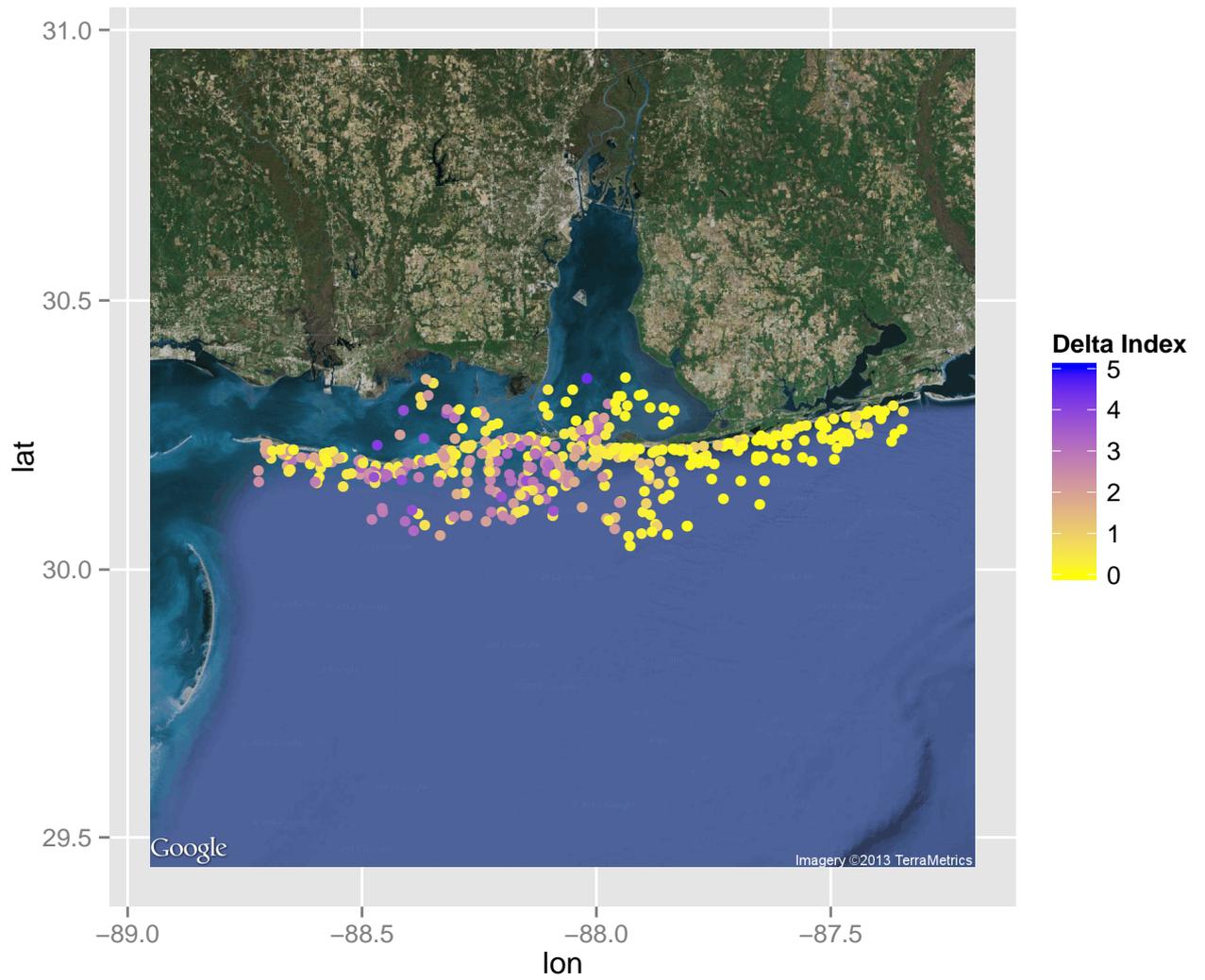


Figure 10: Spatially-explicit delta-index fitted values at each sampling location using a boosted regression tree (BRT) model.

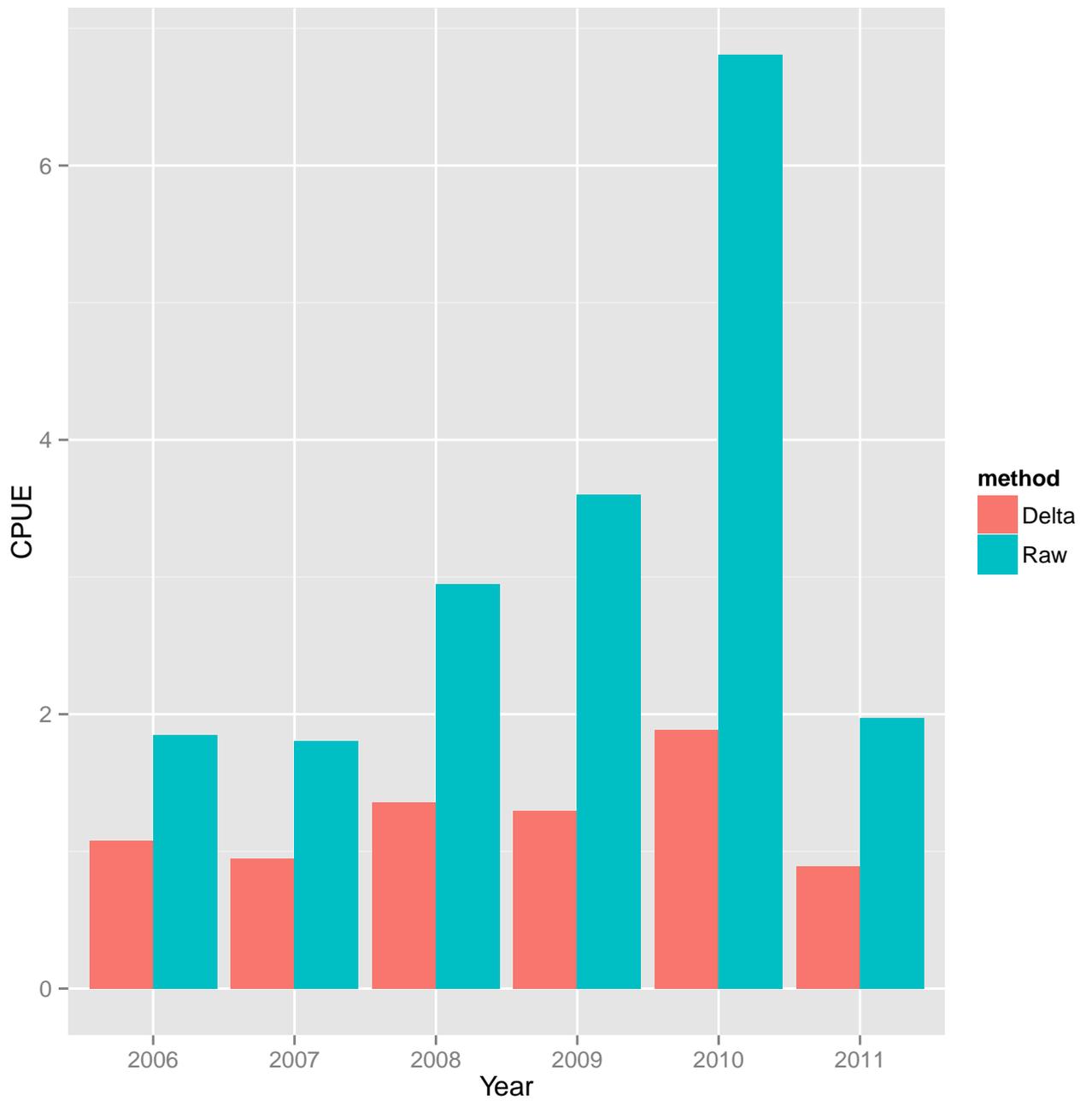


Figure 11: Comparison of annual (raw) and standardized (delta) CPUE estimates for Atlantic sharpnose shark using boosted regression trees (BRT).

AUC = 0.73 (Figure 13). Model selection for log-normal GLM sub-model also indicated that Longitude, Year, Block, and Depth predictors should be retained in the model (Table 2) and inspection of residuals indicated an acceptable model fit (Figure 14). To provide further detail about the model fit, a QQplot was also produced (Figure 15).

3.3.1 delta-glm index

The delta-glm model (combining binomial and negative binomial models) is the product of fitted values from the binomial and negative binomial model for each observation (Figure 10). Annual fits were computed as the mean of delta fits for each sample year and compared to raw estimates of CPUE (Figure 16). Comparitively, the delta index has less annual variability than the raw CPUE and the estimates and the abrupt increase in 2010 are moderated, although less than with the delta-brt model (Table 4).

4 Discussion

This report summarizes a time series of catch per unit effort (CPUE) for Atlantic Sharpnose Shark off the coast of Alabama, USA, from 2006 to 2011. Standardized indices of abundance were computed to account for changes in CPUE that are assumed independent of population size (Froeschke et al. 2013). Two modeling approaches were used: generalized linear models (GLMs) and boosted regression trees (BRTs). Whereas generalized linear models are the most common method for obtaining standardized indices of abundance, boosted regression is a comparable, yet rarely employed, technique. This report examines both approaches in terms of model fit and results. Both methods of standardization resulted in a relatively stable CPUE across years, with the exception of a notable increase in 2010. Standardizing the indices moderated the increase in 2010, particularly in the index constructed with boosted regression. This suggests that both sampling and biological effects may have influenced the elevated catch rates in 2010. In terms of model fit, the BRT outperformed the GLM approach for both the binomial and log-normal sub-models, indicating that the increased model flexibility of this approach may be useful in the development of future CPUE indices. That said, it is difficult to evaluate model accuracy per se as the true population trend over time is unknown. This is something that could be further explored as a simulation exercise.

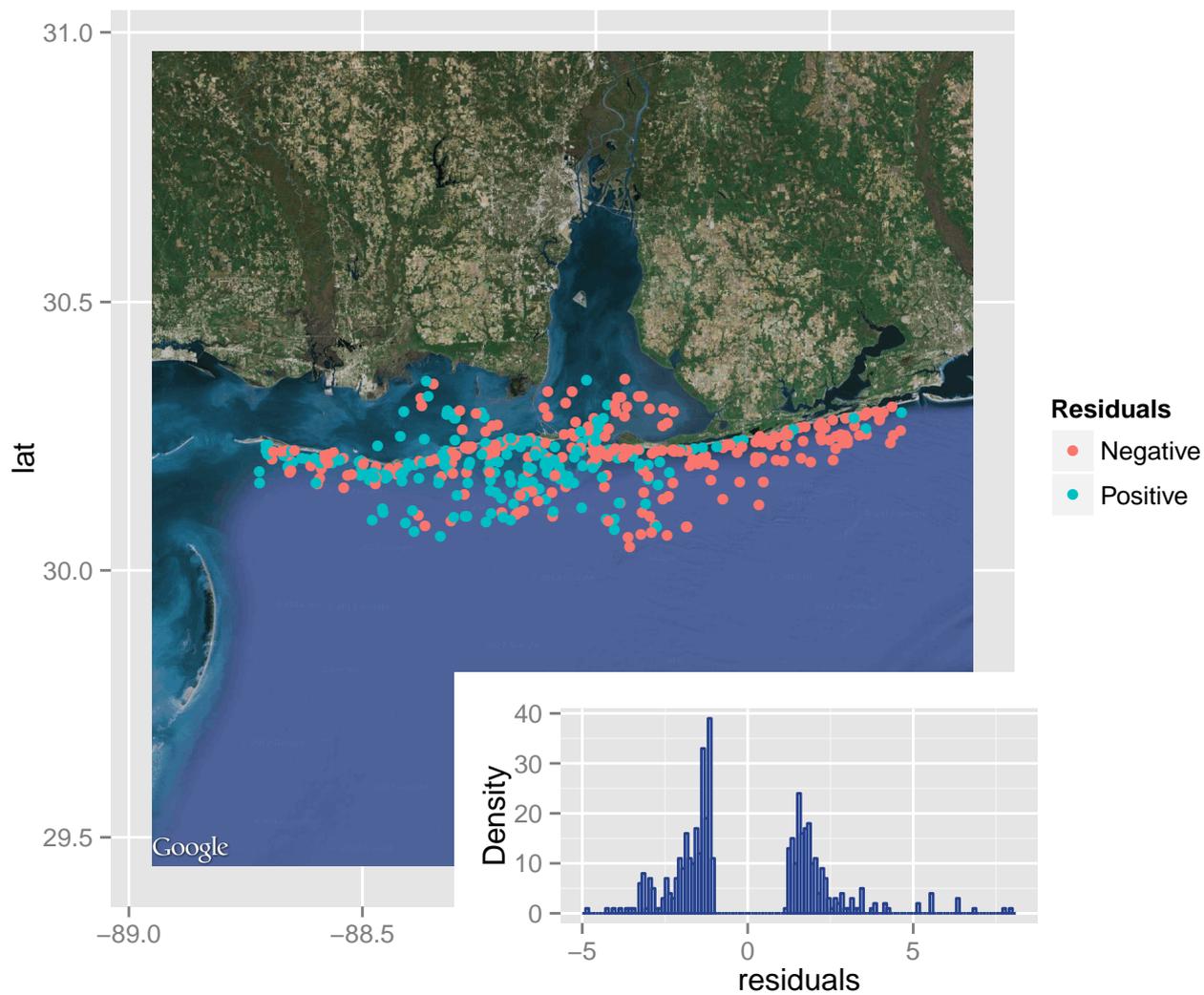


Figure 12: Map and plot of residuals from binomial portion of fitted BRT model.

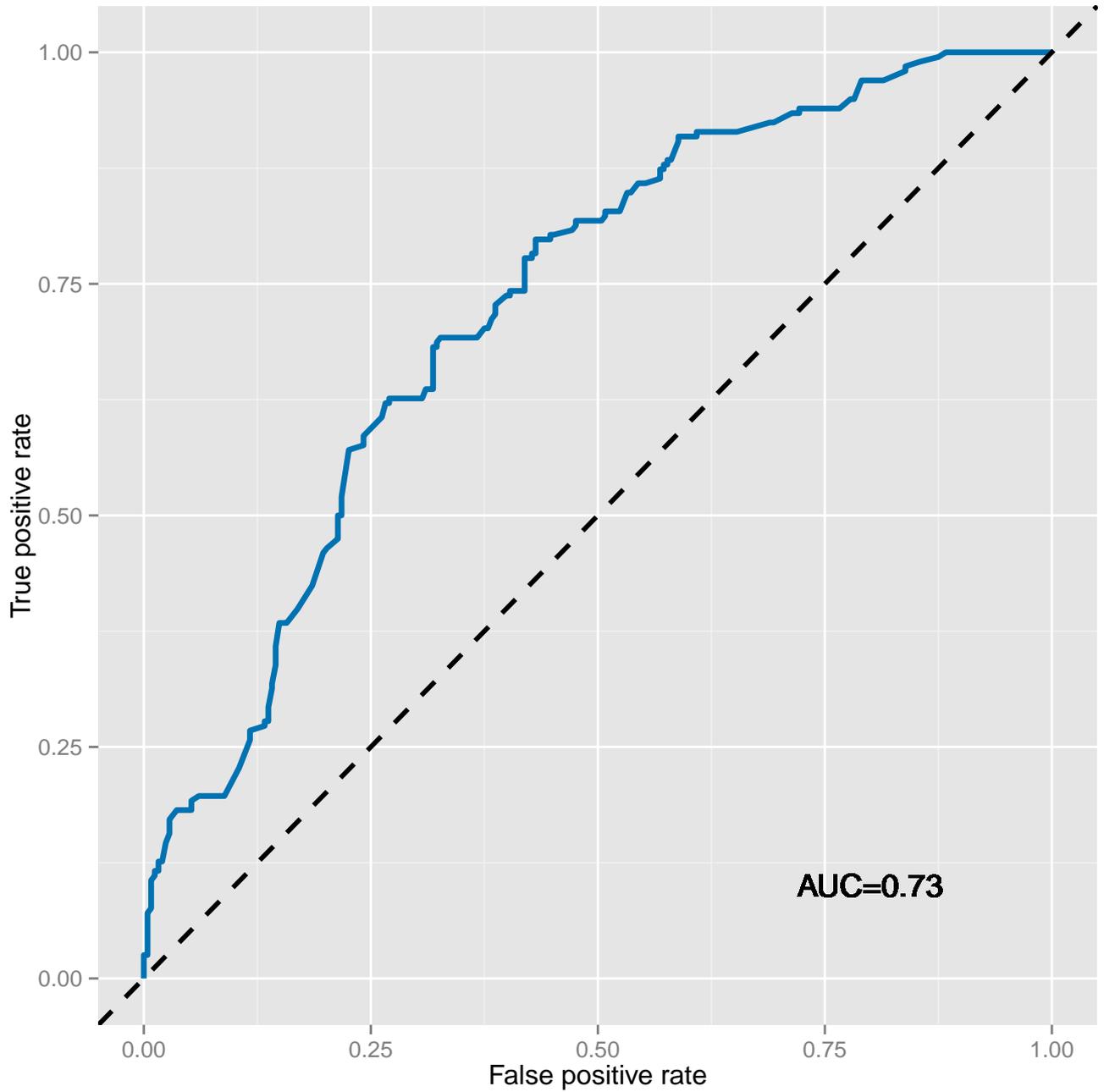


Figure 13: A receiver operating characteristic (ROC) curve illustrating the performance of the GLM sub-model. Model performance can be estimated using the area under the curve (AUC) where AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). Model values approaching 1 indicate superior model performance.

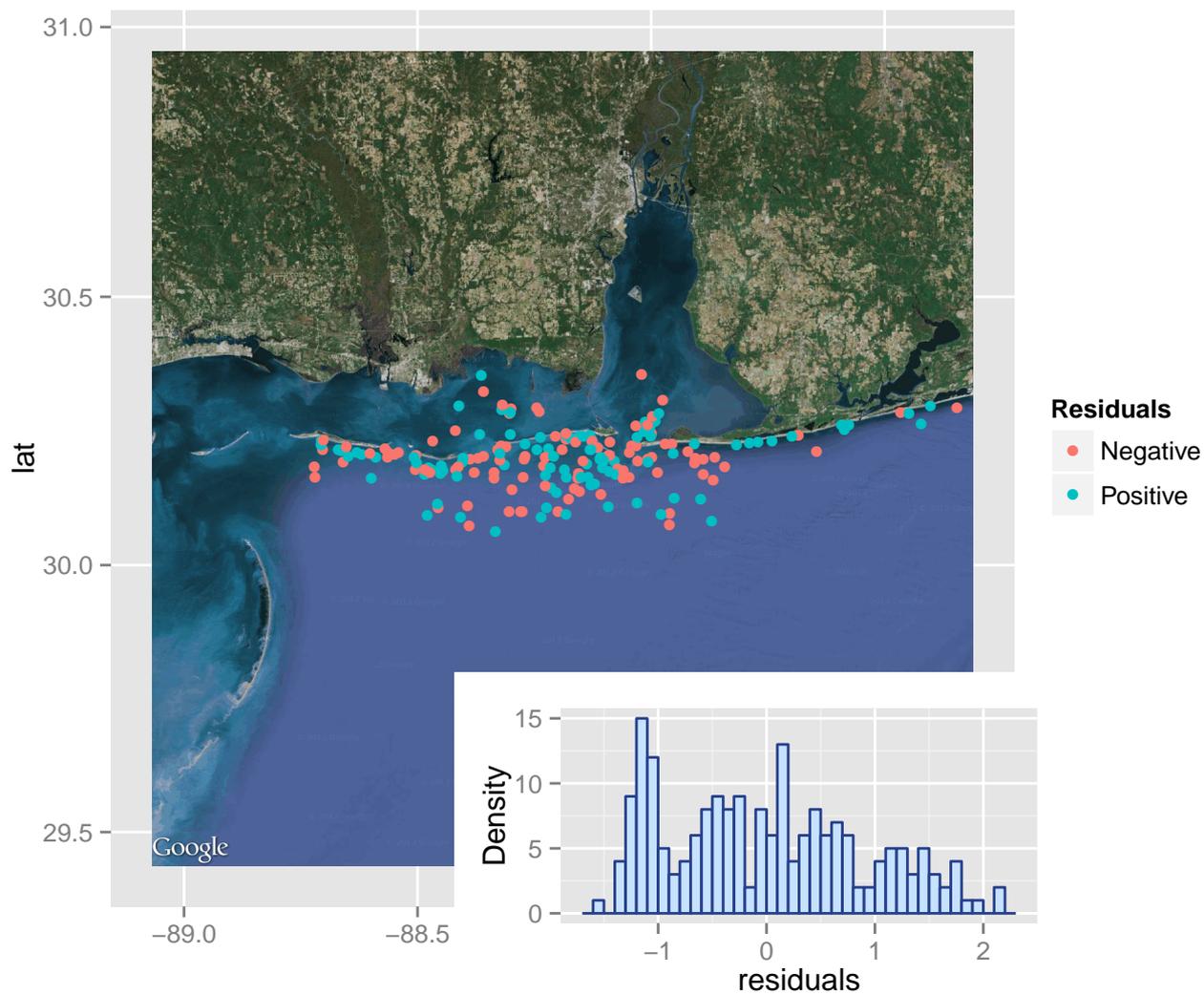


Figure 14: Map and plot of residuals from log-normal portion of fitted GLM model.

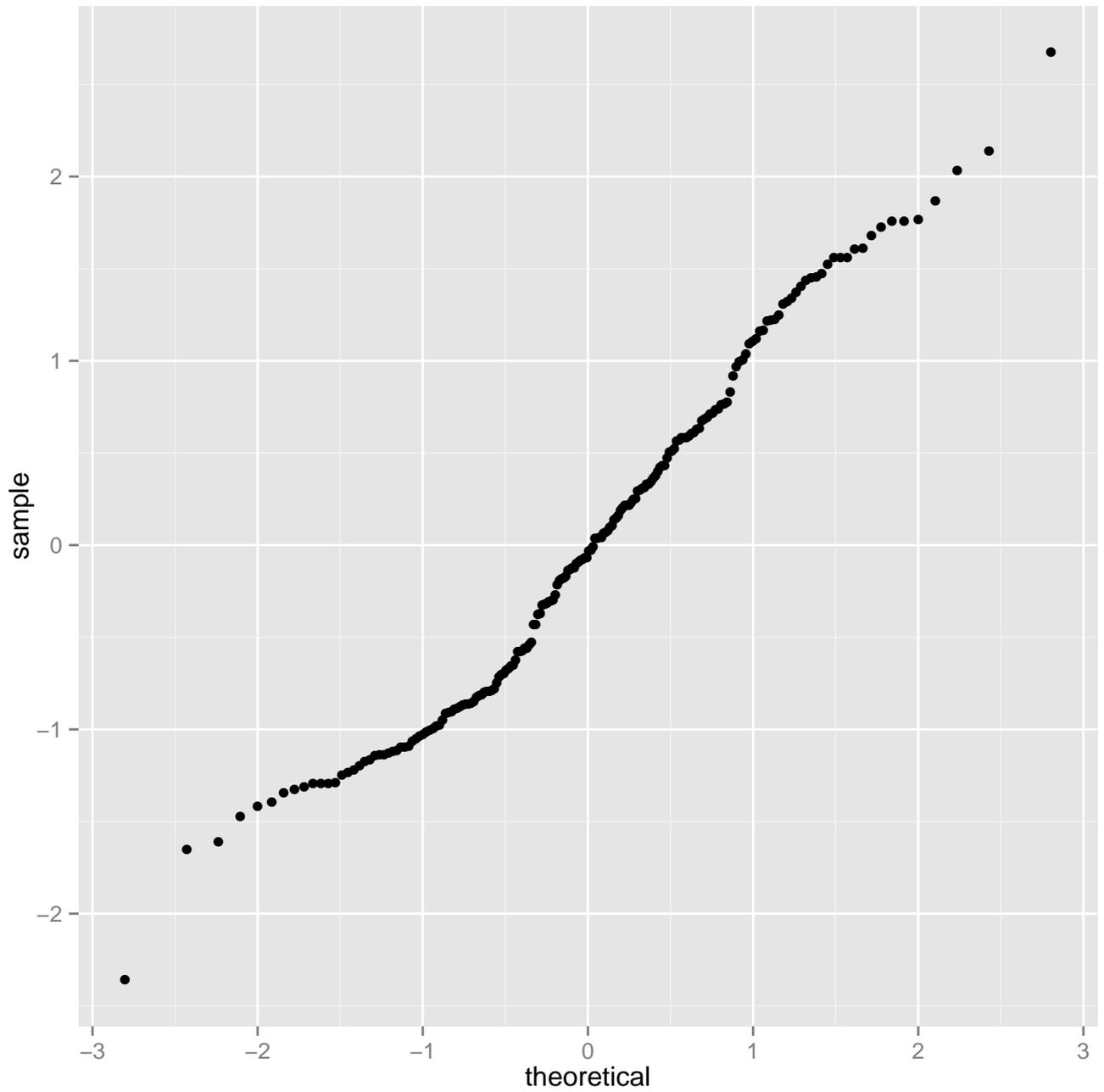


Figure 15: Diagnostic QQ plot used to assess model fit. Inspection of model residuals suggest an imperfect but acceptable fit to these data.

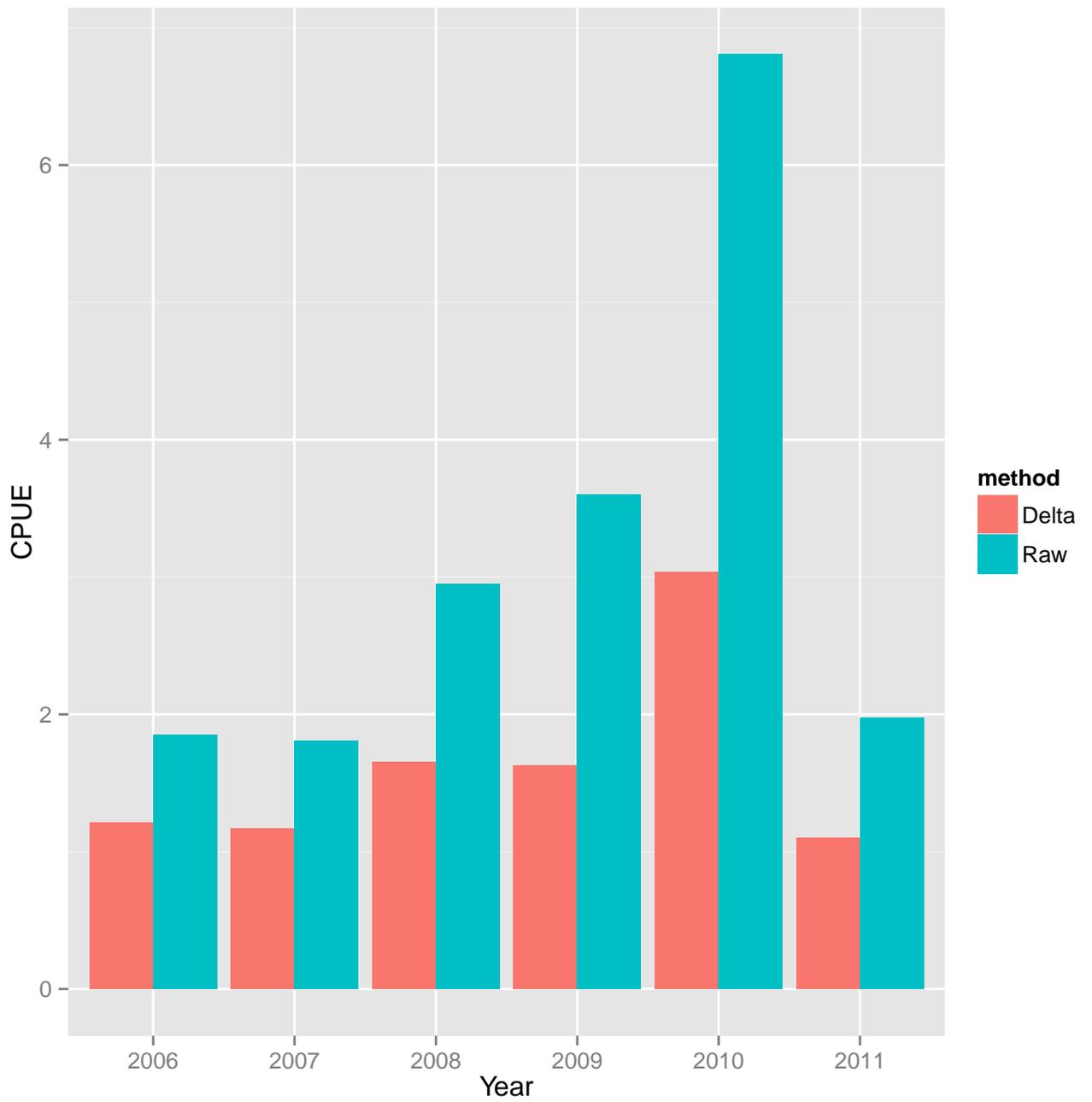


Figure 16: Comparison of annual (raw) and standardized (delta) CPUE estimates for Atlantic sharpnose shark using generalized linear models (GLMs).

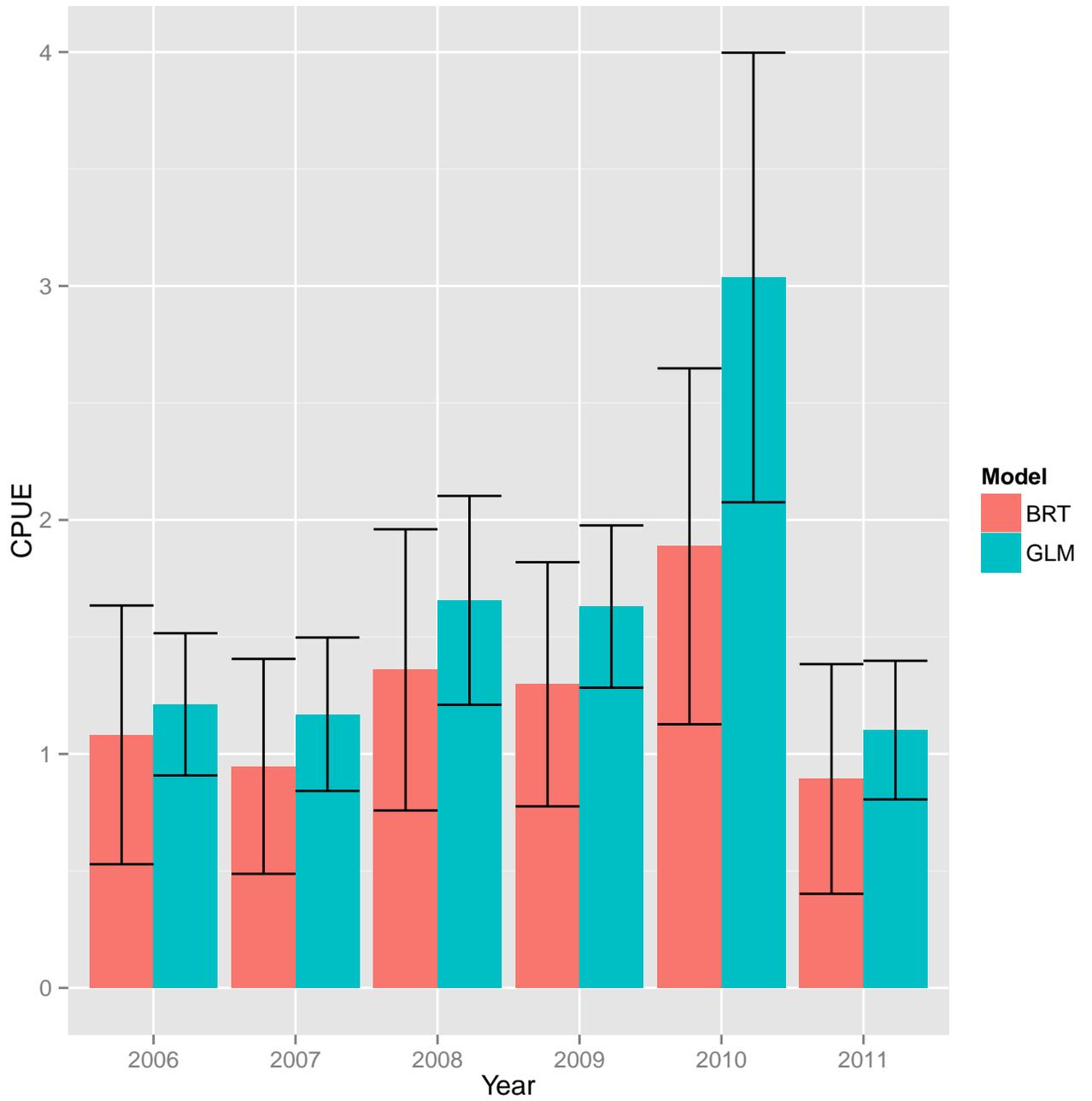


Figure 17: Comparison of mean annual delta index values using generalized linear models (GLMs) and boosted regression trees (BRTs).

	Longitude	Year	Month	Block	Depth	Criterion
1	FALSE	FALSE	TRUE	TRUE	FALSE	559.21
2	FALSE	FALSE	TRUE	TRUE	TRUE	559.62
3	TRUE	FALSE	TRUE	TRUE	FALSE	561.03
4	TRUE	FALSE	TRUE	TRUE	TRUE	561.48
5	FALSE	TRUE	TRUE	TRUE	FALSE	561.49
6	FALSE	TRUE	TRUE	TRUE	TRUE	561.69
7	TRUE	TRUE	TRUE	TRUE	FALSE	563.30
8	TRUE	FALSE	TRUE	FALSE	TRUE	563.43
9	TRUE	TRUE	TRUE	TRUE	TRUE	563.55
10	TRUE	FALSE	TRUE	FALSE	FALSE	563.59
11	TRUE	TRUE	TRUE	FALSE	TRUE	565.17
12	TRUE	TRUE	TRUE	FALSE	FALSE	565.63
13	FALSE	FALSE	FALSE	TRUE	FALSE	566.20
14	FALSE	FALSE	FALSE	TRUE	TRUE	566.30
15	TRUE	FALSE	FALSE	FALSE	TRUE	567.75
16	TRUE	FALSE	FALSE	TRUE	FALSE	567.83
17	TRUE	FALSE	FALSE	FALSE	FALSE	567.85
18	TRUE	FALSE	FALSE	TRUE	TRUE	568.00
19	FALSE	TRUE	FALSE	TRUE	FALSE	569.22
20	FALSE	TRUE	FALSE	TRUE	TRUE	569.62

Table 1: Summary of variables selected for binomial GLM model. Model selection used an exhaustive search algorithm and models were selected using Akaike’s information criterion (AIC). The 20 best models are illustrated and the model with the lowest AIC value that included year as a covariate was selected as ”best”.

	Longitude	Year	Month	Block	Depth	Criterion
1	TRUE	FALSE	FALSE	TRUE	TRUE	-3.56
2	FALSE	FALSE	FALSE	TRUE	TRUE	-2.94
3	TRUE	FALSE	TRUE	TRUE	TRUE	-1.57
4	FALSE	FALSE	TRUE	TRUE	TRUE	-0.96
5	FALSE	FALSE	FALSE	TRUE	FALSE	0.66
6	TRUE	FALSE	FALSE	TRUE	FALSE	1.22
7	FALSE	FALSE	FALSE	FALSE	FALSE	2.24
8	TRUE	TRUE	FALSE	TRUE	TRUE	2.55
9	TRUE	FALSE	FALSE	FALSE	FALSE	2.57
10	FALSE	FALSE	TRUE	TRUE	FALSE	2.64
11	FALSE	TRUE	FALSE	TRUE	TRUE	3.18
12	TRUE	FALSE	TRUE	TRUE	FALSE	3.21
13	TRUE	TRUE	FALSE	FALSE	FALSE	3.28
14	FALSE	FALSE	FALSE	FALSE	TRUE	3.39
15	FALSE	TRUE	FALSE	FALSE	FALSE	3.61
16	TRUE	FALSE	FALSE	FALSE	TRUE	3.94
17	FALSE	FALSE	TRUE	FALSE	FALSE	4.24
18	FALSE	TRUE	FALSE	TRUE	FALSE	4.31
19	TRUE	TRUE	TRUE	TRUE	TRUE	4.50
20	TRUE	FALSE	TRUE	FALSE	FALSE	4.53

Table 2: Summary of variables selected for log-normal GLM model. Model selection used an exhaustive search algorithm and models were selected using Akaike’s information criterion (AIC). The 20 best models are illustrated and the model with the lowest AIC value that included year as a covariate was selected as ”best”

	Year	Nominal	Sets	PPS	BRT	LCI	UCI
1	2006	1.85	93.00	0.42	1.08	0.53	1.63
2	2007	1.81	124.00	0.40	0.95	0.49	1.41
3	2008	2.95	115.00	0.50	1.36	0.76	1.96
4	2009	3.60	45.00	0.49	1.30	0.78	1.82
5	2010	6.81	31.00	0.58	1.89	1.13	2.65
6	2011	1.97	38.00	0.34	0.89	0.40	1.38

Table 3: Abundance index statistics for Atlantic sharpnose shark sampled during the Dauphin Island Sea Lab (DISL) bottom longline survey, 2006-2011. Shown are nominal CPUE (catch/100 hooks/hour), total sets per year, proportion positive sets (PPS), standardized delta-glm index, and standardized delta-glm index.

	Year	Nominal	Sets	PPS	BRT	LCI	UCI
1	2006	1.85	93.00	0.42	1.21	0.91	1.52
2	2007	1.81	124.00	0.40	1.17	0.84	1.50
3	2008	2.95	115.00	0.50	1.66	1.21	2.10
4	2009	3.60	45.00	0.49	1.63	1.28	1.98
5	2010	6.81	31.00	0.58	3.04	2.08	4.00
6	2011	1.97	38.00	0.34	1.10	0.81	1.40

Table 4: Abundance index statistics for Atlantic sharpnose shark sampled during the Dauphin Island Sea Lab (DISL) bottom longline survey, 2006-2011. Shown are nominal CPUE (catch/100 hooks/hour), total sets per year, proportion positive sets (PPS), standardized delta-brt index, and standardized delta-glm index.

References

- [1] Cass-Calay SL, Schmidt TW (2009) Monitoring changes in the catch rates and abundance of juvenile goliath grouper using the ENP creel survey, 1973-2006. *Endangered Species Research* 7: 183-193.
- [2] Clark E, von Schmidt K (1965) Sharks of the central Gulf coast of Florida. *Bulletin of Marine Science* 15: 13-83.
- [3] Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77:802-813.
- [4] Efron B, Tibshirani RJ (1993) *An introduction of the Bootstrap*. Chapman and Hall/CRC.
- [5] Friedman JH (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29:1189-1232.
- [6] Friedman JH, Meulman, JJ (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicine* 22: 1365-1381.
- [7] Froeschke JT, Stunz GW, Wildhaber MW (2010) Environmental influences on the occurrence of coastal sharks in estuarine waters. *Mar Ecol Prog Ser* 407:279-292.
- [8] Froeschke JT, Froeschke BF (2011) Spatio-temporal predictive model based on environmental factors for juvenilen spotted seatrout in Texas estuaries using boosted regression trees. *Fish. Res.* 111:131-138.
- [9] Froeschke JT, Froeschke BF, Stinson CM (2013) Long-term trends of bull shark (*Carchrhinus leucas*) in estuarine waters of Texas, USA. *Can J. Fish. Aquat. Sci.* 70:1-9.
- [10] Hastie T, Tibshiriani R, Friedman JH (2001). *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York.
- [11] Ingram GW Jr., Richards WJ, Lamkin JT, Muhling B (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.
- [12] Lane JQ, Raimondi PT, Kudela RM (2009) Development of a logistic regression model for the prediction of toxigenic *Pseudo-nitzschia* blooms in Monterey Bay, California. *Mar. Ecol. Prog. Ser.* 383: 37-51.

- [13] Leathwick JR, Elith J, Francis MP, Hastie T, Taylor, P. (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Mar Ecol Prog Ser* 321: 267–281
- [14] Leathwick JR, Elith J, Chadderton WL, Rowe D, Hastie T (2008) Dispersal, disturbance, and the contrasting biogeographies of New Zealand’s diadromous and non-diadromous fish species. *Journal Biogeogr* 35: 1481-1497
- [15] Lo, C.N., Jacobsen, L.D., and Squire, J.L. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. *Can. J. Fish. Aquat. Sci.* 49: 2515-2526.
- [16] Maunder, M., and Punt, A. 2004. Standardizing catch and effort data: a review of recent approaches. *Fisheries Research* 70: 141-159.
- [17] Parisien MA, Moritz MA (2009) Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecol. Monogr.* 79: 127:154.
- [18] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.