2009

Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico lonline [sic] fishery

Shane Abeare
Louisiana State University and Agricultural and Mechanical College

COMPARISONS OF BOOSTED REGRESSION TREE, GLM AND GAM PERFORMANCE IN THE STANDARDIZATION OF YELLOWFIN TUNA CATCH-RATE DATA FROM THE GULF OF MEXICO LONLINE FISHERY

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Oceanography and Coastal Sciences

by
Shane M Abeare
B.Sc., University of Michigan, 1999
M.Sc., University of Pretoria, 2004
December 2009

Acknowledgements

I would like to thank my major professor, Dr. Joseph Powers, for his guidance and suggestions over the past couple of years leading up to the successful completion of my thesis. I would, also, like to thank my committee members Drs. Kenneth Rose, Larry Rouse and Bin Li for the constructive comments and feed-back received on my manuscript. A special thanks going to Dr. Bin Li for the time invested in discussions of the statistical aspects of my project. Lastly, I would like to extend my sincere gratitude to all those in the Department of Oceanography and Coastal Sciences instrumental in providing me the opportunity to pursue the master's program at Louisiana State University.

Table of Contents

List of Tables

List of Figures

Abstract

Recent advances in statistical understanding have focused fisheries research attention on addressing the theoretical and statistical issues encountered in standardizing catch-rate data. Similarly, the present study evaluates the performance of boosted regression trees (BRT), the product of recent progress in machine learning technology, as a potential tool for catch-rate standardization. The BRT method provides a number of advantages over the traditional GLM and GAM approaches including, but not limited to: robust parameter estimates as a result of the integrated stochastic gradient boosting algorithm; model structure learned from data and not determined *a priori*, thereby avoiding assumptions required for model specification; and easy implementation of complex and/or multi-way interactions. Performance of the BRT method was evaluated comparatively, where GLM, GAM and BRT main-effects models, and a BRT two-way model, were trained using zero-truncated, lognormal catch-rate data, with identical predictors and dataset. Data used were observer-collected records of yellowfin tuna catch from the Gulf of Mexico longline fishery, 1998-2005. Model comparisons were based, primarily, on percent deviance explained by the trained models and prediction error using a test dataset, measured as root mean squared error (RMSE). Secondarily, the relative influence of model predictors and handling of spatially correlated error structures by each of the four models were examined. Fitted GLM, GAM, BRT and BRT two-way models accounted for 19.56%, 25.10%, 26.10% and 37.3% of total model deviance, respectively. RMSE values for the GLM (0.3552), GAM (0.3554), BRT (0.3546) and BRT two-way (0.3509) models indicate that the BRT-based models performed marginally better than the traditional GLM and GAM methods, with lower prediction error. Indices of predictor influence and spatial analysis of model residuals, for the main-effects models, suggest GAM and BRT models perform comparably in the partitioning of variance amongst predictors and handling of autocorrelated variance structures. Overall, results of the

main-effects models indicate that the BRT method is as equally adept as GAMs in fitting non-linear responses, however unlike the GAM, the BRT avoided overfitting the data, thereby providing more robust estimates. The BRT two-way interaction model further demonstrates: the ability of the BRT method in fitting complex models, while avoiding overfitting; the ease with which interactions can be incorporated and specific terms extracted, such as the year term; and the potential role of complex interactions in accounting for non-stationary processes. Although the results presented here are not definitive, for every measure of performance examined the BRT-based models performed as equally well or better than the traditional GLM/GAM standardization methods, thereby confirming the utility of the BRT method for catch standardization purposes.

Introduction

In fisheries management, policy decisions are based on considerations of the current state of a fish stock and near-term population projections. Projections of the future state of a stock are determined using a population dynamics model, commonly a surplus production or age structured model. Population models used in stock assessments incorporate, either explicitly or implicitly, essential population parameters, such as growth rate, fecundity and mortality rate. Often, these parameters are calibrated to the fluctuating year-term of a statistical model used for standardizing the raw data, thereby converting the year-effect into more biologically relevant terms, biomass. Once the estimates of biomass are obtained, biological reference points pertinent to management can be calculated, such as maximum sustainable yield.

In the assessment process, the catch rate standardization step provides an important function as a statistical filter by removing bias and accounting for unexplained variation in the catch time-series. The resultant "clean" year-effect, or relative index of abundance, is then extracted from the standardization model, forming the baseline upon which a population model may be calibrated. Unexplained variation or persisting biases, resulting from the use of biased data, poorly fitted statistical distributions (Shono, 2008; Dick, 2004; Ortiz & Arocha, 2004; Terceiro, 2003) or misspecified standardization models will be passed on to the population dynamics model. Propagation of error from the standardized dataset to the population model may ultimately lead to a loss of accuracy in model predictions and management advice (Bishop, 2006).

The importance of statistical filtering becomes apparent when considering how widely the quality of catch rate data may vary. Fisheries independent data are collected during systematic research surveys, where aspects of fishing gear, temporal and spatial extent, and species recorded are all stringently controlled, providing highly-detailed data of limited bias.

Less stringently controlled data are those collected by trained fisheries observers, where observers have no control over directing the spatial or temporal extent of the fishing effort, yet high-quality, detailed data are collected on aspects of gear, environmental conditions and harvested species. Self-reported logbook data, or fisheries dependent data, may be considered the poorest of data sources due to its high level of spatio-temporal bias (*i.e.* targeting) and potentially high unintentional, or intentional, misreporting rates. Although fisheries dependent data may be considered of lesser quality, it is by far the most abundant data available, and as such, is often used in the stock assessment process.

Catch Rate Standardization, in Theory

At the core of the assessment process, a fundamental assumption exists, namely, that the year-effect resulting from the fitted standardization model, representing the change in catch over time, is proportional to changes in the underlying population abundance. This assumption follows the basic relationship:

$C_t = E_t q N_t'$ or

$\text{CPUE} = C_t/E_t = qN_t'$,

where, C_t, denotes catch in year t, E_t is effort, the catchability coefficient, q, and the local population size, N_t'. However, the relationship between catch rate and true abundance may be far more complex than the simple linearity implied in the formula above (Harley et al., 2001; Maunder & Punt, 2004; Gaertner & Dreyfus-Leon, 2004; Haggarty & King, 2006). Research has shown that catchability can be influenced by: spatiotemporal scale, climatic conditions (Rouyer *et al*., 2008), environmental conditions (Gordoa *et al*., 2000; Ziegler *et al*., 2003; Damalas *et al.*, 2007) and exploitation history (Anderdson *et al*., 2008). Moreover, catchability may be species-specific, density dependent (Tsuboi & Endou, 2008), sex and age-specific (Goni *et al.*, 2003; Solmundsson *et al*., 2003), and affected by complex interactions, thereof (Simpfendorfer *et al*.,

2002; Rouyer *et al*., 2008; Planque *et al*., in press; Perry *et al*., in press). Arreguin-Sanchez (1996) provides a good review of many of the issues surrounding catchability, its interpretation and examples from previous studies of the variability observed in its mathematical formulation.

Catch Rate Standardization, in Practice

In addition to the issues of data quality and the uncertainty that exists surrounding the CPUE-abundance relationship, the catch standardization process is further complicated by the nature of the data itself and the choices that must be made at the implementation stage of statistical modeling. Empirical distributions of catch rate data are known for being highly right-skewed, zero-inflated, overdispersed and serially-correlated making the choice of probability distribution function and statistical analyses difficult using standard parametric methods. Moreover, the level of data aggregation and spatiotemporal scale may also influence the choice of the statistical approach adopted. Often, an additive, regression-based, main-effects model is used, typically a generalized linear model (GLM) or generalized additive model (GAM), and less commonly, regression tree methods.

GLMs are used extensively in fisheries for standardizing catch rate data (*e.g.* Glazer & Butterworth, 2002; Battaile & Quinn, 2004; Maunder & Punt, 2004; Venables & Dichmont, 2004). However, empirical distributions of fisheries catch data often cannot be characterized by any of the standard, theoretical distributions of the exponential family. To overcome this problem, two-step mixture models were developed to model the zero-inflated portion of the distribution separately from the count portion. Traditionally, the zero-inflated portion is modeled using a binomial GLM, whereas the count data, or zero-truncated portion of the distribution, are modeled using a lognormal GLM, known as the delta-lognormal method (Lo et al., 1992). The delta-lognormal method has been used extensively in fisheries research, and is commonly used for stock assessment purposes, for example, yellowfin tuna (ICCAT, 2008) and swordfish

(ICCAT, 2007) stock assessments. In addition to the delta-lognormal method, other GLM-family models used in fisheries research include: delta-Poisson models (Ortiz & Arocha, 2004), delta-gamma (Sousa et al., 2007), zero-truncated models, where only the count portion of the distribution is considered (Baum *et al*., 2003) and generalized linear mixed models (GLMM) (Rodriguez-Marin, et al., 2003). Also, with the rising interest in spatial fisheries ecology (Ciannelli et al., 2008), GLM-based methods have been equipped to handle spatially auto-correlated data (Nishida & Chen, 2004), an approach akin to kriging.

In recent years, the use of GAMs to model catch rate data has become increasingly popular, given their flexibility in fitting nonlinear relationships. Much of this interest is likely attributable to the concurrently increasing interest in the spatial aspects of fisheries management. Inclusion of environmental and/or positional (*e.g.* latitude/longitude) covariates in statistical models often invoke(s) a nonlinear response (Bigelow et al., 1999; Zagaglia et al., 2004; Cheng & Gallinat, 2004; Howell & Kobayashi, 2006; Hazin & Erzini, 2008), which are easily fitted using GAM methods. Furthermore, as in the case of GLMs, GAM-based approaches have been developed to resolve the ever-present problem of zero-inflation and overdispersion. For example, Minami et al. (2007) used a zero-inflated negative binomial model with smoothing to model shark bycatch.

Despite these recent developments in adapting the GLM/GAM methods to the unique challenges presented by fisheries abundance data, in fitting any of the linear regression-based models there are a number of issues that must still be considered: 1) the order in which explanatory variables are introduced into the model, 2) potentially correlated explanatory variables, or multicollinearity, 3) variable selection, 4) outlier detection and removal, 5) model overfitting and 6) model misspecification due to either missing important variables, or a flawed mathematical form (i.e. additive versus multiplicative). The importance of variable order,

multicollinearity, variable selection and outlier detection are widely recognized concerns, where techniques to detect or mitigate these effects are commonly used. Similarly, precautions are taken to avoid overfitting by limiting the number of explanatory variables used, including interaction terms. However, in many cases, concerns of overfitting a model may lead to the needless exclusion of important interaction terms. Moreover, interactions involving categorical variables, such as "year", may be often excluded due to the difficulties that would ensue in calculating the year-effect. Regardless, missing significant variables is a form of misspecification that could result in significant losses of prediction accuracy.

Simple regression tree (SRT), or decision tree, approaches to statistical modeling have traditionally been used for classification purposes (Breiman et al., 1984, Hastie et al., 2001). In recent years, this technique has received some interest for applications in ecology (e.g. De'ath & Fabricius, 2000; Vayssieres et al., 2000; De'ath, 2002) and in fisheries research (Walsh & Kleiber 2001). Interest in SRT-based methods is largely due to the ease with which simple tree models can be visualized, and their ability at handling predictor and response variables of any type, without the need for transformations. Moreover, SRTs are insensitive to outliers, ignore insignificant predictors and automatically model complex interactions (Elith *et al.*, 2008). Despite these advantages over GLM/GAM methods, SRTs have received only limited attention in the fisheries and ecological literature due, principally, to several critical shortcomings: the difficulty in interpreting large, complex trees; the step-like response functions that result, which may be considered ecologically unrealistic by some; and the poor predictive performance of SRTs on novel datasets due to a tendency to produce overfitted models.

Boosted regression trees (BRT) offer a promising alternative approach to catch rate standardization, both conceptually and practically. Conceptually, the BRT approach is very different from the approach used for the fully-parameterized GLM and GAM methods. For

GLMs and GAMs, the structure of the model (*e.g.* additive versus multiplicative and the relationship between interaction terms) must be determined *a priori*. Hence, for fully-parameterized models, specification of the model itself or the (in)significance of a particular predictor is more the focus, rather than maximizing explained variance. The BRT approach, on other hand, is rooted in advancements in machine learning algorithms, where the final model is "learned" from the data and not predetermined. Thus, fewer parameters requiring precise specification reduces the risk of misspecification, which is of particular importance given the uncertainty of the catch-abundance relationship, and the likely existence of complex interactions. Given that catch rate standardization is more a data filtering exercise, with the objective of reducing bias and maximizing explained variance, than a test of theory, the BRT approach may be better-suited for this purpose.

With recent developments in machine learning technology, the "boosting" component of the boosted regression tree method is able to overcome many of the weaknesses of the earlier SRT method. Boosting, or more precisely, stochastic gradient boosting, increases predictive performance by reducing the over-learning, or overfitting, that commonly occurs with SRTs. Similar to the GAM method, fitted BRT functions may be linear, curvilinear or non-linear, where the choice of error distribution includes normal, binomial and Poisson (De'ath, 2007 and Elith *et al.*, 2008 for details). However, unlike the GLM or GAM methods, in fitting a BRT model there is no need for concern regarding outliers, the number or order of predictors, missing predictor values, nor variable selection, which also suggests a potential degree of immunity to the detrimental effects of multicollinearity. Moreover, interactions are easily implemented, without concern for potentially complicated calculations of the standardized year effect. Given these advantages of the BRT method, there has been recent interest in tree-based models for ecological applications (Cappo *et al*., 2005; Leathwick *et al.*, 2006; De'ath, 2007; Elith *et al.*, 2008).

In the present study, a test of the suitability and performance of BRTs as a potential technique for the standardization of catch-rate data was performed. Performance of the BRT method was evaluated comparatively, where primary comparisons were based on the results of GLM, GAM and BRT main-effects models fitted using identical data and explanatory variables. Secondarily, due to ease with which interactions can be introduced using the BRT method, a two-way interaction BRT model was fitted to illustrate: the importance of considering all interaction terms in catch standardization models, and the ability of the method in fitting complex models, without suffering from the commonly associated problem of overfitting. Comparisons focused on the relative influence of predictor variables and percent deviance explained by each of the respective modeling methods. Residuals from each of the modeling techniques were examined for conformity to model assumptions, with particular attention given to the assumption of spatial independence, a critical assumption of all modern statistical methods. Using training and testing datasets, predictions from each of the trained models were evaluated for prediction accuracy. Note, given that true abundance is unknown, the prediction accuracy of the statistical models, using a test dataset, is assumed to reflect, globally, the accuracy of parameter estimates. Therefore, the year term extracted from the model with the greatest prediction accuracy would be used for expressing annual relative abundance. Lastly, relative indices of abundance for all models examined here were calculated and plotted for visual comparison.

Methods

Data Sources and Description

The pelagic longline catch data used here were collected during the course of the Pelagic Observer Program (POP) and provided by the Southeast Fisheries Science Center. Records of longline sets available in the POP database extend from 1992-2005, covering the Gulf of Mexico, the Caribbean and Atlantic fishing zones. Details of the gear, environmental measurements, locations of sets and catch are divided amongst three separate files: the "gear-log", "haul-log" and "animal-log". Analyses performed here limited the database to information contained within the haul-log and animal-log, and therein, to records collected from the Gulf of Mexico (GOM-POP) during the period of 1998-2005. During this period, the total number of sets remained relatively stable (*ca.* 10,000 sets per year for all regions), however observer coverage increased from *ca.* 3% to 7% in later years (NOAA Technical Memorandum, NMFS-SEFSC-562).

POP Database-derived Variables

The GOM-POP animal-log dataset comprised records of catch for 114 identified species. Of the species frequently caught, yellowfin tuna (*Thunnus albacares*) were the most abundant (Figure 1). Due to the high rate of catch success (positive sets/ total sets= 1452/1603= 0.91), yellowfin tuna (YFT) were selected as the dependent variable for the statistical models compared here. Despite the relatively low zero-count, the empirical distribution of the YFT catch/set data was still too zero-inflated and over-dispersed to fit a Poisson distribution (Figure 2). Although catch/set data fit a negative binomial distribution well, not all the statistical modeling methods, used here, accommodate the negative binomial distribution (Figure 3). As is commonly done in such situations, the count data were converted to a zero- truncated, lognormal distribution (Figure 4) in using $\log(CPUE_{yft})$ of the positive count data as the response variable in the models,

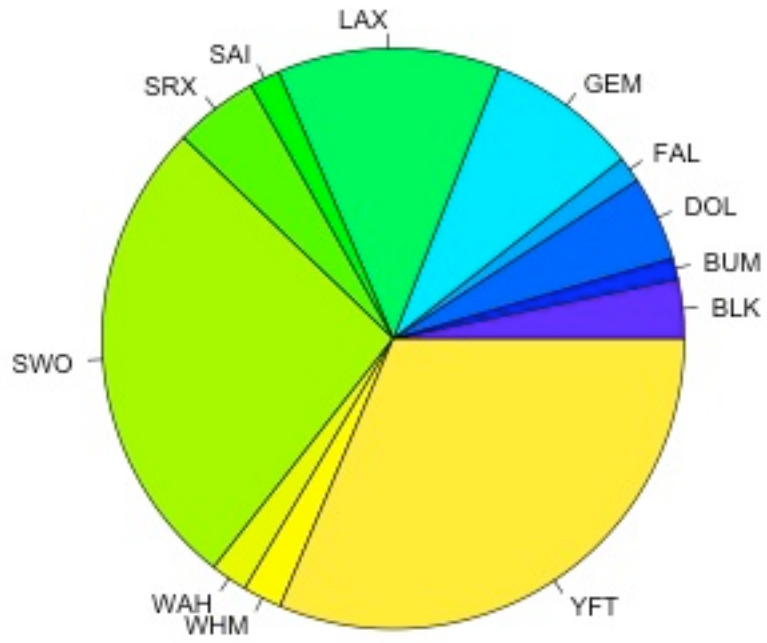


Figure 1. Relative abundance of species regularly caught on GOM longlines: blue marlin (BUM), blackfin tuna (BLK), dolphin fish (DOL), silky shark (FAL), escolar (GEM), lancetfish (LAX), Atlantic sailfish (SAI), skates/rays (SRX), swordfish (SWO), wahoo (WAH), white marlin (WHM) and yellowfin tuna (YFT)

where $CPUE_{yft} = Catch_{yft}$ / (soak time*number of hooks/1000).

From the haul log, information retained to serve as explanatory variables in the statistical models, included: year, season, target, bait kind, minimum hook depth, maximum hook depth, number of floats, number of lights and bait weight. The importance of accounting for targeting behavior in catch standardization models is widely recognized (*e.g.* Quirijns *et al*., 2008), albeit the designation of an "official" target may be rather subjective and influenced by various external forces, including fisheries policy. Of all possible targets recorded in the POP database (*i.e.* SWO= swordfish, TUN= tuna, YFT= yellowfin tuna, BET= bigeye tuna, SHX= sharks, DOL= dolphin fish or MIX=multiple species targeted), only sets where YFT, TUN, SWO or MIX are the declared target were analyzed here. Figure 5 illustrates changes in the proportion of sets for each target and bait used over the study period, where the total sets per year indicate the

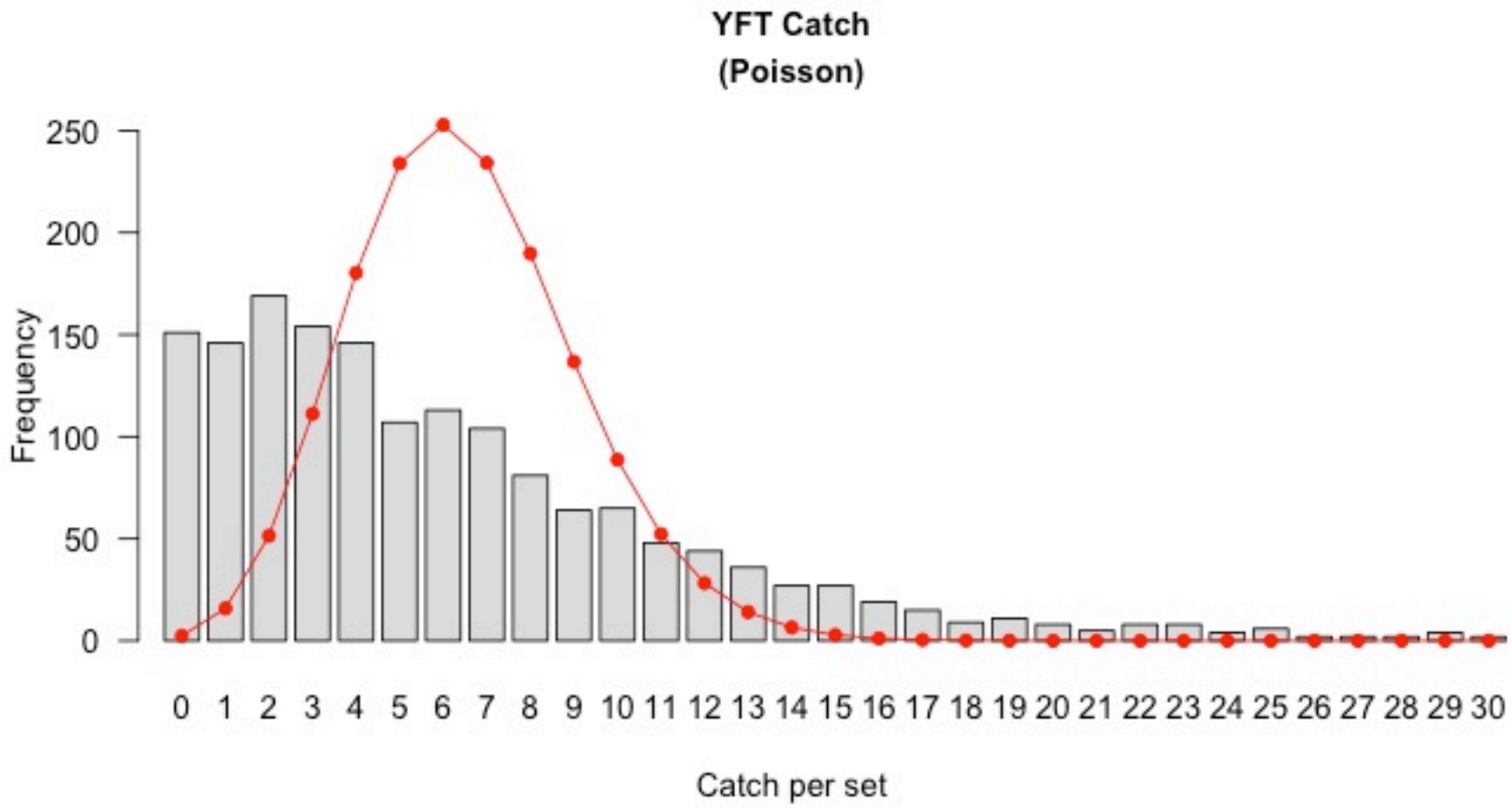


Figure 2. YFT catch/set data: red line represents the theoretical Poisson distribution (λ= 6.48); gray bars represent empirical distribution of catch/set data; x-axis truncated at 30 YFT/set for graphing purposes, max= 63 YFT/set

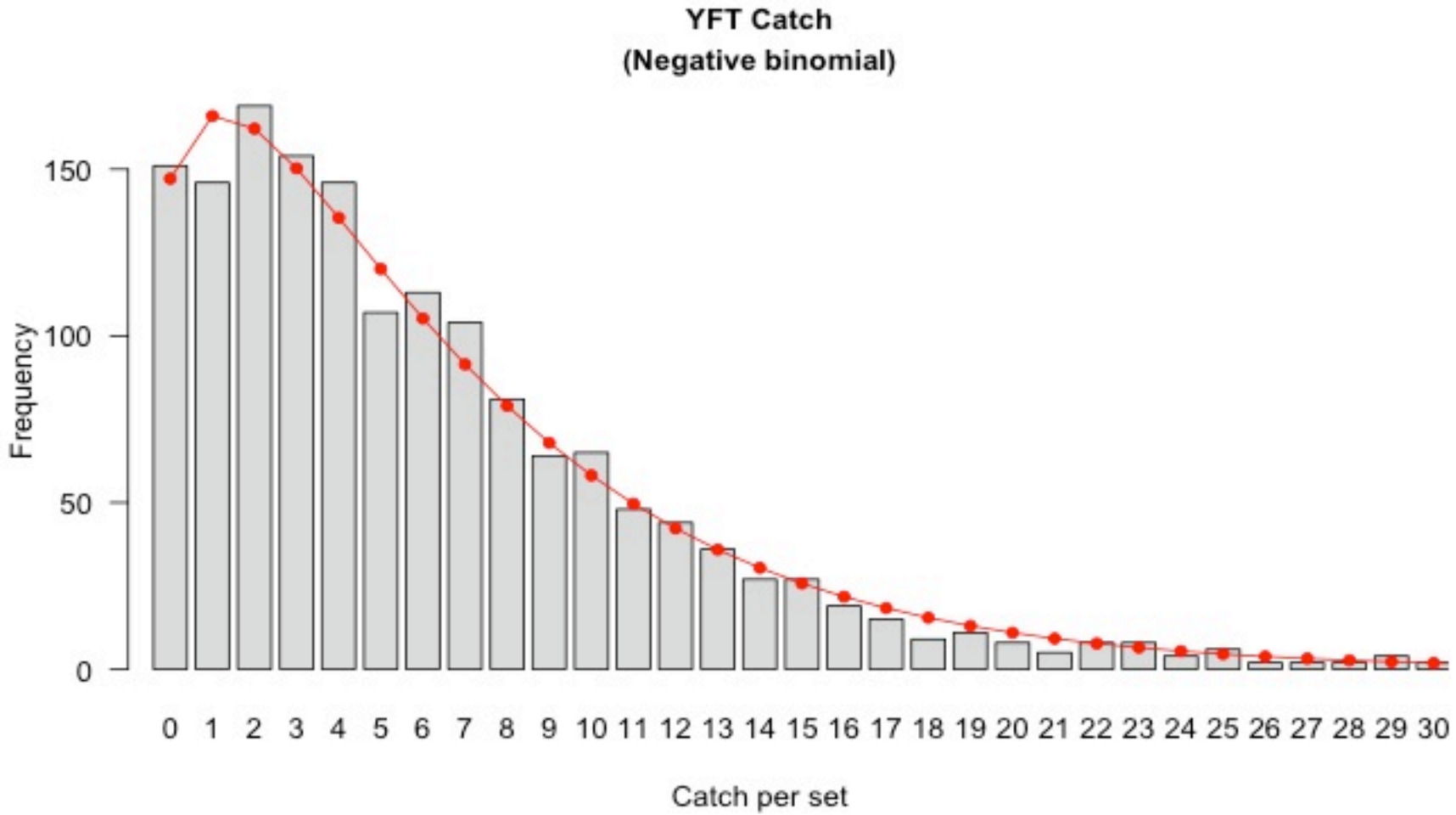


Figure 3. YFT catch/set data: red line represents the theoretical negative binomial distribution (size= 1.37, μ= 6.48); gray bars represent empirical distribution of catch/set data; x-axis truncated at 30 YFT/set for graphing purposes, max= 63 YFT/set

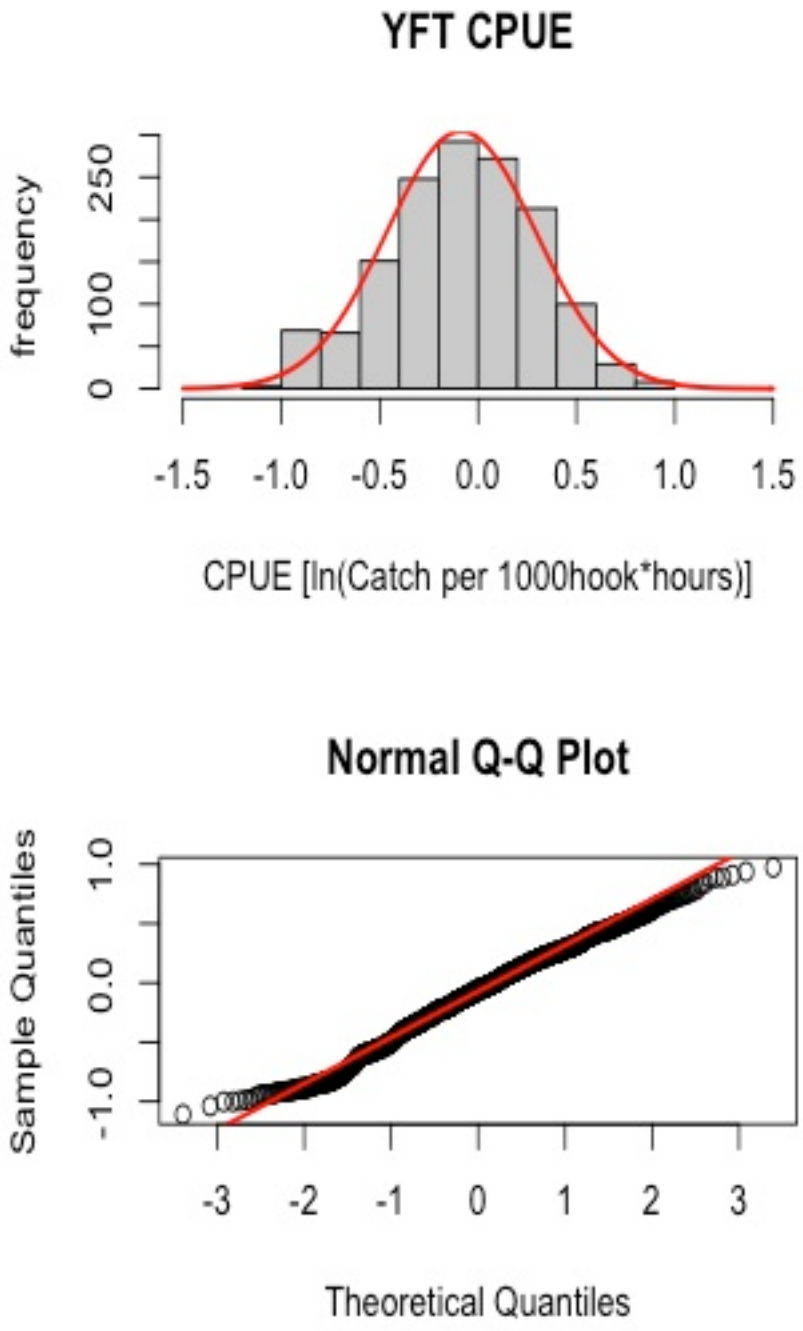


Figure 4. Log($CPUE_{yft}$): actual and theoretical normal distributions, $N(-0.086, 0.145)$ (above); actual and theoretical normal quantiles (below)

changes in observer coverage. Bait kind and bait weight were also included in the model to account for variance due to changes in targeting behavior that may be indiscernible in using the Target variable alone.

GIS-derived Variables

Also, included in the GOM-POP haul log was positional information for each longline set. Latitudinal and longitudinal coordinates were recorded for: the beginning of the sets, the end of the sets, the beginning of hauls and the end of the hauls. The four coordinates for each set were plotted to a map of the Gulf of Mexico, using ArcGIS 9.x (ESRI, 2005). For each set of four points, a polygon was delimited using the minimum convex polygon (MCP) tool from the Hawth's Tools toolbox (Beyer, 2004). Each polygon was considered to represent the approximate "area swept" by the longline during each set (Figure 6). The MCP layer was then projected to a custom, equidistant-conic projection covering the northern Gulf of Mexico, in order to obtain reliable areal estimates, in meters squared. Areal estimates for the MCP polygons were included as an explanatory variable in the statistical models (x-variable= area). Centroids of each polygon were calculated from which the xy-coordinates were included in the model (x-variables= X-cent and Y-cent), and used as the sampling points for the GIS-derived explanatory variables described below (Figure 6, inset).

In addition to gear-related variables, often catch rate standardization models include some measure of primary productivity and other environmental measurements, such as chlorophyll concentrations, sea surface temperatures, and bottom depth (Bigelow et al., 1999, Zagaglia *et al*., 2004, Cheng & Gallinat, 2004, Howell & Kobayashi, 2006, Bigelow & Maunder, 2007, Hazin & Erzini, 2008). Here, a measure of net primary productivity (x-variable= NPP) was incorporated into the models, where chlorophyll, sea surface temperature and solar irradiance were vertically integrated to the euphotic depth, following the standard Vertically Generalized Production

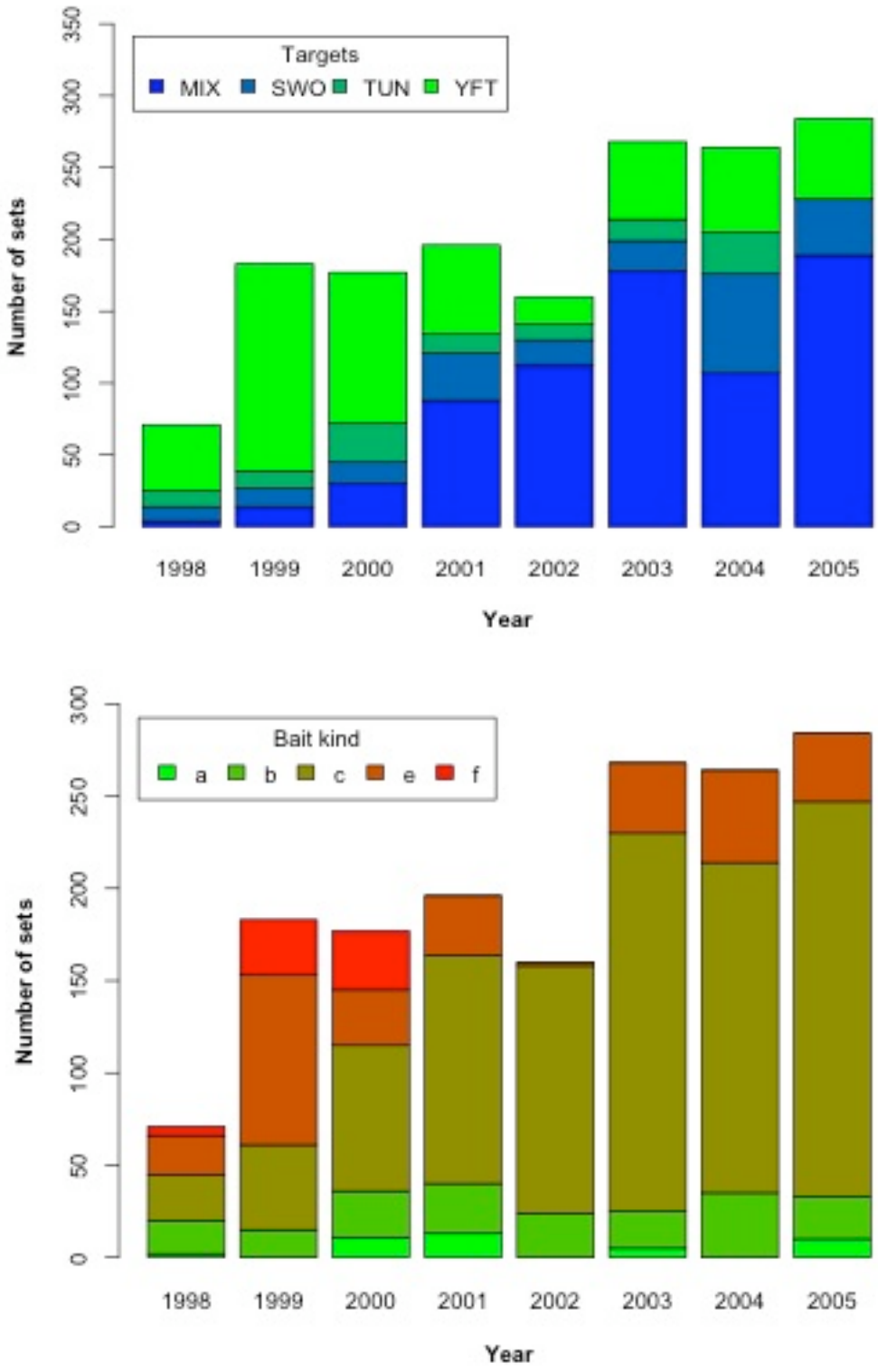


Figure 5. Number of longline sets observed per year in the POP-GOM dataset; change in declared target over the study period (above), where MIX= mixed target, SWO= swordfish, TUN= tuna and YFT= yellowfin tuna; change in bait used over the study period (below), where a= mackerel, b= herring, c= squid, e= sardine and f= scad

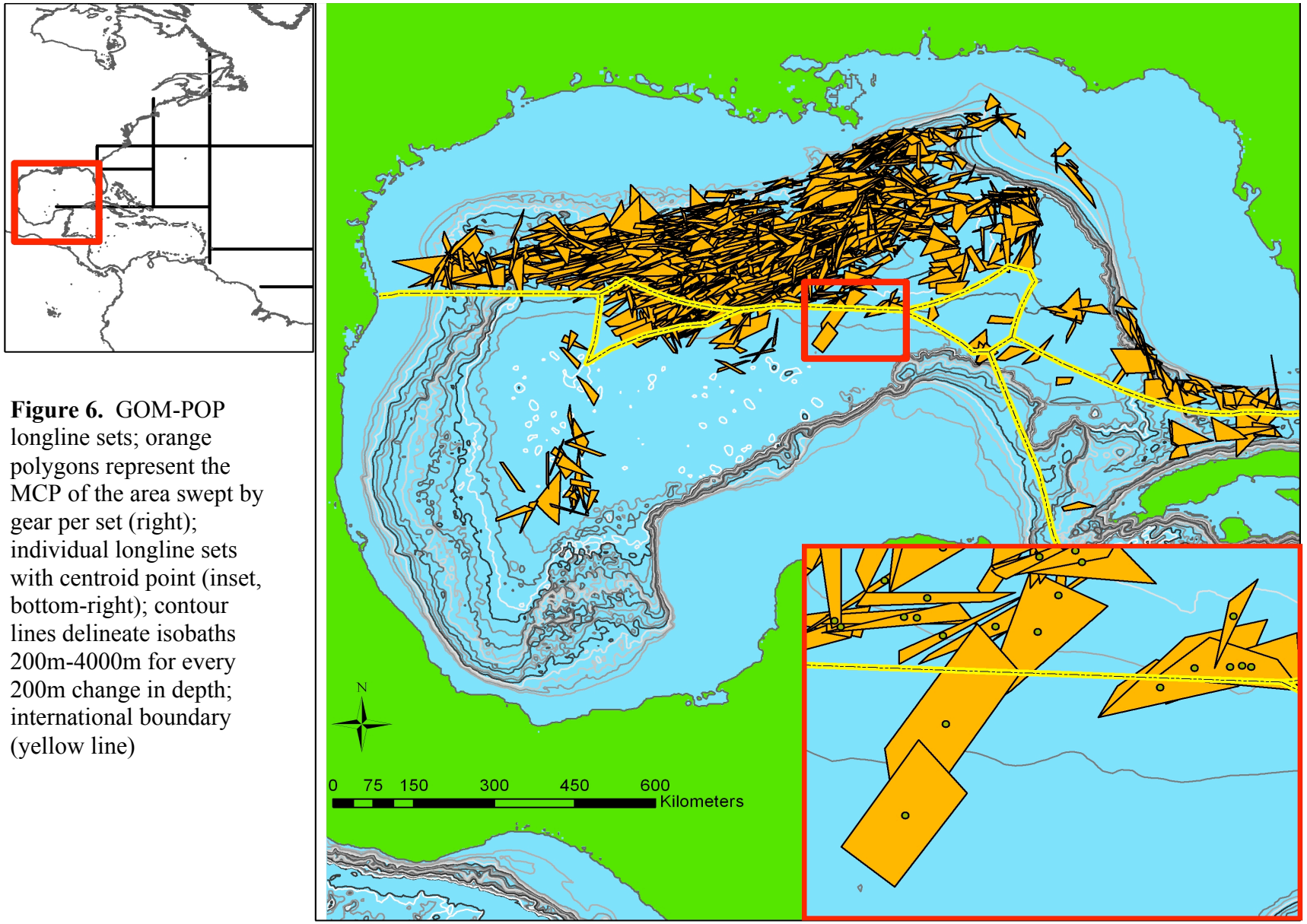


Figure 6. GOM-POP longline sets; orange polygons represent the MCP of the area swept by gear per set (right); individual longline sets with centroid point (inset, bottom-right); contour lines delineate isobaths 200m-4000m for every 200m change in depth; international boundary (yellow line)

Model (VGPM) of Behrenfeld and Falkowski (1997). Monthly averages of the VGPM-based NPP estimates (resolution= 1080 x 2160= 10-min= *ca.* 18.5 km) were downloaded for the study period, 1998-2005, from the Ocean Productivity homepage. For a more detailed description of the calculations and satellite data used, see the Ocean Productivity home page (http://www.science. oregonstate.edu/ocean. productivity/index.php).

Estimates of bottom depth (x-variable= depth) were obtained from bathymetry data downloaded from the NOAA National Geophysical Data Center, Geophysical Data System (http://www.ngdc.noaa.gov/mgg/bathymetry /relief.html). Downloaded bathymetry data, originally of ETOPO 2-minute resolution, were resampled using bilinear interpolation to a 10-min grid for compatibility with NPP grids. Data values for the NPP and bathymetry variables were then obtained by sampling the grids at the centroid locations. Sampling of the bathymetry data was performed using the nearest neighbor procedure, whereas sampling of the 96 monthly, NPP layers was performed using the bilinear interpolation method. All grid-sampling procedures were performed using the Marine Geospatial Ecology Tools toolbox in ArcGIS 9.x (Roberts, *et al.*, in review).

Statistical Models

Before models were fitted, the dataset (N=1452 sets) used was split 70/30 into a training (N=1016 sets) and test set (N=436 sets), respectively, using stratified random sampling. Given that the original dataset was unbalanced, re-sampling was based on year strata to avoid further biasing of the results, particularly calculations of year-effect. Models fit to the training set were used to determine the year-effect from the training set data, whereas the test set was used for making predictions and evaluating performance measures. In order to maintain comparability between the modeling methods, the process of selecting, or reducing, the number of explanatory variables was not performed. Thus, all the same explanatory variables were used in all models

examined here. Similarly, statistical outliers were not removed, as an outlier for one method may not be an outlier for another. The different modeling techniques and functions used for fitting the models are described below.

All explanatory variables were log transformed, except X-cent and Y-cent, for use in the final statistical models. Table 1 presents a summary of conversions, transformations and descriptions of explanatory variables and factor levels included in the models. The basic formula used for all statistical models compared here followed the form:

$CPUE_{yft}$ = year + season + target + bait kind + *log(depth)* + *log(NPP)* + *log(area)* +
log(min. hook depth) + *log(max. hook depth)* + *log(number floats)* +
log(number lights) + *log(bait weight)* + *X-cent* + *Y-cent*,

where continuous predictors are italicized.

Generalized Linear Model

The most commonly used statistical models in fisheries management are generalizations of the ordinary least-squares method (OLS), generalized linear models (GLM) (Nelder and Wedderburn, 1972) and generalized additive models (Hastie and Tibshirani, 1987). In GLM models, a differentiable, monotonic link function $g()$ is used to relate the response variable, y, to the predictor variables, x, such that:

$$g(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon_i \quad (1)$$

The link function, acting as an intermediate between the linear predictor and response portions of the model, allows for the use of non-normal distributions, including normal, Poisson, gamma and binomial distributions. As such, response variables may be either discrete or continuous, and explanatory variables either quantitative or categorical. Moreover, the rigid assumption of

Table 1. Summaries and descriptions of all explanatory variables used in the statistical models: variable, variable type, unit conversions, transformations, summary statistics (minimum, mean and maximum) and descriptions of factor-level code

Variable	Type	Conversion	Transformation	Min.	mean	Max.	Description
year: 1998	factor	-	-	-	-	-	1998
1999	factor	-	-	-	-	-	1999
2000	factor	-	-	-	-	-	2000
2001	factor	-	-	-	-	-	2001
2002	factor	-	-	-	-	-	2002
2003	factor	-	-	-	-	-	2003
2004	factor	-	-	-	-	-	2004
2005	factor	-	-	-	-	-	2005
Season: FA	factor	-	-	-	-	-	Fall (Oct, Nov, Dec)
WI	factor	-	-	-	-	-	Winter (Jan, Feb, Mar)
SP	factor	-	-	-	-	-	Spring (Apr, May, June)
SU	factor	-	-	-	-	-	Summer(July, Aug, Sept)
Target: MIX	factor	-	-	-	-	-	Mixed
TUN	factor	-	-	-	-	-	Tuna
SWO	factor	-	-	-	-	-	Swordfish
YFT	factor	-	-	-	-	-	Yellowfin tuna
Bait kind: a	factor	-	-	-	-	-	Mackeral
b	factor	-	-	-	-	-	Herring
c	factor	-	-	-	-	-	Squid
e	factor	-	-	-	-	-	Sardine
f	factor	-	-	-	-	-	Scad
Ldepth	continuous	-	log	2.29	3.28	3.58	log(depth) in meters
LNPP	continuous	-	log	2.28	2.63	3.70	log(NPP)
Larea	continuous	-	log	0.79	2.49	3.60	log(area) of MCP in sq. kms.
LminHD_m	continuous	fathoms->meters	log	0.56	1.76	2.02	log(min. hook depth) in meters
LmaxHD_m	continuous	fathoms->meters	log	1.44	1.77	2.11	log(max. hook depth) in meters
Lnum_fl	continuous	-	log	0.30	2.23	3.32	log(number of floats)
Lnum_light	continuous	add 1E-7	log	-7.00	-0.66	3.07	log(number of lights) + C; due to high zero count
Lbait_wght	continuous	-	log	0.90	2.24	2.97	log(bait weight)
X-cent	continuous	-	-	-96.17	-90.73	-82.20	Centroid x-coordinate
Y-cent	continuous	-	-	22.02	26.76	29.46	Centroid y-coordinate

linearity from the OLS method is relaxed to include a limited set of non-linear relationships, as defined by the distribution-specific link function (*e.g.* Gaussian = link(identity), Poisson = link(log), binomial = link(logit), *etc*). Fitting of a GLM model is achieved through optimization of maximum likelihood estimates by an iteratively reweighted least-squares mechanism.

In the present study, the fitting of the main-effects GLM was obtained using the "stats" library of R statistical software, version 2.9.1. In fitting the GLM, default parameters were used with the distribution family= "Gaussian" and link= "identity". Although parameter estimates obtained from a Gaussian GLM fit are similar to those obtained from OLS regression, the GLM method was preferred in order to maintain comparability with the GAM, as optimization for both methods is likelihood-based.

Generalized Additive Model

Generalized additive models (GAM) provide for an even wider generalization of the response-predictor relationship than the limited set of nonlinear relationships afforded by GLM link functions. Rather than specifying a link-relationship, a nonparametric smoothing function *f* approximates the individual response-predictor relationships, such that:

$$g(y_i) = \beta_0 + \sum_{j=1}^{p} f_i(x_{ji}) + \varepsilon_i . \qquad (2)$$

Given the use of a nonparametric smoothing function, yet the need to specify a distribution for the response variable, GAMs are more aptly considered as semi-parametric statistical models (Guisan et al., 2002). The nonparametric smoothing functions allow for highly nonlinear relationships, depending on the degree of smoothing and smoothing function used. The cubic spline function produces the closest fit, whereas the loess smoother is based upon a locally weighted averaging technique. Model estimates are obtained by minimizing a penalized negative log-likelihood function.

Here, a main-effects GAM was fitted using the default parameters of the "gam" function from the "mgcv" library (Wood, 2001), where family= "Gaussisan" and link= "identity". By default, the smooth terms of a model are represented using penalized regression splines, with selection of smoothing parameters determined by the minimization of an internal generalized cross validation function. Parameter estimates for models are obtained through a penalized likelihood maximization problem solved by penalized iteratively reweighted least squares (See Wood, 2004 and Wood, 2008 for further details).

Boosted Regression Tree

Boosted regression trees (BRT) are a combination of two powerful statistical techniques: boosting and regression trees. Boosting is a machine learning technique similar to model averaging, where the results of several competing models are merged. Unlike model averaging, however, boosting uses a forward, stage-wise procedure, where tree models are fitted iteratively to a subset of the training data. Subsets of the training data used at each iteration of the model fit are randomly selected without replacement, where the proportion of the training data used is determined by the modeler, the "bag fraction" parameter. This procedure, known as stochastic gradient boosting, introduces an element of stochasticity that improves model accuracy and reduces overfitting (Elith *et al.*, 2008).

Initially, 50 trees are fitted in the normal manner, using recursive binary partitioning of the data. Residuals from the initial fit are then fitted with another set of 50 trees, these residuals are then fitted with another set of trees, and so forth, whereby the process focuses more-and-more on extreme observations. Trees are fitted iteratively until a specific loss function is minimized, verified through n-fold cross-validation. In the case of regression trees, the loss function minimized is model deviance. Final fitted values are based on the entire dataset and

computed as the sum of all trees multiplied by the learning rate (See Elith et al., 2008; De'ath, 2007 and references therein for further details).

In fitting a BRT, two parameters must be specified, the learning rate and the tree complexity. The learning rate determines the contribution of each successive tree to the final model, as it proceeds through the iterations. The tree complexity fixes whether the model will be main effects only (tree complexity =1), or whether interactions should be included (tree complexity= 2, 3, …). Ultimately, the learning rate and tree complexity combined determine the total number of trees in the final model.

Fitted BRT models, compared here, were obtained using the BRT script provided by Elith *et al.* (2008), which references the "gbm" library (Ridgeway, 2007) in R. Default parameters of the BRT script were used, where learning rate= 0.01 and tree complexity= 1 and cross-validation= 10-fold. However, the bag fraction was changed from the default value, 0.75, to 0.5. Parameters for the two-way interaction model were the same as those above, except tree complexity= 2.

Model Comparisons

For each of the fitted models, the pseudo-R^2, or D^2, was calculated for comparison, where: D^2 = 1 – (residual deviance/total deviance). Here, it was not necessary to calculate an adjusted D^2 because all the models have the same number of terms. In addition, a measure of relative influence (*RI*) was calculated for each of the predictor terms (x_p) in the model to facilitate comparisons of term-wise contributions in deviance reduction, where:
RI (x_p) = (reduction in deviance/ *df* (x_p)) / total explained deviance. This formula was used for the calculation of relative influence for the fitted GLM and GAM models, whereas for the BRT-based models, the calculations requiring the summation of contributions in the regression tree were performed using the BRT script provided by Elith *et al.* (2008).

Diagnostics were performed on each of the model fits to confirm conformity to standard regression assumptions, with particular attention given to the assumption of data independence. Given that spatial autocorrelation is commonly a problem in modeling species abundance, semivariograms were used to test for the presence of autocorrelation in the response data ($CPUE_{yft}$). Empirical semivariograms were also calculated for the residuals of each of the four models in order to compare how each of the models handled spatially structured variance. The empirical semivariogram (γ) is calculated as follows:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [y(u_\alpha + h) - y(u_\alpha)]^2, \quad (3)$$

where h is the lag Euclidean distance between data point pairs in decimal degrees, $N(h)$ the number of pairs separated by lag h, the variable of interest, y, and a specific point location, u_α. Semivariogram calculations were done using the "geoR" library (Ribeiro & Diggle, 2001) in R.

Lastly, the predictive performance of the four statistical modeling techniques was evaluated. Root mean square errors (RMSE) were calculated for predictions of $\log(CPUE_{yft})$ made by the fitted models, using the predictor values of the test dataset. Paired t-tests were then used to test the actual versus predicted $\log(CPUE_{yft})$, whereas F-tests were used to test for differences in the variance of predictions between the modeling techniques. The final product of the catch rate standardization process is a plot of the year term, α_t, which is extracted from the fitted statistical model (*i.e.* $\exp(\alpha_t)$) and centered to its mean, referred to as the relative index of abundance. For a final visual comparison of the four modeling techniques, the relative indices of abundance derived from each model were plotted.

Results

Fitted Models

Generalized Linear Model (GLM): Main Effects Model

Overall, the GLM accounted for 19.56% of the unexplained model deviance (explained deviance/null deviance = 28.18/144.08). Despite the modest reduction in unexplained deviance, the full GLM was significantly distinct from the null model ($F=8.895$, $df=27$, $p<0.001$; AIC=735.66). Variables determined to be significant, or marginally so, in predicting $CPUE_{yft}$ included, in descending order of significance: ln(area) ($p<0.001$), target:TUN ($p<0.001$), X-cent ($p=0.001$), ln(maxHD_m) ($p=0.018$), target:YFT ($p=0.02$), season:SU ($p=0.05$), bait kind:f ($p=0.067$), ln(minHD_m) ($p=0.068$) and year:2004 ($p=0.089$) (Table 2). The five most influential variables, as determined by the RI calculation were: 1) target (8.45%), 2) ln(area) (5.80%), 3) year (5.48%), 4) bait kind (4.86%) and 5) X-cent (4.01%) (Table 2; See Figure 7).

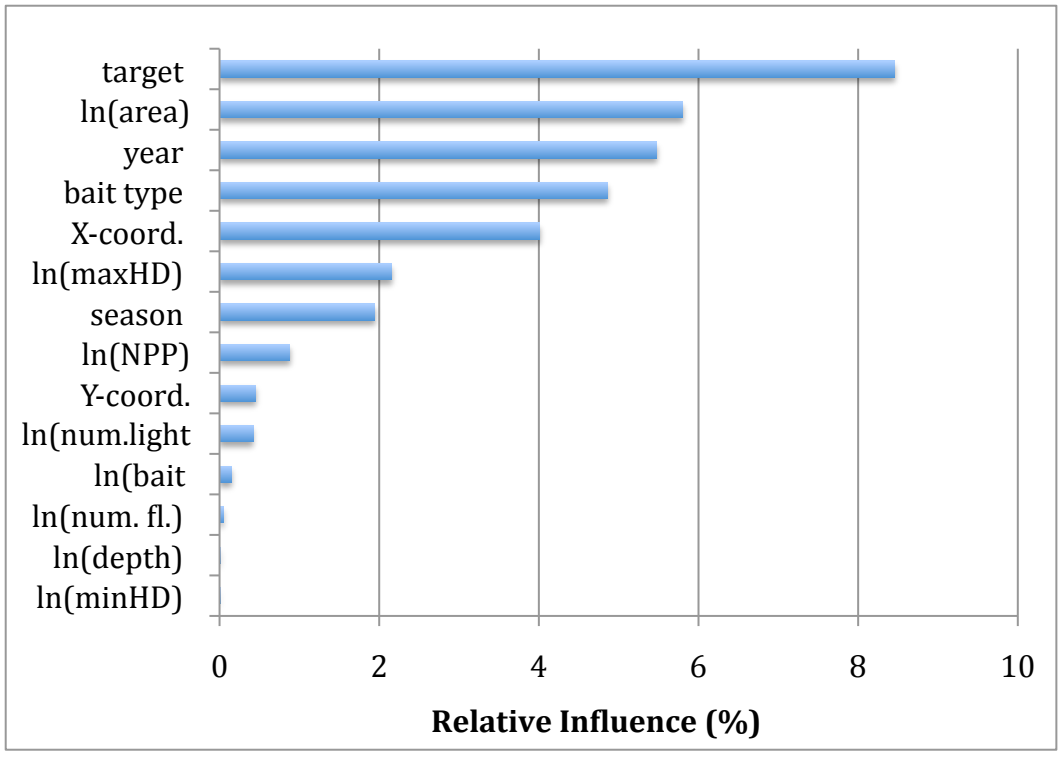


Figure 7. GLM: Ranked relative influence of model terms as a percentage of total deviance explained per df

Table 2. GLM: Significance levels of model terms; "." = 0.1>p>0.05, "*" = 0.05>p>0.01, "**" =0.01>p>0.001, "***" = p<0.001

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-1.514	1.017	-1.490	0.137	
year 1999	0.016	0.063	0.248	0.804	
2000	-0.001	0.066	-0.008	0.994	
2001	-0.075	0.066	-1.150	0.250	
2002	-0.045	0.069	-0.652	0.515	
2003	-0.092	0.064	-1.429	0.153	
2004	-0.108	0.064	-1.702	0.089	·
2005	-0.066	0.065	-1.022	0.307	
season SP	0.025	0.032	0.796	0.426	
SU	-0.065	0.033	-1.960	0.050	·
WI	-0.032	0.039	-0.826	0.409	
Target SWO	0.033	0.038	0.854	0.393	
TUN	0.219	0.065	3.377	0.001	***
YFT	0.108	0.047	2.328	0.020	*
bait kind b	-0.081	0.124	-0.654	0.513	
c	-0.031	0.123	-0.250	0.803	
e	0.081	0.125	0.648	0.517	
f	0.248	0.135	1.832	0.067	·
ln(depth)	0.071	0.085	0.833	0.405	
ln(NPP)	-0.150	0.100	-1.496	0.135	
ln(area)	-0.103	0.027	-3.837	0.000	***
ln(minHD_m)	0.422	0.231	1.825	0.068	·
ln(maxHD_m)	-0.571	0.240	-2.375	0.018	*
ln(num. fl.)	-0.009	0.092	-0.102	0.919	
Ln(num.light)	-0.005	0.006	-0.790	0.429	
ln(bait wght.)	-0.044	0.069	-0.645	0.519	
X-cent.	-0.020	0.006	-3.260	0.001	**
Y-cent.	0.018	0.017	1.043	0.297	

Table 3. GLM: Relative influence of model components, as determined by term-wise reductions in deviance per degree of freedom

	Df	Deviance	Residual Df	Residual Deviance	Relative Influence (%)
NULL			1015	144.077	-
Year	7	10.805	1008	133.272	5.48
Season	3	2.192	1005	131.080	2.59
Target	3	7.145	1002	123.934	8.45
bait type	4	4.108	998	119.827	3.64
ln(depth)	1	0.005	997	119.822	0.02
ln(NPP)	1	0.248	996	119.574	0.88
ln(area)	1	1.634	995	117.940	5.80
ln(minHD_m)	1	0.000	994	117.939	0.00
ln(maxHD_m)	1	0.606	993	117.333	2.15
ln(num. fl.)	1	0.014	992	117.319	0.05
ln(num.light)	1	0.118	991	117.201	0.42
ln(bait wght.)	1	0.041	990	117.160	0.15
X-cent.	1	1.131	989	116.029	4.01
Y-cent.	1	0.128	988	115.901	0.45

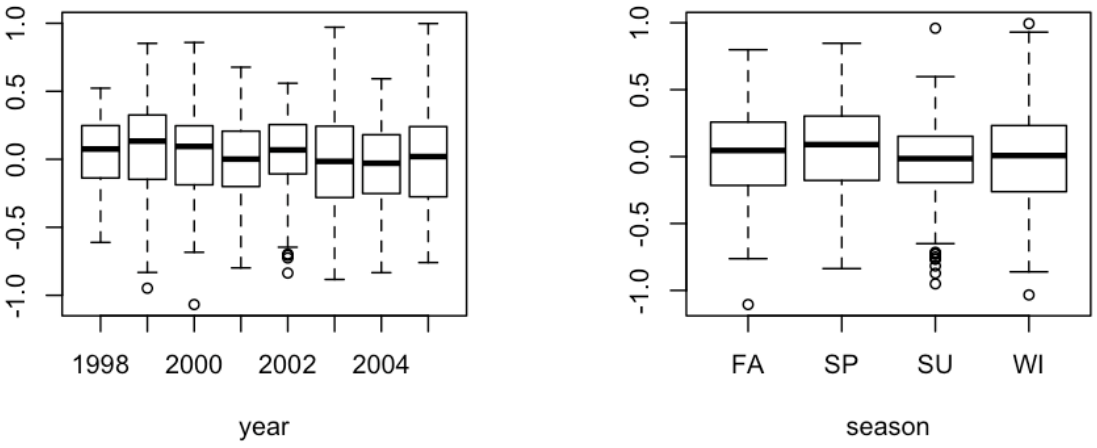
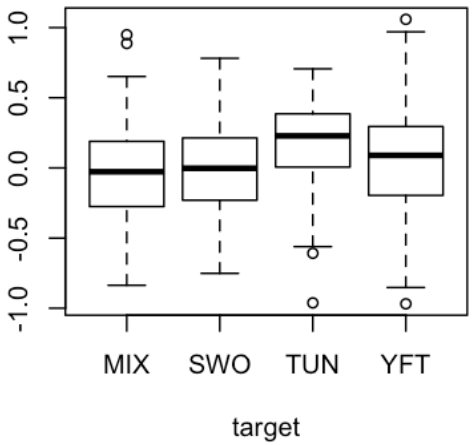
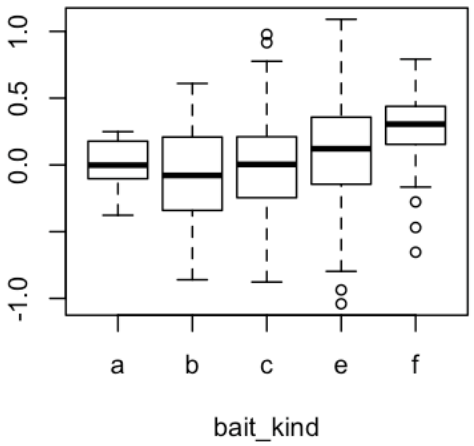
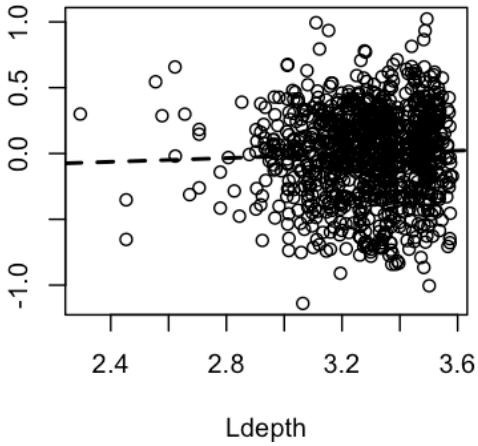


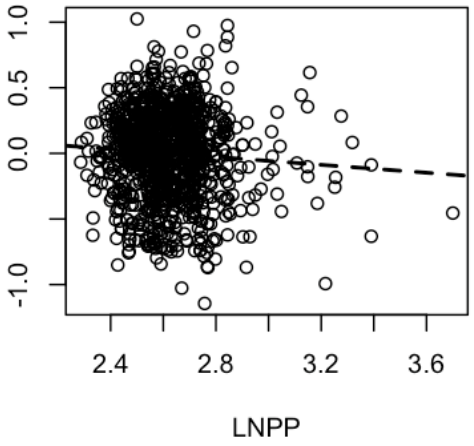
Figure 8. GLM: Partial residual plots; dashed lines represent least-squares fit, continued below

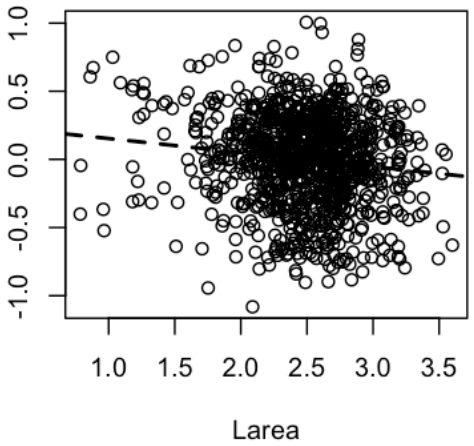
figure continued











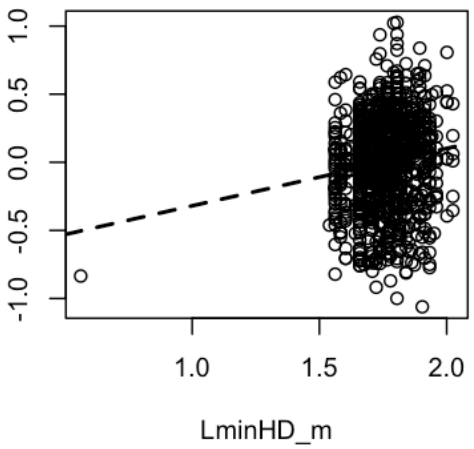
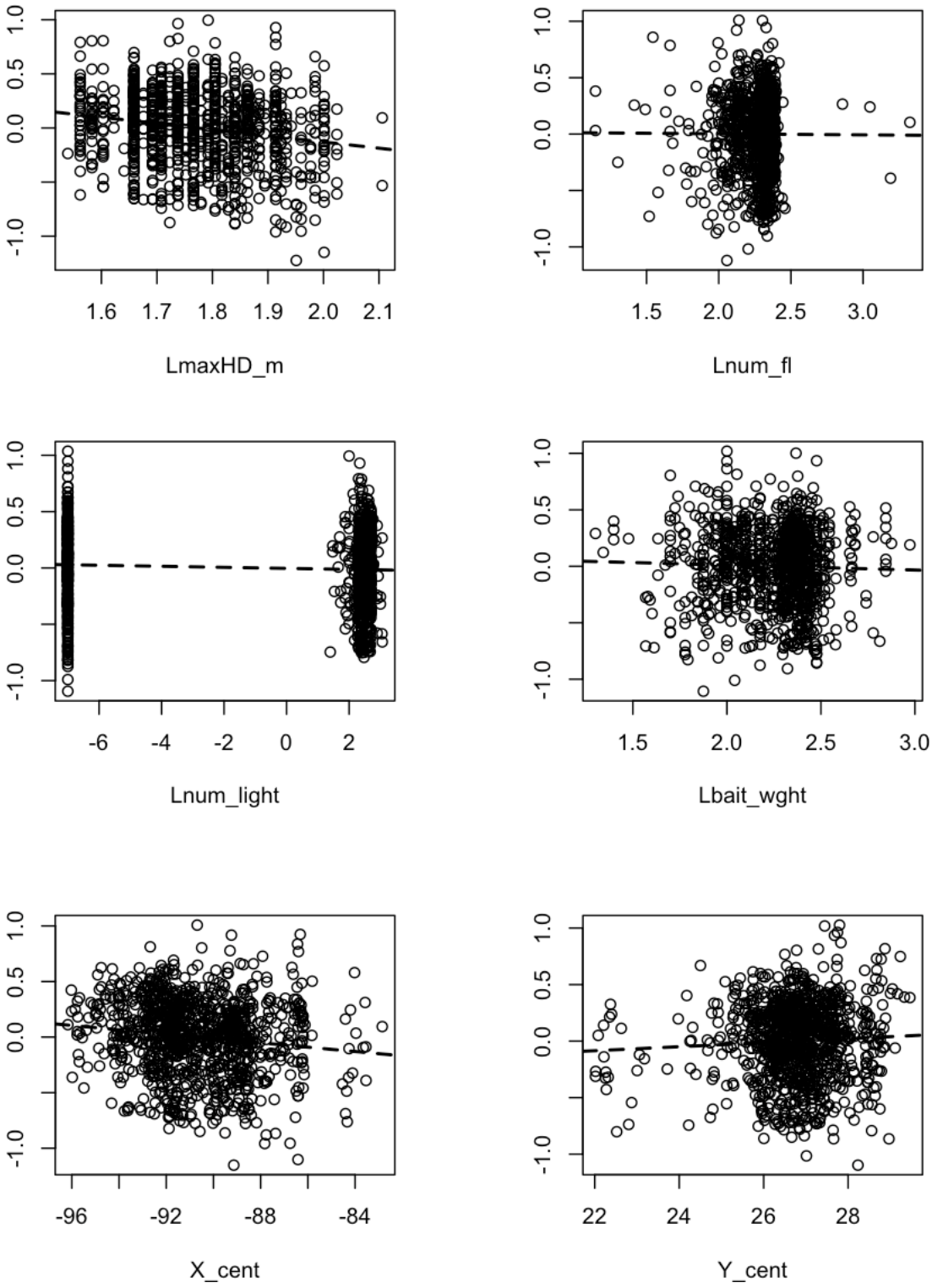


figure continued



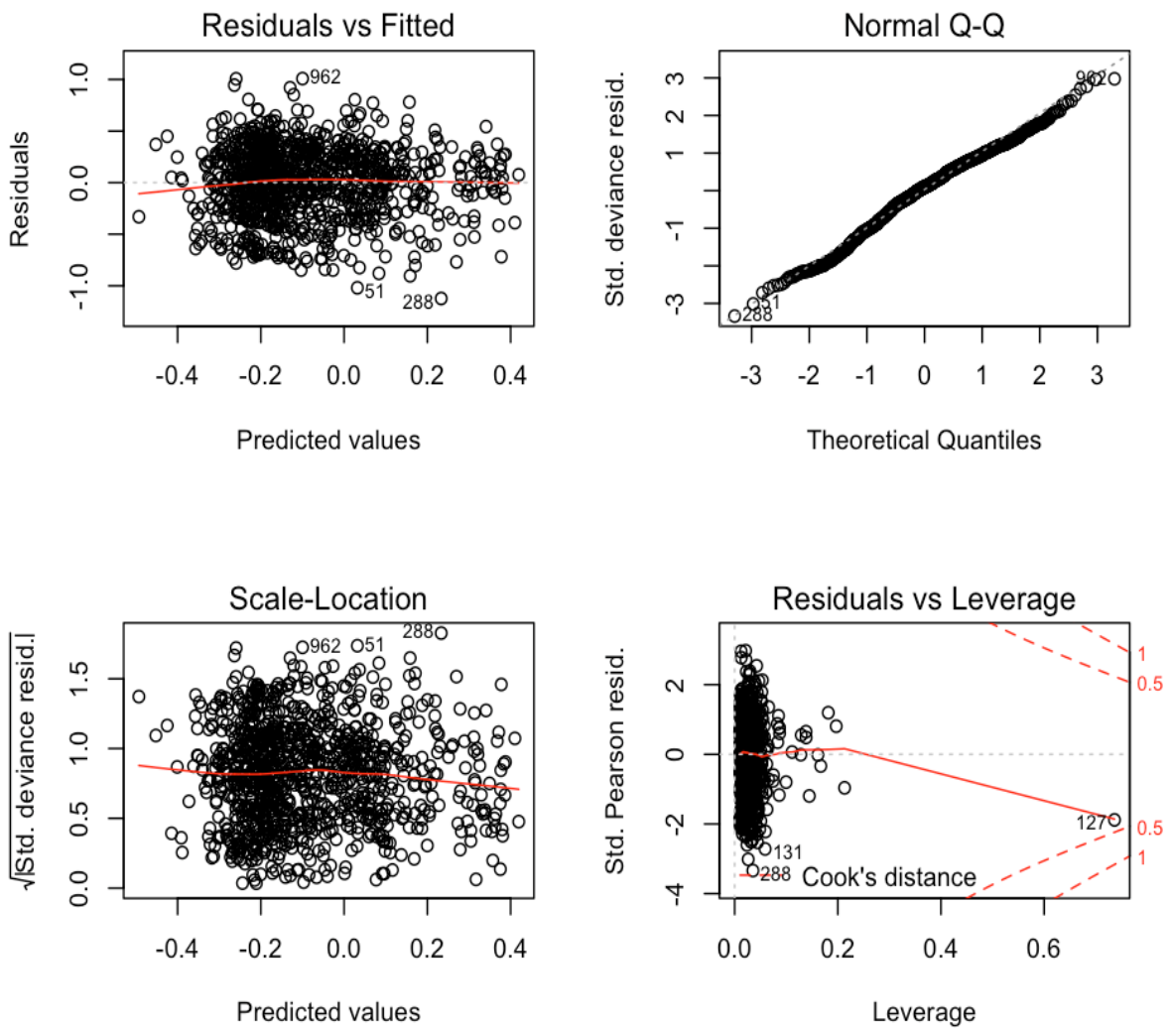


Figure 9. GLM: Model diagnostics; red line represents smooth fit; in the leverage plot dashed red line represents Cook's distance

As can be seen in the partial residual plots, the relationship between the log transformed response and explanatory variables appears well-approximated by a linear relationship (Figure 8). Partitioning of variance about the least-squares fit (Figure 8, dashed line) appears equal, which is congruent with the overall homoscedasticity observed in the standard diagnostic plots (Figure 9). Additionally, the diagnostic plots of model residuals, and standardized residuals, versus predicted values illustrate the absence of any trend or pattern. The presence of which would indicate possible autocorrelation in the data, or important variables missing from the model. Conformity to the assumption of normality is confirmed by the Q-Q plot of model residuals versus normal quantiles. The leverage plot indicates the presence of one outlier (observation #127), however, to maintain comparability between the modeling methods, no changes were made to the data, nor the variables used.

Generalized Additive Model (GAM): Main Effects Model

Results of the GAM fit indicated a significant overall reduction in model deviance of 25.10% (explained deviance/null deviance=36.13/144.077) ($F=7.46$, $df=44$, $p<0.001$). Significant and marginally significant variables in the model, included: s(Larea) ($p<0.001$), s(Y-cent) ($p<0.001$), target:TUN ($p<0.001$), s(X_cent) ($p=0.006$), target:YFT ($p=0.012$), year:2004 ($p=0.040$), year:2003 ($p=0.054$), s(Lbait_wght) ($p=0.076$) and s(LminHD_m) (0.092) (Table 4). Variable rankings based on reduction in deviance per *df* for the five most influential explanatory variables were: 1) target (6.59%), 2) ln(area) (6.35%), 3) year (4.27%), 4) bait type (2.84%) and 5) season (2.02%) (Table 5; See Figure 10).

The partial residual plots for the model clearly illustrate for which variables the non-linear fitting ability of the GAM is important (Figure 11). For variables Ldepth, LNPP, Larea, LmaxHD_m, Lnum_fl and Lnum_light, the response-predictor relationship is well approximated as linear, thus a linear model may have been sufficient for these variables alone. However, the

relationship for X-cent is curvilinear, and that for Lbait_wght and Y-cent, non-linear. The ability of a GAM to model linear and non-linear relationships explains, in particular, the increased performance of the GAM over the GLM in accounting for unexplained variation in the models examined here, and also their increased performance over GLMs, in general.

Diagnostic plots produced from the GAM residuals indicate a strong conformity to model assumptions of normality (See Figure 12, Q-Q plot and histogram), with no apparent trend in the residual *versus* linear predictor, or response *versus* fitted values plots. The Cook's distance plot indicates the presence of one potential outlier (observation #124, D_i=0.24; Figure 13), however for Cook's distance values (D_i) that are less than one no action is required.

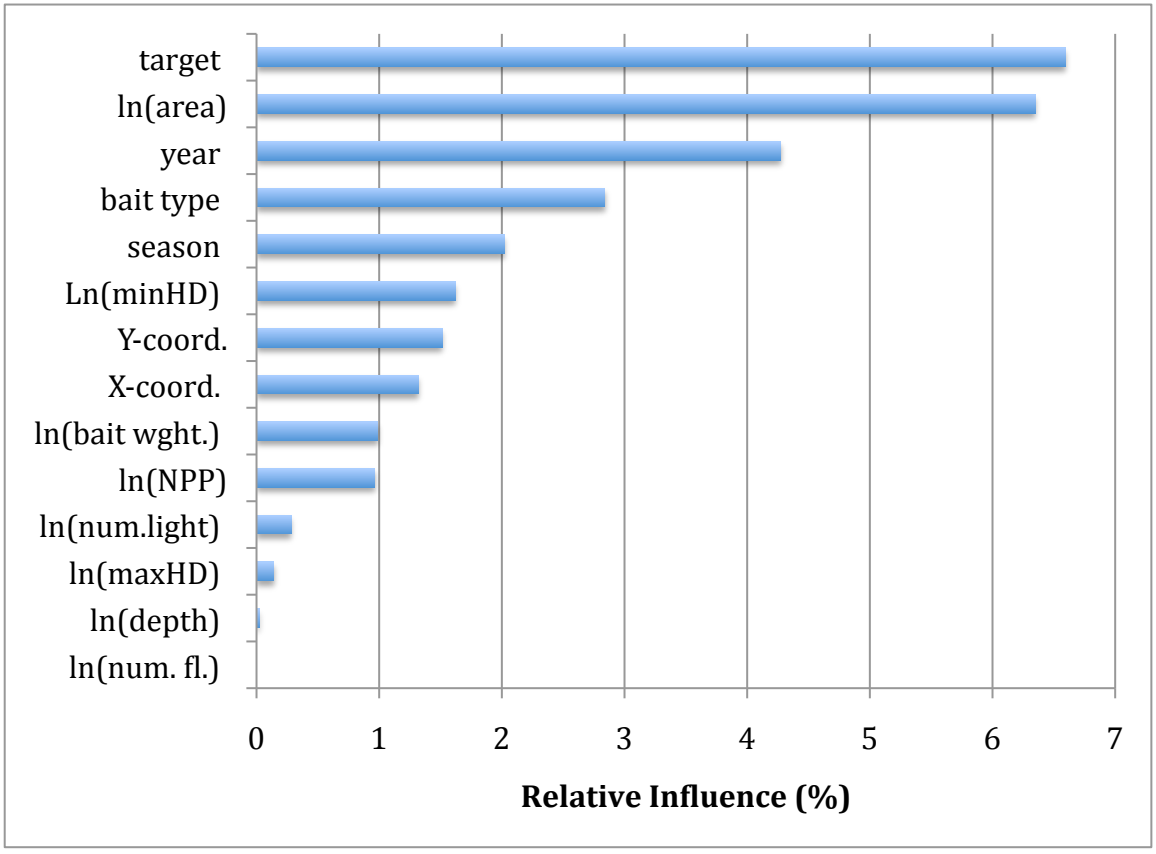


Figure 10. GAM: Ranked relative influence of model terms as a percentage of total deviance explained per df

Table 4. GAM: Significance levels of model terms; "." = 0.1>p>0.05, "*" = 0.05>p>0.01, "**" =0.01>p>0.001, "***" = p<0.001

Parametric coefficients:

	Estimate	Standard Error	t-value	Pr(>|t|)	
(Intercept)	0.016239	0.137038	0.119	0.905695	
year 1999	-0.007042	0.06524	-0.108	0.91406	
2000	-0.03379	0.067539	-0.5	0.616976	
2001	-0.10453	0.068603	-1.524	0.127911	
2002	-0.055999	0.072694	-0.77	0.441283	
2003	-0.132869	0.068934	-1.927	0.054209	.
2004	-0.140609	0.068276	-2.059	0.039721	*
2005	-0.109266	0.06955	-1.571	0.116501	
season SP	0.028646	0.03225	0.888	0.374634	
SU	-0.052296	0.034744	-1.505	0.132605	
WI	-0.04253	0.039776	-1.069	0.285225	
target SWO	0.057529	0.039202	1.468	0.142561	
TUN	0.222188	0.06567	3.383	0.000744	***
YFT	0.1164	0.046384	2.509	0.012252	*
bait_kind b	-0.115979	0.124216	-0.934	0.350698	
c	-0.089788	0.123132	-0.729	0.466055	
e	-0.035228	0.125186	-0.281	0.778457	
f	0.184152	0.134865	1.365	0.172427	

Approximate significance of smooth terms:

	Estimated df	F	p-value	
s(Ldepth)	1	0.125	0.723256	
s(LNPP)	1.39	1.916	0.160925	
s(Larea)	1	17.288	3.50E-05	***
s(LminHD_m)	2.175	2.337	0.092257	.
s(LmaxHD_m)	1.941	1.141	0.318701	
s(Lnum_fl)	1	0.179	0.672488	
s(Lnum_light)	1	0.109	0.74095	
s(Lbait_wght)	5.536	1.949	0.076244	.
s(X_cent)	2.935	4.233	0.005882	**
s(Y_cent)	8.605	3.658	0.000215	***

Table 5. GAM: Relative influence of model components, as determined by term-wise reductions in deviance per degree freedom: df, Deviance, Residual df, Residual Deviance and Relative Influence

	df	Dev.	Resid. df	Resid. Dev.	Rel. Influence (%)
NULL	-	-	1015.00	144.08	
year	7.00	10.81	1008.00	133.27	4.27
season	3.00	2.19	1005.00	131.08	2.02
target	3.00	7.15	1002.00	123.93	6.59
bait kind	4.00	4.10	998.00	119.83	2.84
ln(depth)	1.00	0.01	997.00	119.82	0.02
ln(NPP)	2.27	0.79	994.73	119.03	0.97
ln(area)	0.63	1.44	994.10	117.59	6.35
Ln(minHD)	2.50	1.46	991.60	116.13	1.62
ln(maxHD)	0.80	0.04	990.81	116.09	0.14
ln(num. fl.)	0.96	0.00	989.84	116.09	-0.01
ln(num.light)	0.98	0.10	988.86	115.99	0.28
ln(bait wght.)	6.54	2.34	982.32	113.65	0.99
X-cent.	3.64	1.73	978.68	111.92	1.32
Y-cent.	7.26	3.97	971.42	107.95	1.51

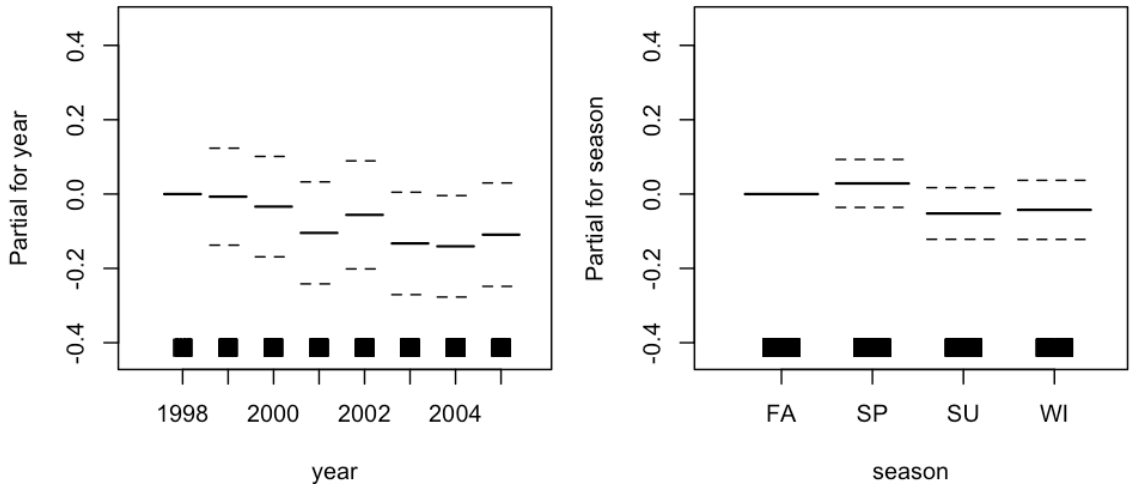


Figure 11. GAM: Partial residual plots of model terms; cubic spline fitted functions with 95% confidence intervals (shaded), continued below

figure continued

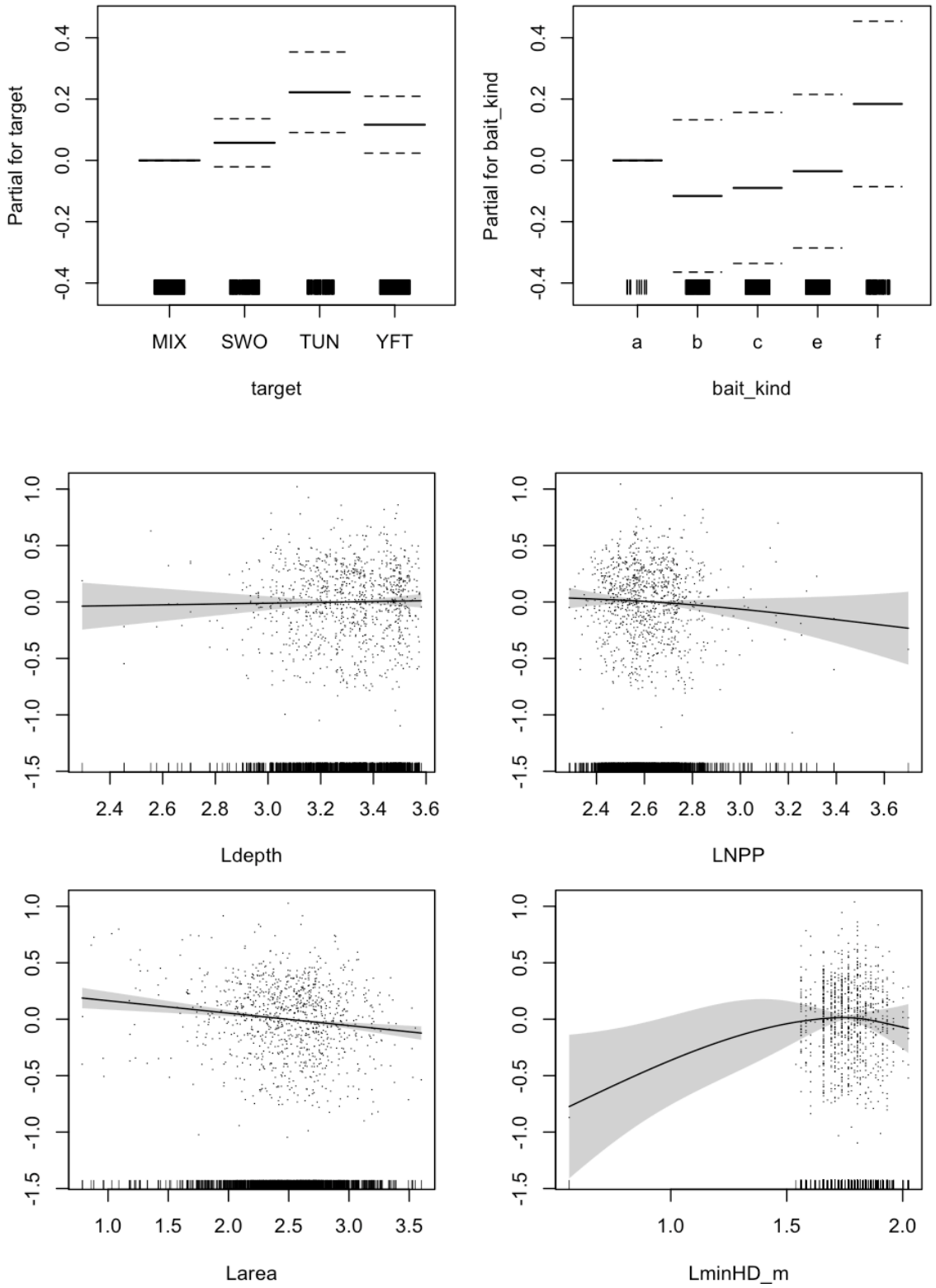
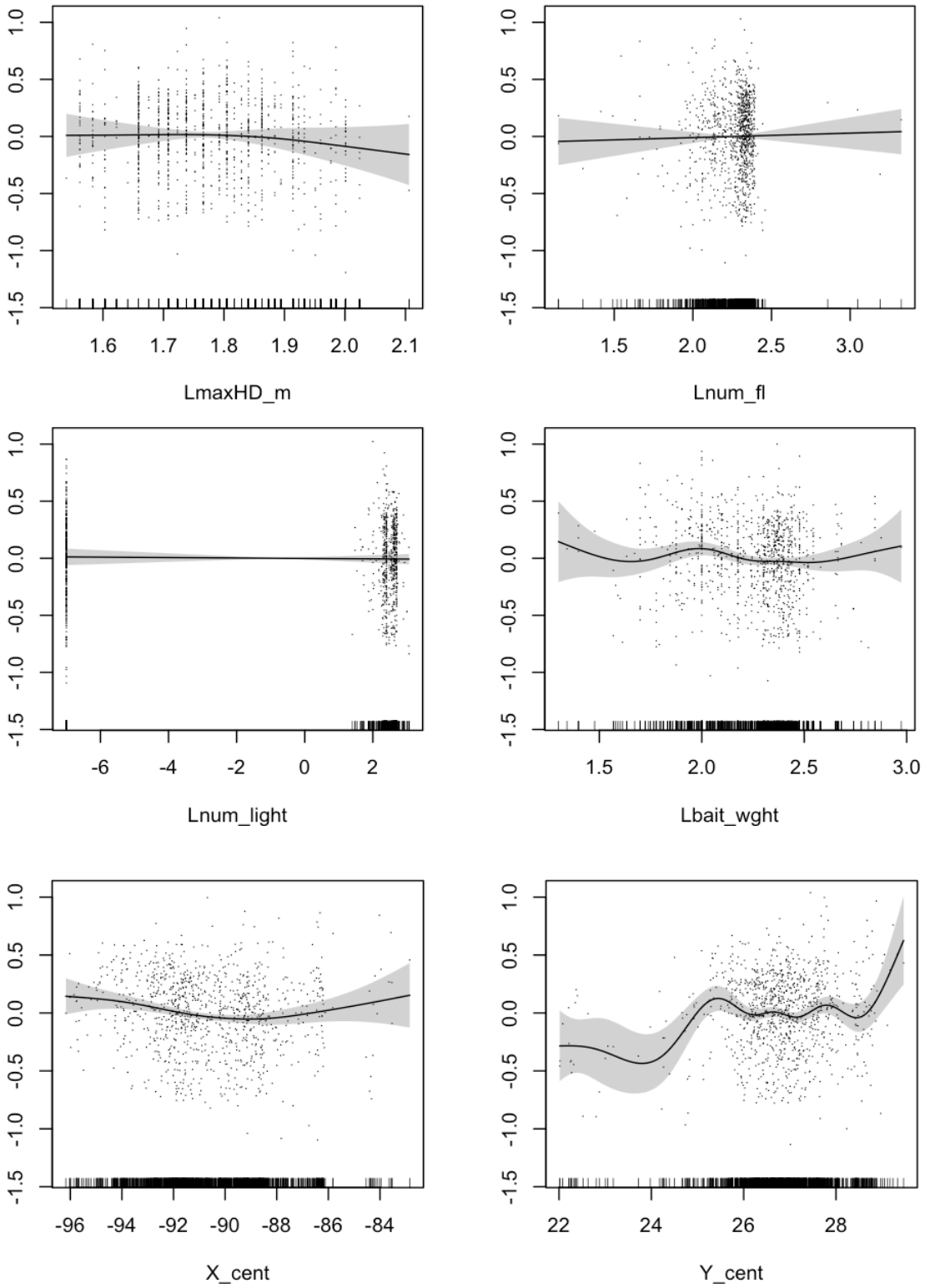


figure continued



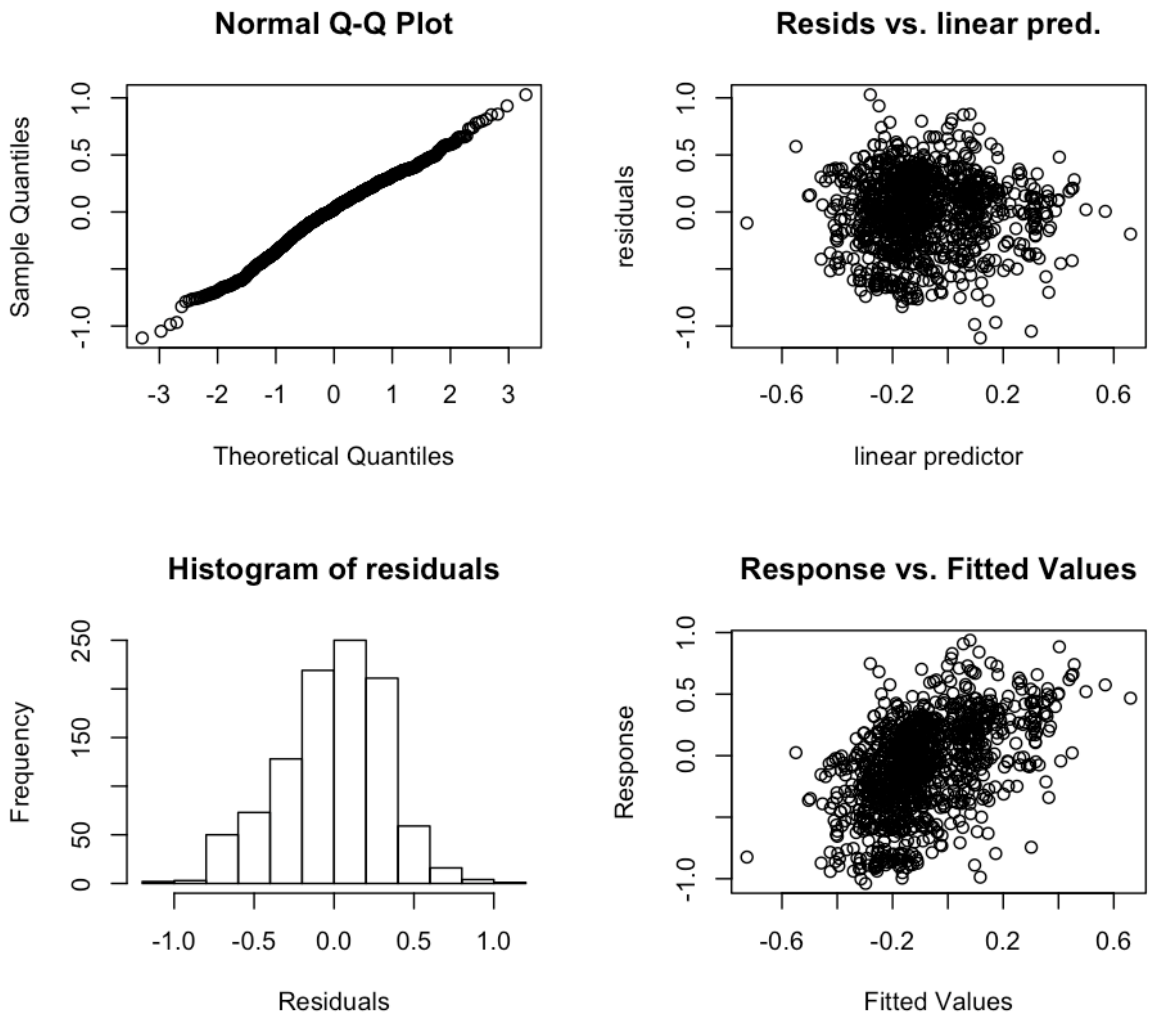


Figure 12. GAM: Diagnostic plots

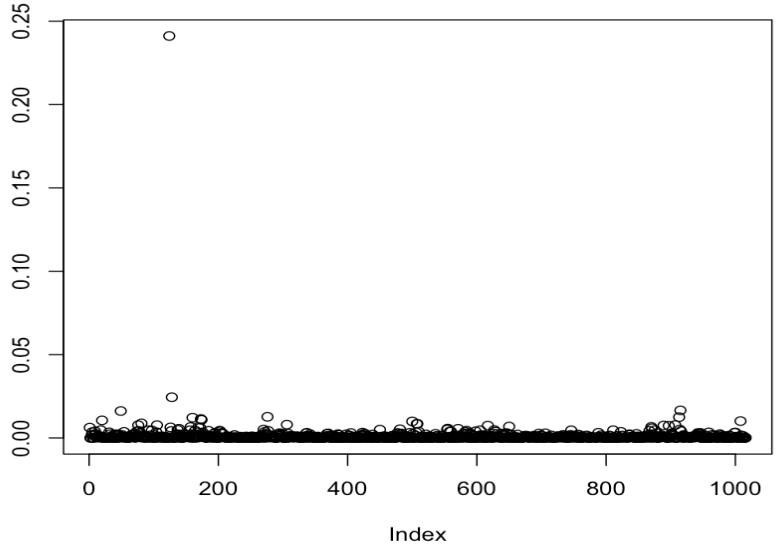


Figure 13. GAM: Cook's distance plot; observation #124, $D_i = 0.24$

Boosted Regression Tree (BRT): Main Effects Model

For the main-effects BRT model (*i.e.* interaction depth=1), where learning rate=0.01 and bag fraction=0.5, the optimal number of trees was reached at trees=1350 (Figure 14). The BRT final model accounted for 26.10% of the mean total deviance (1-mean residual deviance / mean total deviance=1-(0.105 / 0.142)=0.261).

In GLMs and GAMs, the standard output includes component-wise p-values for each of the predictors in the model, which serve to evaluate the contribution of each term in reducing overall model deviance and form the basis upon which variables may be eliminated. For BRTs, however, term-wise p-values are not so easily determined. Instead, an index of relative influence

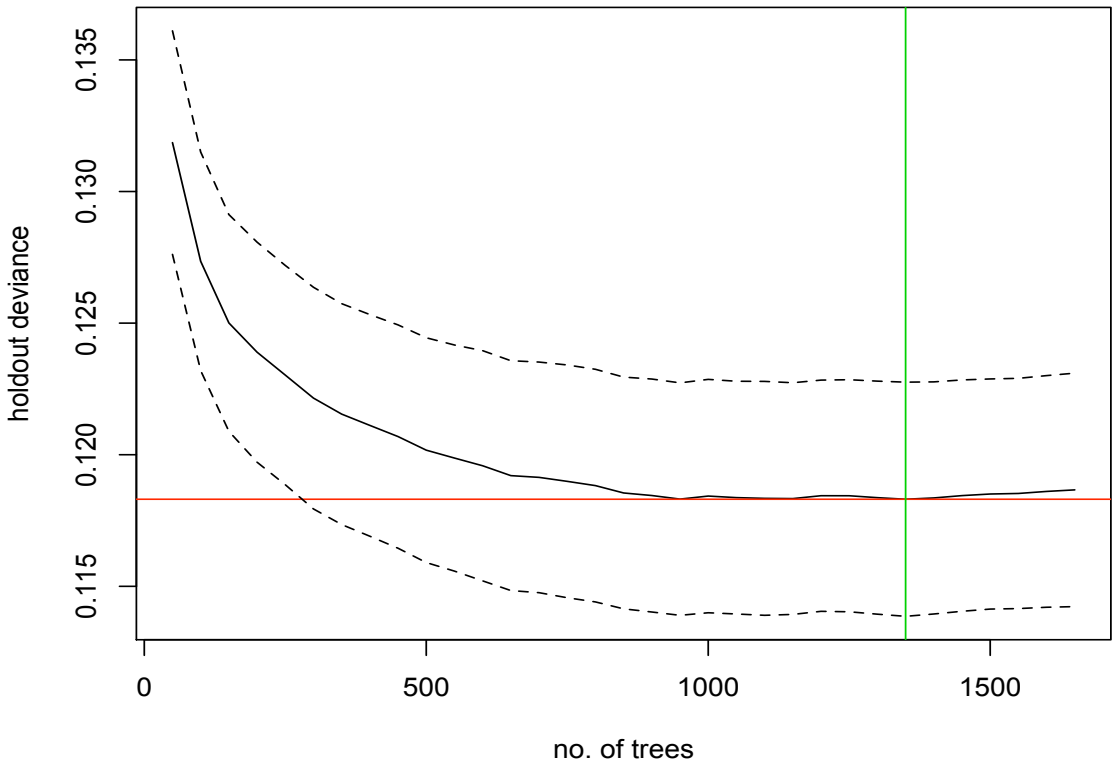


Figure 14. BRT: Optimization plot for the BRT main effects model; minimization of model deviance with the stage-wise addition of trees

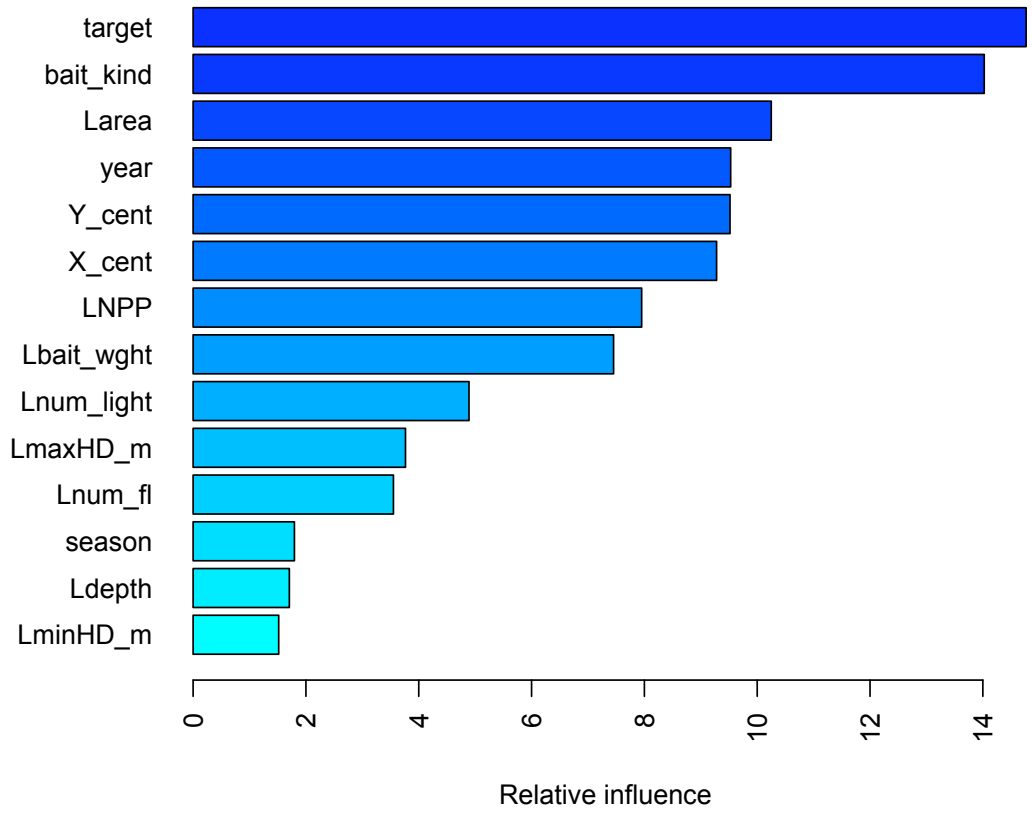


Figure 15. BRT: Relative influence of model terms calculated by the contribution of each term in reducing overall model deviance

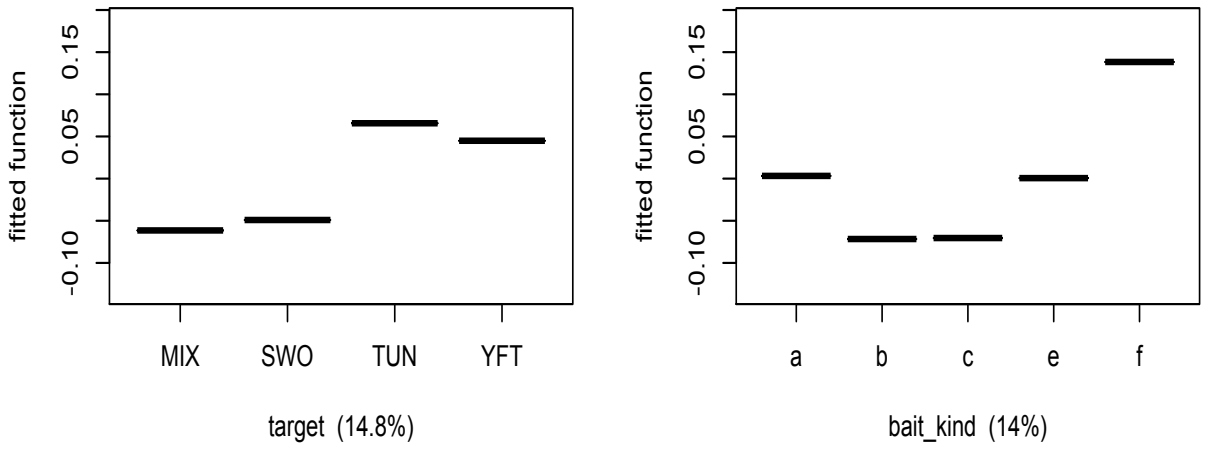


Figure 16. BRT: Fitted functions for each term in the BRT main effects model ordered by relative influence value, in parentheses, continued below

figure continued

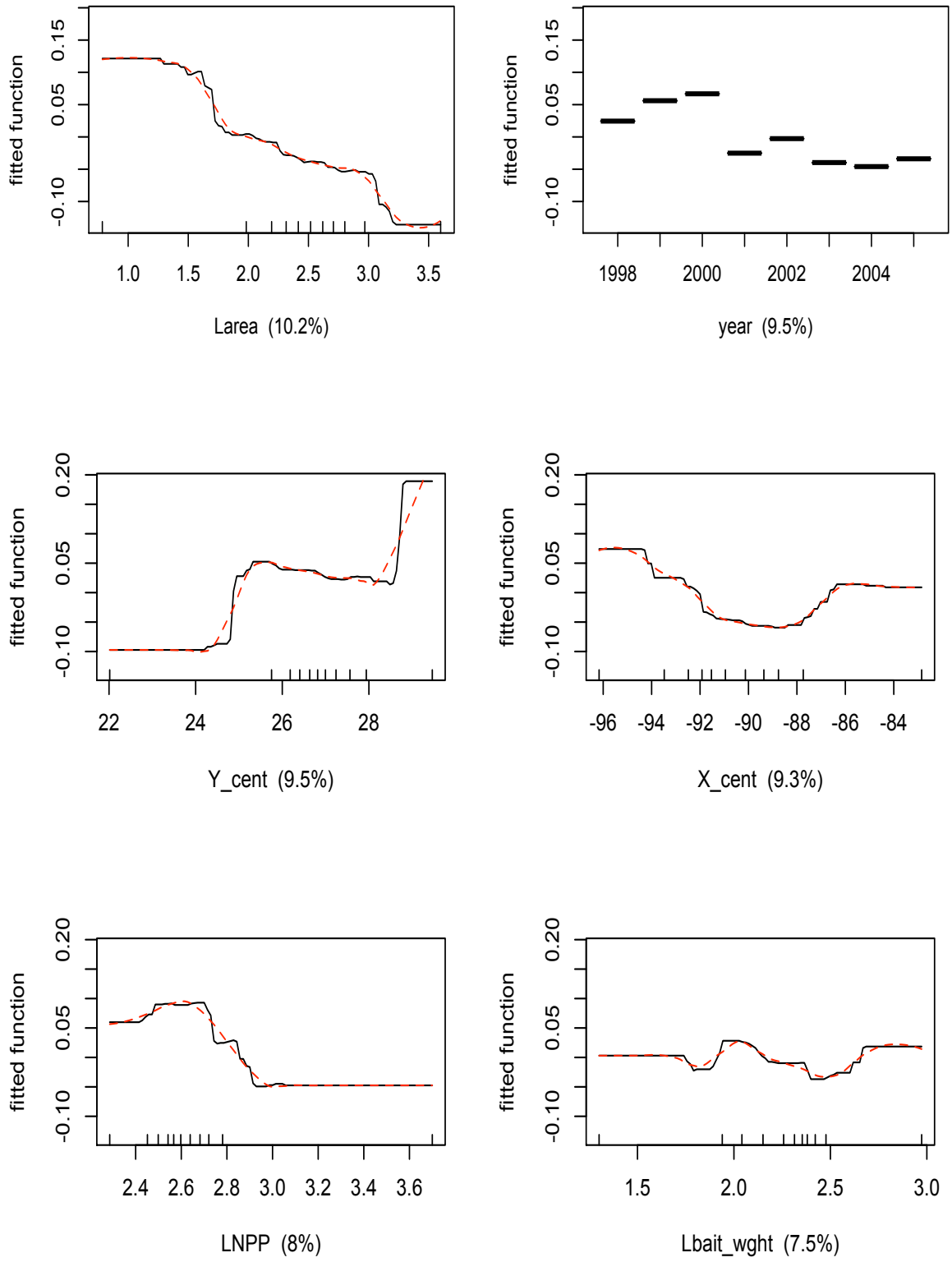
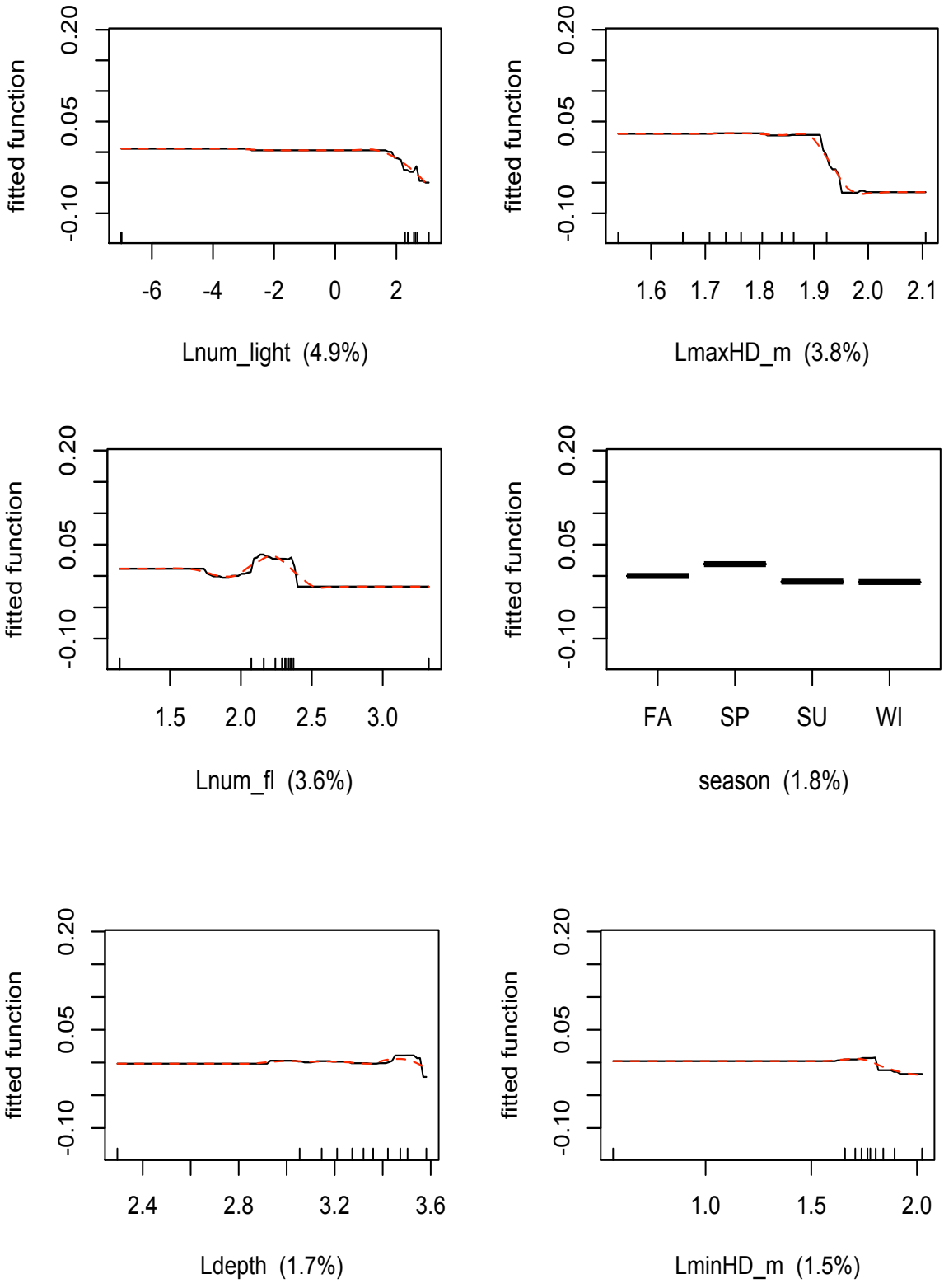


figure continued



may be calculated in summing the contribution of each variable, which is equivalent to summing the branch length for each variable in the regression tree. For the main effects BRT model fitted here, the five most influential variables were: 1) target (14.77%), 2) bait kind (14.02%), 3) Larea (10.25%), 4) year (9.53%) and 5) Y-cent (9.52) (Figure 15).

The information normally gained from examination of the partial residual plots from a GLM or GAM analysis, in the case of BRT, is obtained from term-wise plots of fitted functions *versus* observed values (Figures 16) and predicted *versus* observed values (See Appendix, A1.1-A1.3). Clearly, the form of the response functions fitted by the BRT model used here differed, from a greater to lesser degree, to those fitted by the GLM or GAM models. Examples of these differences are more apparent for those response functions that are fitted with a linear relationship by one method and not by another. For example, results of the GAM indicate a linear relationship for bottom depth (Ldepth) and primary productivity (LNPP), however, results of the BRT model indicate a non-linear relationship. Conversely, the function fitted for the minimum hook depth predictor (LminHD_m) in the BRT model appears linear, yet in the GAM a curvilinear function was fitted.

Boosted Regression Tree: Two-way Interaction Model

The loss function for the two-way interaction model (interaction depth=2) was minimized at trees=1350, the same number of trees required for the main effects model (Figure 17). The final two-way, BRT model accounted for 37.30% of the mean total deviance (1-mean unexplained deviance / mean total deviance=0.089 / 0.142=0.373). The index of relative predictor influence for the two-way BRT illustrates how the inclusion of interactions can significantly change the contributions of the terms in the final model, in particular spatiotemporal variables. Intuitively, time-space interacts with all things, thus, to some extent, with all predictors in a standardization model. Therefore, in order to adequately "capture" the

contribution of spatiotemporal predictors, a minimum of two-way interactions should be included in all standardization models, especially given the importance placed on the year term. Rankings for the five most influential variables in the two-way model examined here were: 1) year (11.23%), 2) bait_wght (11.16%), 3) Y-cent (10.28%), 4) bait kind (9.96%) and 5) target (9.79%) (Figure 18). As can be seen in the table of ranked interaction sizes, or strengths, the spatiotemporal predictors take part in seven of the ten most influential interactions, accounting for the rise of "year" and "Y-cent" in the ranks of relative influence, as compared to their positions in the main-effects model (Table 6). Interestingly, the bait weight term (Lbait wght) proves to be an important "interactor", increasing in rank from eight to two in relative influence from the BRT main-effects model to the BRT two-way interaction model, respectively. The response surface for the X-cent—Lbait wght cross term, the strongest of the two-way interactions, indicates that the $CPUE_{yft}$ for a particular geographic region was responsive to increases in the weight of the bait used by longliners (Figure 19).

The fitted functions for the two-way model present the same general trends as those observed in the main-effects model. Nonetheless, the presence of interactions does affect to some extent the form of each of the fitted response functions, in reducing the overall smoothed appearance. Differences between the two fits are most notable for the Y-cent predictor, where changes in slope throughout the data range are much sharper for the interaction model than for the main-effect model (Figure 20). For fitted *versus* observed plots, see appendix (A2.1-A2.3).

Residual Analysis

Summaries of model deviances were calculated in order to characterize their empirical distributions, results are presented in Table 7. It is worth noting from the data summary that means calculated for the GLM and GAM models (2.88E-16 and 2.48E-13, respectively) were substantially less than those for the two BRT-based models, 3.79E-04 and 1.35E-04 for the

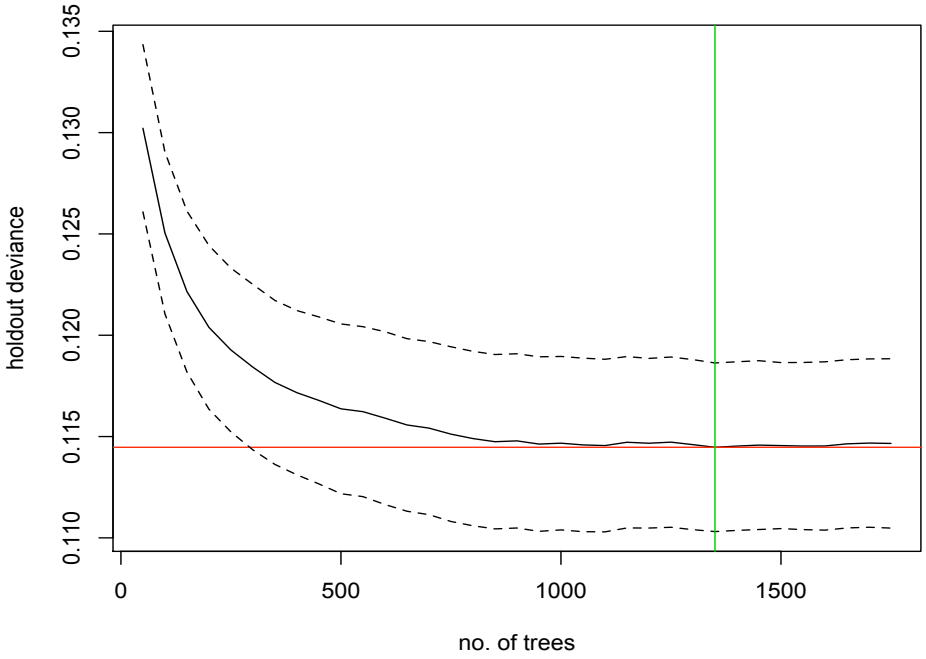


Figure 17. BRT 2-way: Optimization plot for the BRT two-way model; minimization of model deviance with the stage-wise addition of trees

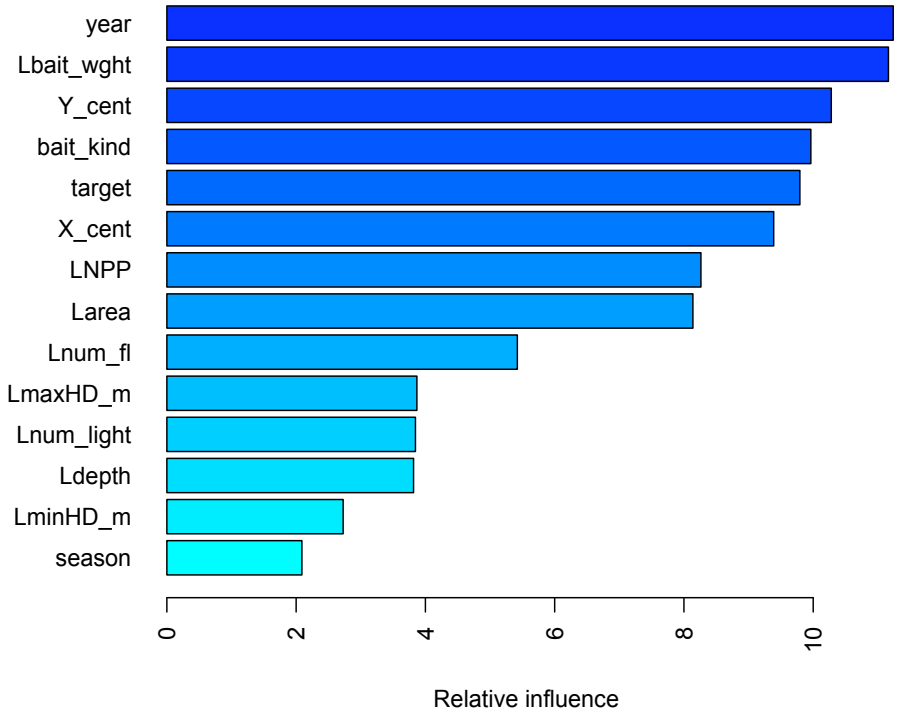


Figure 18. BRT 2-way: Relative influence of model terms calculated by the contribution of each term in reducing overall model deviance

Table 6. BRT 2-way: Ranked interaction sizes of the ten most important pair-wise interactions

Rank	Variable 1	Variable 2	Interaction size
1	X-cent	Lbait wght	0.48
2	Lbait wght	Year	0.25
3	Lnum fl	LNPP	0.12
4	X-cent	LNPP	0.10
5	Larea	Year	0.10
6	Y-cent	Year	0.05
7	LNPP	Year	0.05
8	bait kind	Target	0.05
9	Y-cent	Lbait wght	0.04
10	Lbait wght	LmaxHD_m	0.04

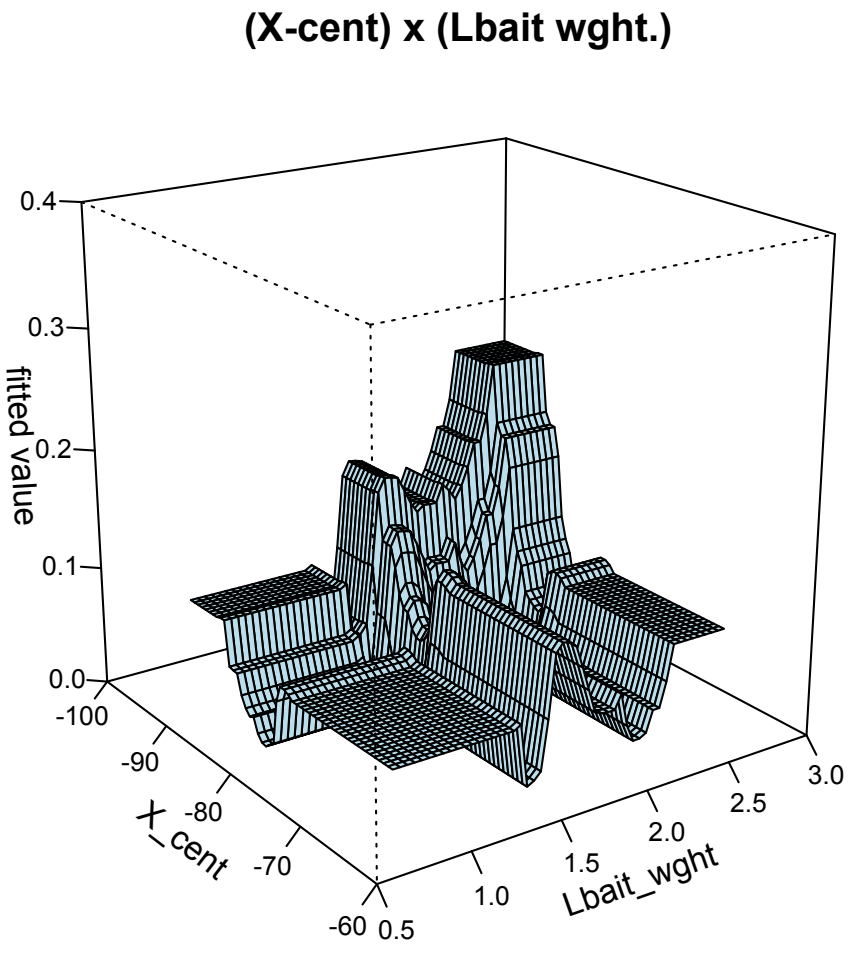


Figure 19. BRT 2-way: Interaction plot for the X-cent x Lbait wght cross-term; the most important interaction in the BRT 2-way model; interaction size= 0.48

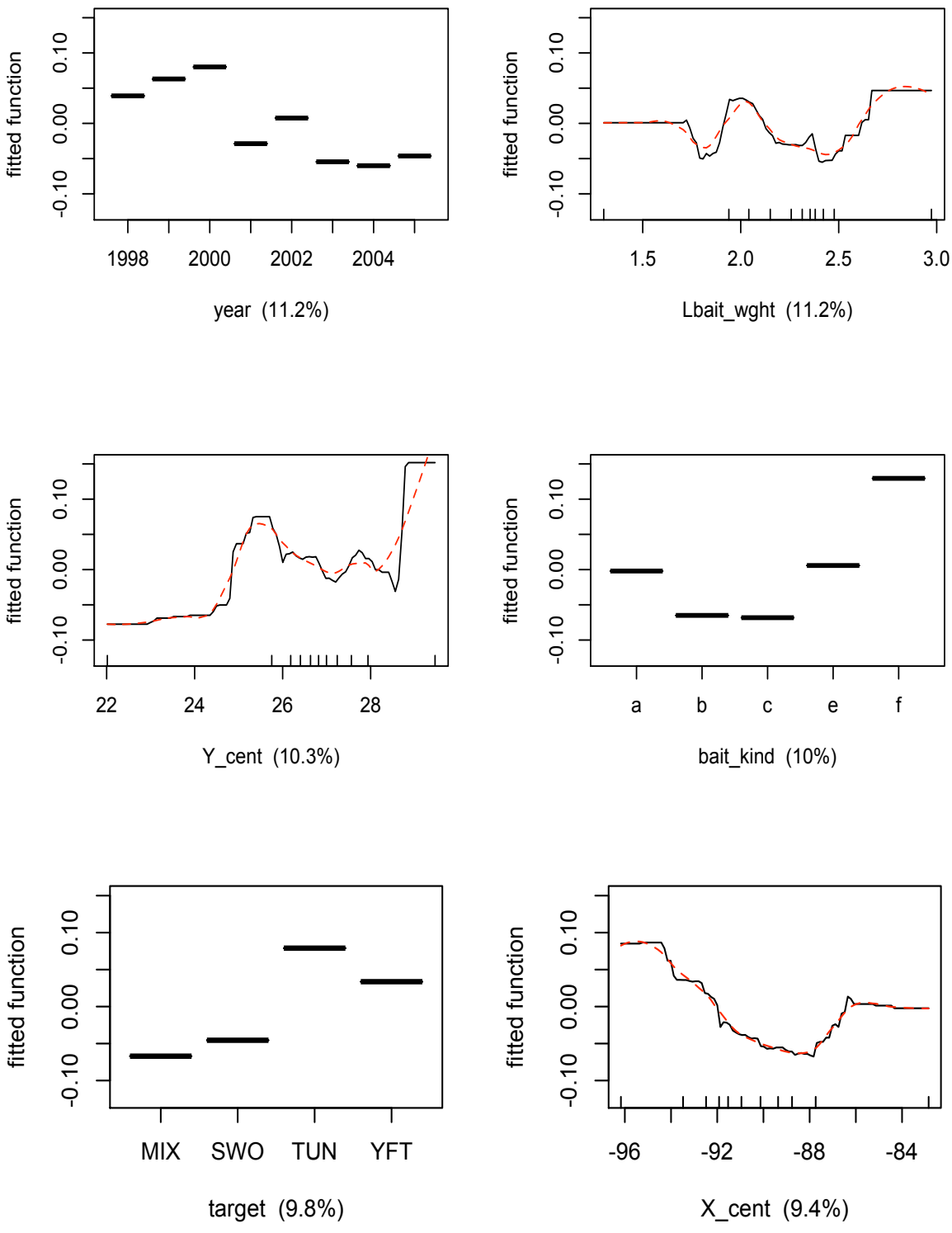


Figure 20. BRT 2-way: Fitted functions for each term in the 2-way factorial BRT model; in descending order of relative influence, in parentheses, continued below

figure continued

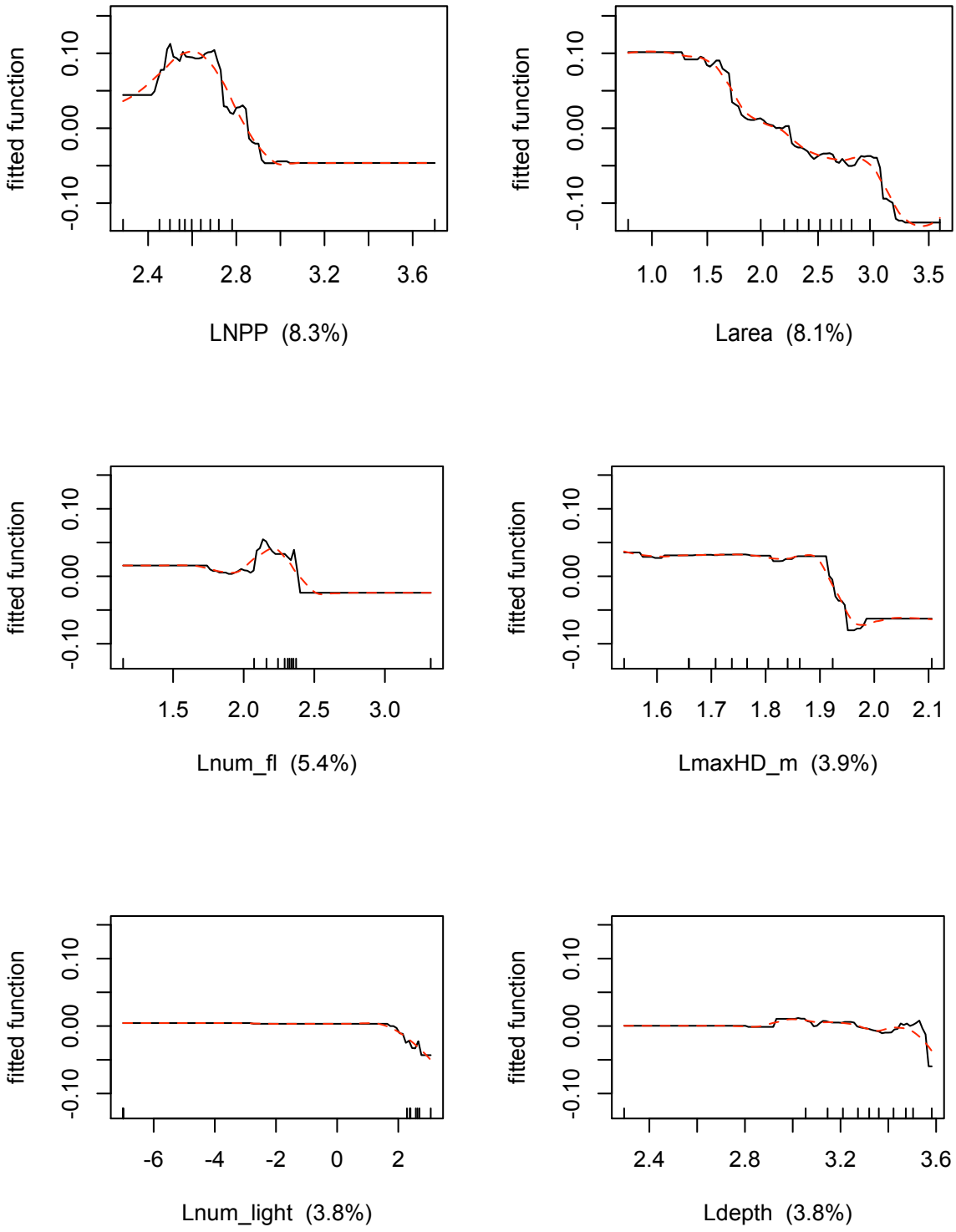


figure continued

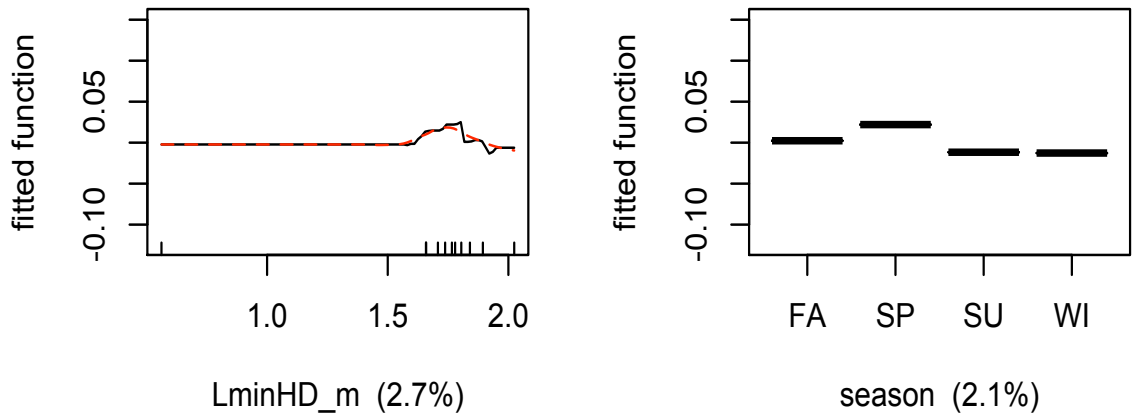


Table 7. Summary of deviances from the four modeling techniques

Model	Mean	S^2	Min.	25% Q	Median	75% Q	Max.	Σ |deviance|
GLM	2.88E-16	0.34	-1.12	-0.22	0.03	0.24	1.01	273.69
GAM	2.48E-13	0.33	-1.05	-0.20	0.02	0.23	1.03	263.32
BRT	3.79E-04	0.32	-1.39	-0.22	0.03	0.24	0.95	264.11
BRT 2-way	1.35E-04	0.30	-1.07	-0.19	0.03	0.21	0.92	242.20

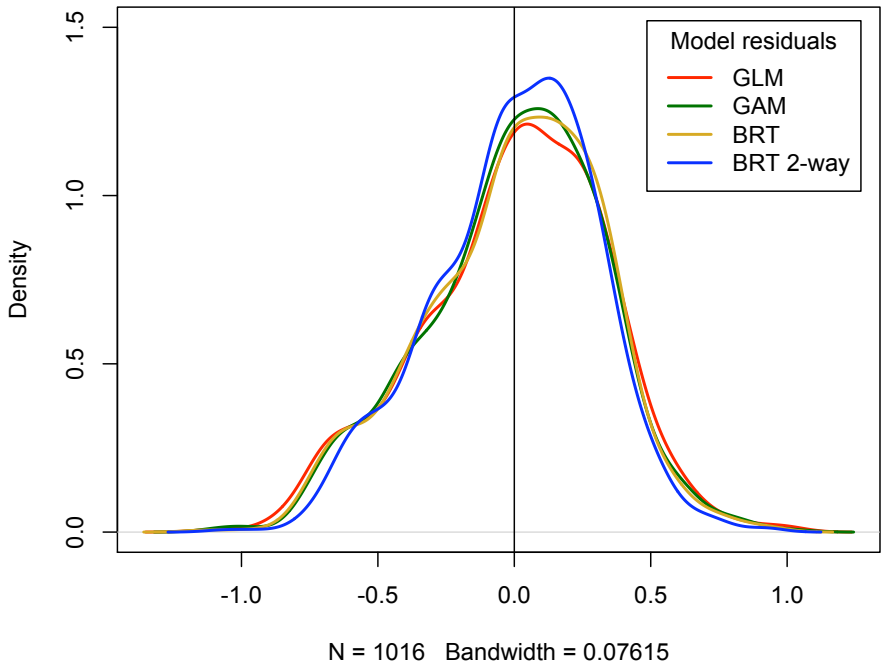


Figure 21. Density functions for the GLM, GAM, BRT and BRT 2-way model residuals; values calculated using the default Gaussian kernel

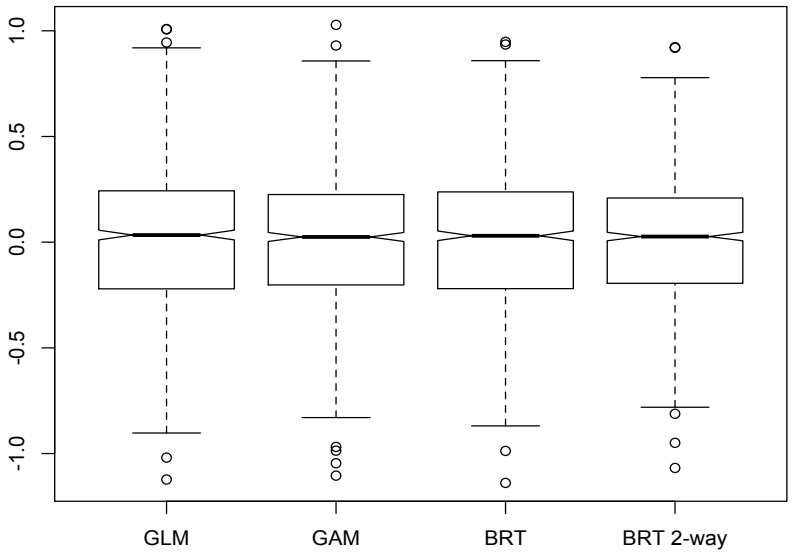


Figure 22. Boxplots of the residuals for the four modeling methods; notches represent a robust estimate of the medians

BRT and BRT 2-way, respectively. While the means were less for the GLM and GAM deviances, sample variances and maximum values for the BRT methods were lower. However, the sum of absolute deviances of the GAM is actually less than that for the BRT. These differences in the empirical distributions of the model deviances were visualized using a density plot and a boxplot (Figures 21 and 22, respectively).

The density plot shows that, indeed, all the distributions are quite similar for all the main-effects models, namely the GLM, GAM and BRT. The most notable difference is between the main-effects models and the BRT 2-way model. The density of the BRT 2-way residuals is slightly more peaked, with more density at the center of the distribution and less in the tails, hence the lower value for the variance (Figure 21, Table 7). The boxplots illustrate more clearly the slight changes in variance resulting from the four models (Figure 22). The overlapping notches of the boxplot indicate that no significant differences exist between any of the residual groups. The matrix plot of correlations (Figure 23) illustrates the general similarity of the different residual groups.

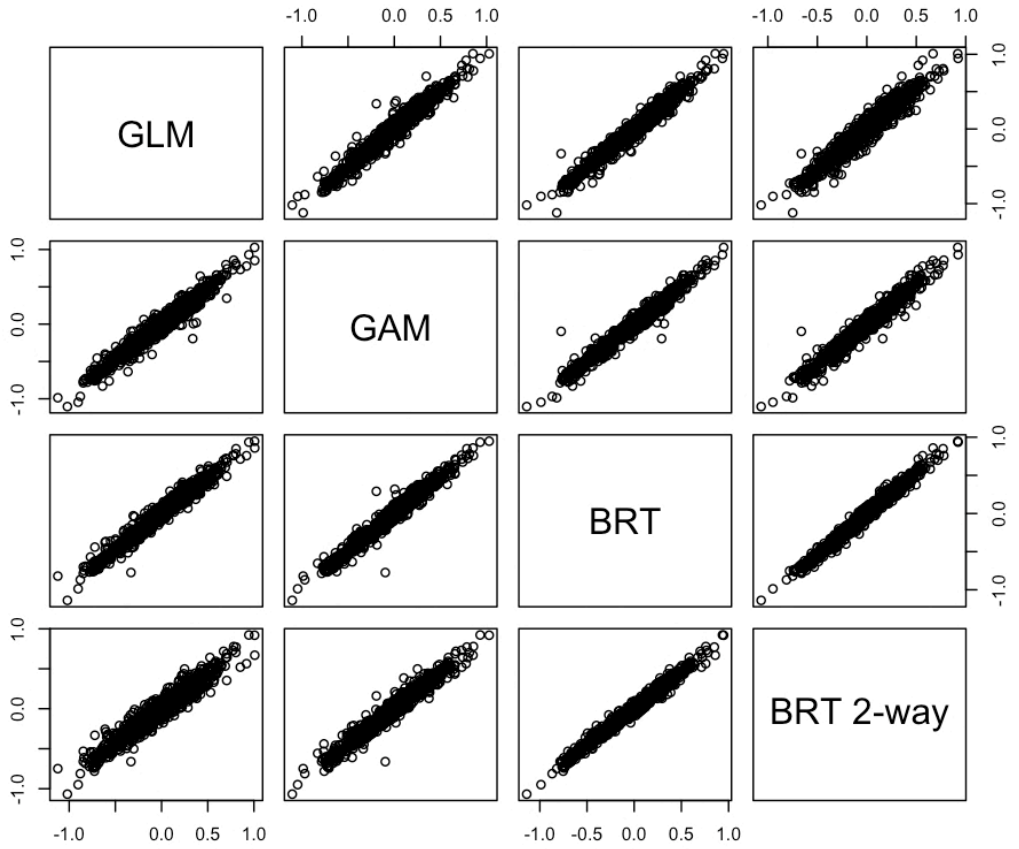


Figure 23. Correlation matrix of the residuals for the four modeling techniques, illustrating similarities/dissimilarities between models

For the final analysis of the model residuals, semivariogram plots were used to test this assumption. In Figure 24, the omnidirectional (black line with points) and four directional semivariograms of the raw catch rate data were plotted to test for spatial autocorrelation globally and along the four principal axes. From the plot, there is evidence of some degree of spatial dependence in first four lags of data. As none of the four directional semivariograms appear to contribute any additional information, the omnidirectional semivariogram was selected for the analysis of model residuals.

Empirical semivariogram plots of model residuals visually confirm that which was previously established, namely, the ability of the four models in reducing variance relative to the native variance (*i.e.* nominal $CPUE_{yft}$ variance), and also, relative to each other (Figure 25). In

Figure 25, the black line represents the semivariogram values for the nominal $CPUE_{yft}$, where the dashed line depicts the semivariogram sill whose value is equal to the overall sample variance ($s^2[CPUE_{yft}] = 0.114$). Thus, the sills for the residual semivariograms (not drawn here) would visually represent the overall ability of the four modeling techniques in reducing variance.

More importantly, however, are the relative abilities of these techniques in handling autocorrelated data. Although none of the regression models used here are structurally equipped to deal with autocovariance, as are spatial regression techniques, the level of autocorrelation in the model residuals does appear to decrease to some extent relative to native levels. For the GLM residuals, there is a slight decrease in the slope of the semivariogram line, and even more so for the semivariograms of the GAM and BRT residuals. The greatest decrease in the level of autocorrelation, evidenced by the flattening of the semivariogram line, results from the BRT 2-way model (Figure 25). Therefore, some proportion of the observed spatial autocorrelation is the result of interaction effects. These results indicate the importance of including interactions in catch rate standardization models not only for deviance reduction purposes, but also, for mitigating the adverse effects of modeling autocorrelated data. Figures 26, illustrate the spatial distribution of the Pearson residuals for the four models. The clustering of residuals illustrates the areas for which the models perform poorly, which is the geographic representation of the spatially autocorrelated residuals, or spatial non-stationarity (e.g. see boxed areas, Figure 26).

Predictive Performance

Predictions were made with each of the four modeling techniques using the test dataset (N=436 sets). The RMSE for the GLM, GAM, BRT and BRT 2-way models were 0.3552, 0.3554, 0.3546 and 0.3509, respectively (Figure 27). Using the RMSE of the GLM model as the baseline, the percent change in RMSE for each of models were +0.06%, -0.17% and -1.21% for the GAM, BRT and BRT 2-way, respectively. Interestingly, the RMSE for the GLM model is

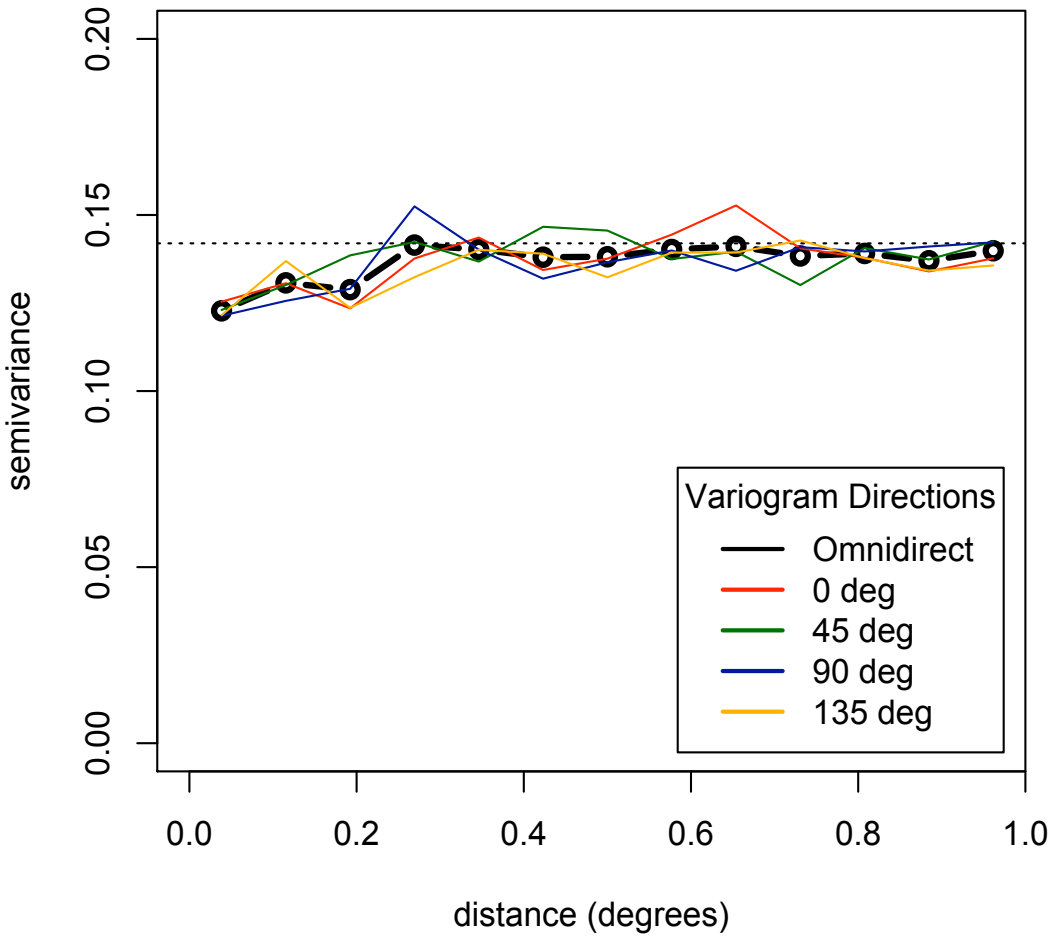


Figure 24. Omnidirectional and directional empirical semivariograms of nominal log(CPUE) data; dashed line represents the semivariogram sill, which is equal to sample variance (s^2=0.142); x-axis= lags measured as Euclidean distance between latitude / longitude coordinates

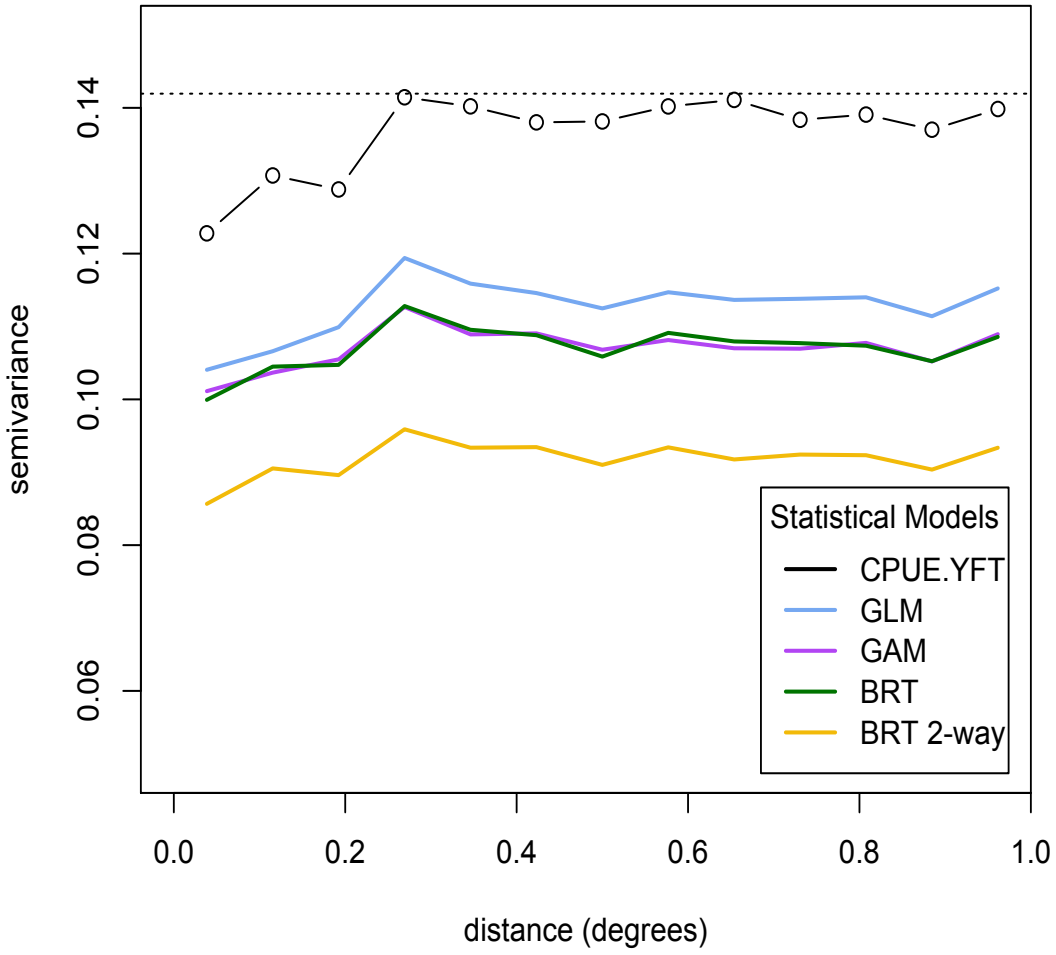


Figure 25. Ominidirectional, semivariograms of model residuals for the four modeling techniques, as compared to the empirical semivariogram of the nominal log(CPUE) (black); dashed line represents the variogram sill = 0.142 = sample variance (s^2); x-axis= lags measured as Euclidean distance between latitude / longitude x,y-coordinates; Note: scale of y-axis changed from Figure 24 to highlight differences

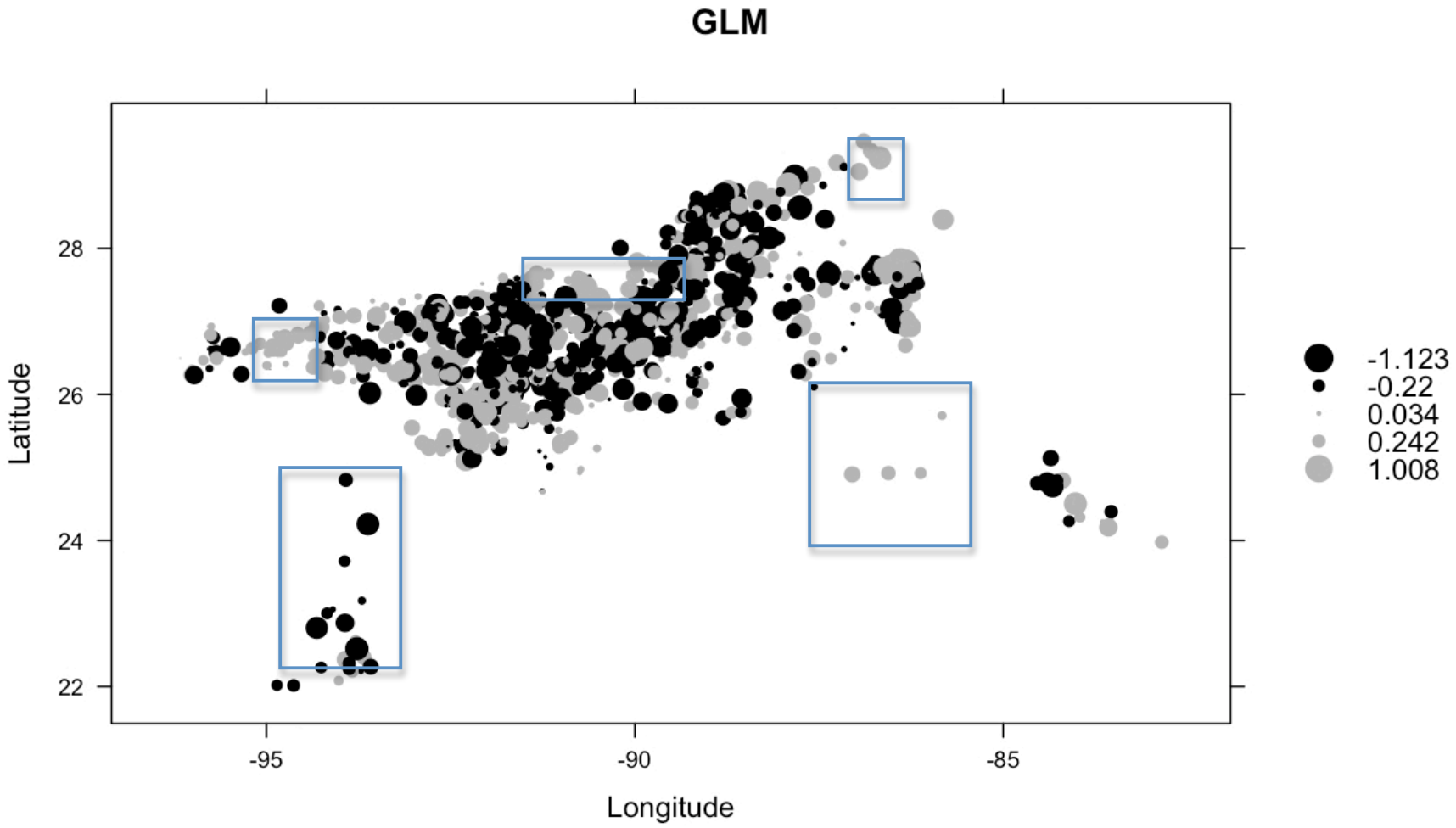


Figure 26. Spatial distribution of Pearson residuals for the four fitted models; points scaled to the minimum, 25% Q, 50% Q, 75% Q and maximum values; blue boxes highlight areas of poor model performance, continued below

figure continued

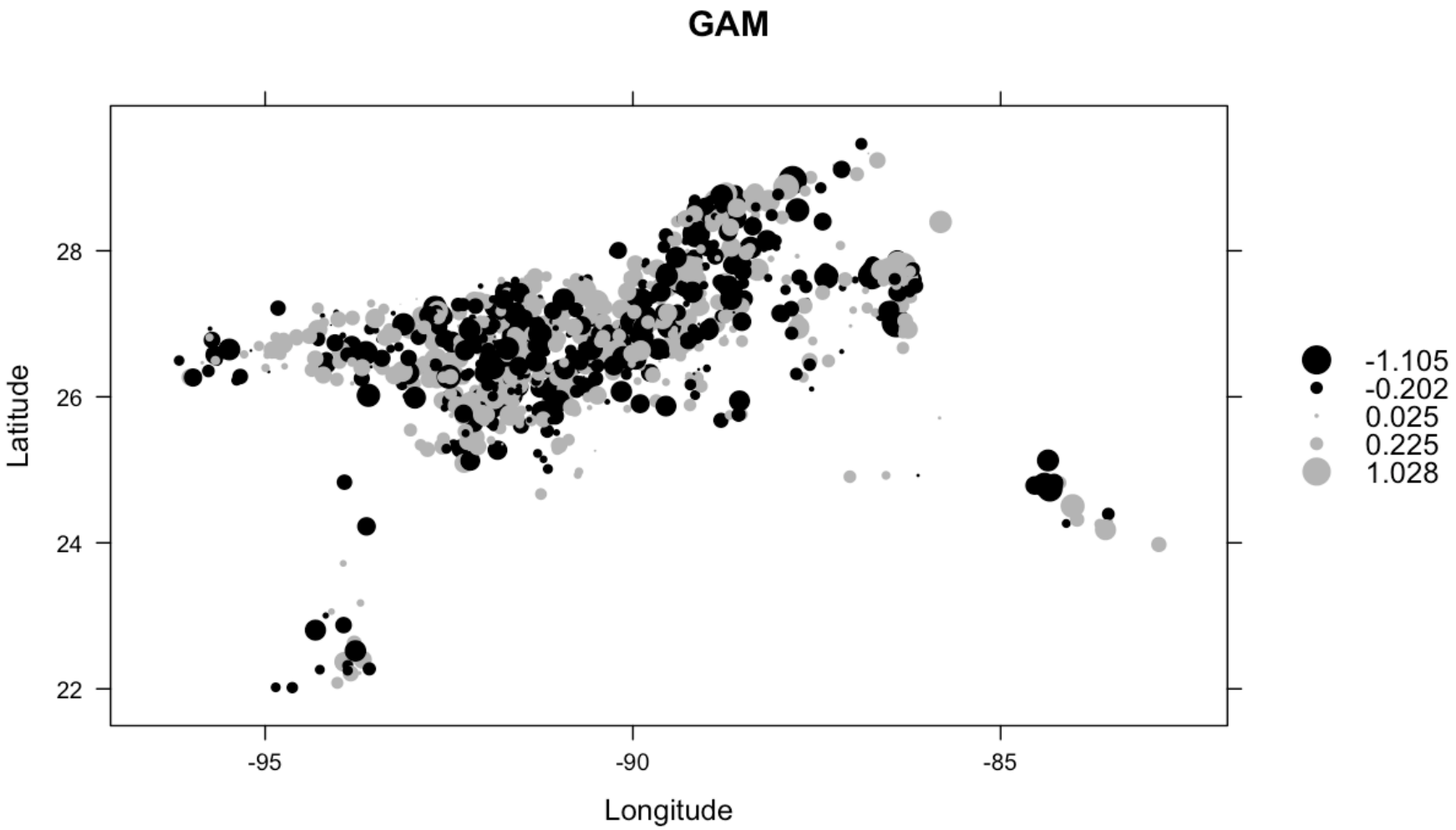


figure continued

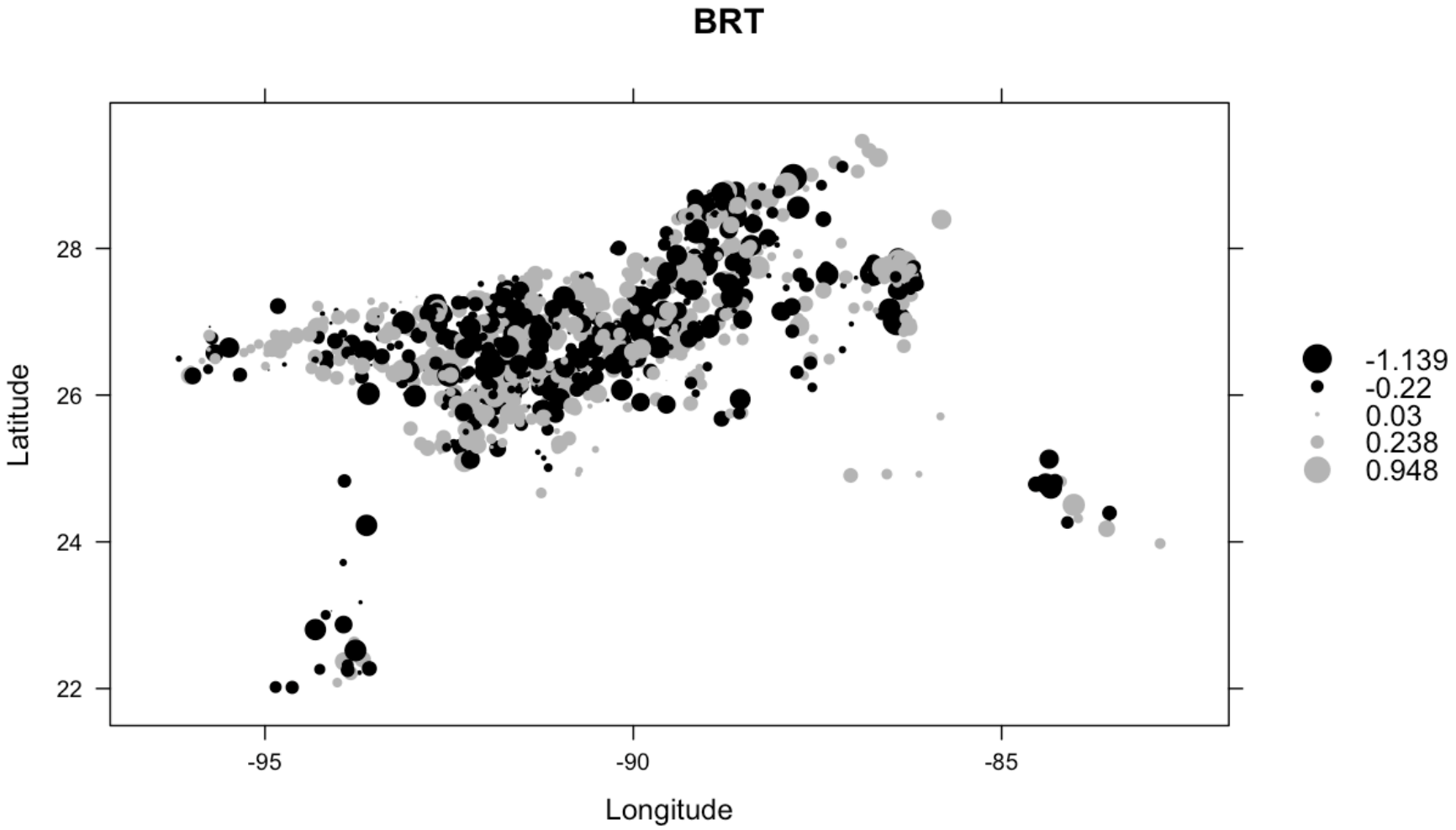
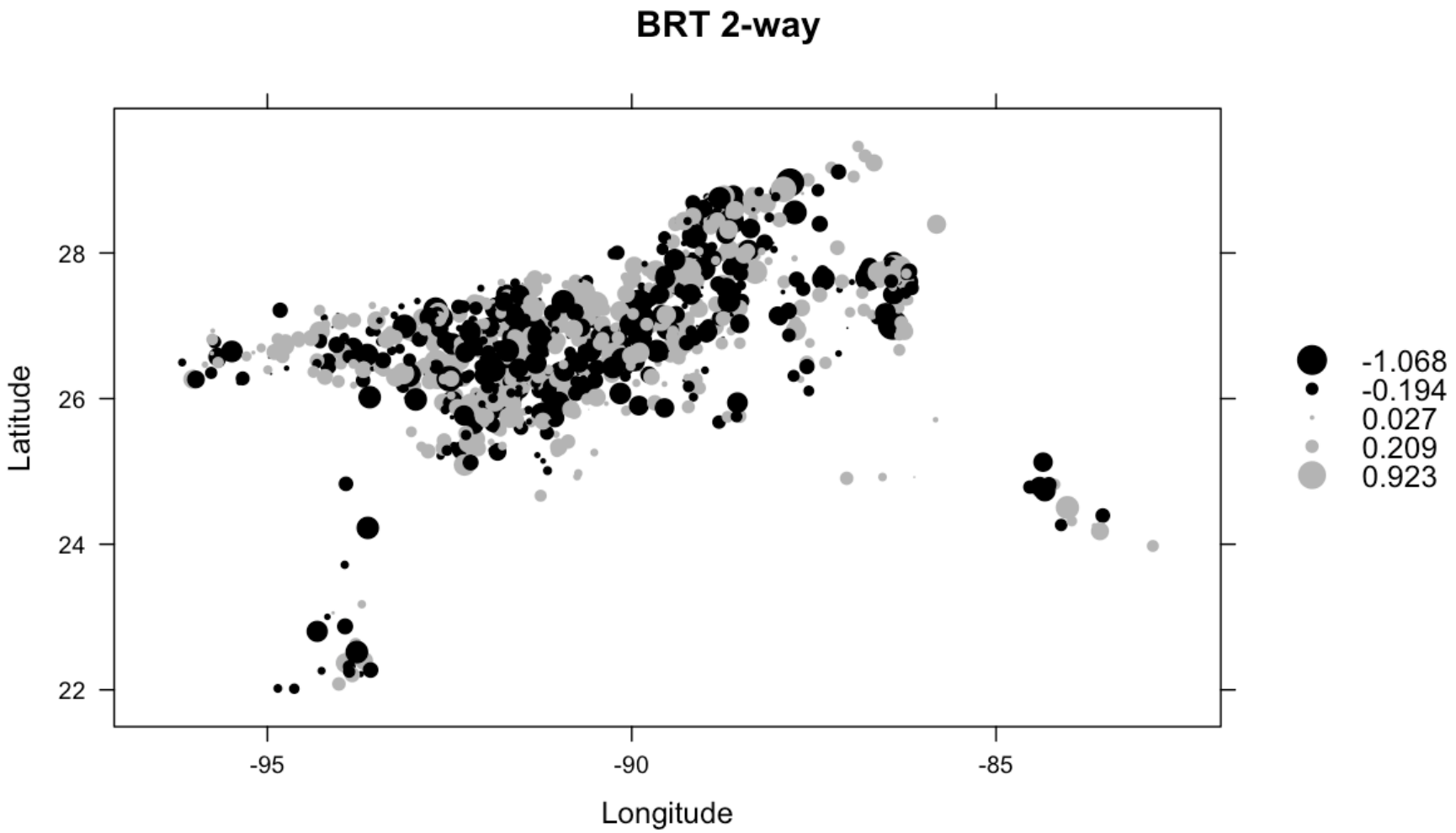


figure continued



lower than that for the GAM, despite the greater reduction in unexplained deviance by the GAM. Although the GAM explained a larger percentage of the model deviance, the variance of the predictions produced by the GAM were greater than that for the GLM (Table 8), indicating that the GAM model likely overfitted the data, hence the greater prediction error. Although differences in model performance for the three main effects models are only marginal, these results are indicative of the ability of the BRT model to reduce unexplained variance, as efficiently as the GAM model, yet without overfitting the data. The greatest improvement in prediction accuracy was achieved by the BRT 2-way model, which would be considered by many as highly overparameterized, yet demonstrated the best predictive success. Again, the stochastic nature of the boosting algorithm is able to avoid overfitting the data, thereby providing more accurate predictions. Paired t-tests performed on actual CPUE values *versus* predicted values, indicated that there were no significant differences. Results of F-tests found significant differences between BRT *versus* GAM ($p<0.001$) and GLM *versus* GAM ($p= 0.04$) prediction variance, whereas no difference was detected between BRT and GLM predictions ($p=0.16$). Figure 28 and Table 8 present summary statistics for model predictions.

As year-effect plots, or relative indices of abundance, are the final objective of the catch standardization process, the indices produced by the four statistical models are plotted, below (Figure 29). In general, the year-effects estimated for the four models are similar. Also, it is worth noting that the BRT model does not provide standard errors for the annual estimates of the year-effect, which hindered formal comparisons of the relative indices of abundance. However, some slight differences in the annual point estimates did exist. For example, the period of 1998-2000 illustrates the largest discrepancies between the fully-parameterized models (*i.e.* GLM and GAM) and the BRT-based models. Estimates then converge for the period of 2001-2003, and diverge again from 2003-2005, where the steeper slopes of GLM/GAM estimates appear more

optimistic for the final years of data. Globally, the trend observed in the relative indices of abundance, calculated here, are consistent with the general trend observed for the Gulf of Mexico indices used in the 2008 Atlantic yellowfin tuna stock assessment (Figure 30) (Brown & Ramirez-Lopez, 2009).

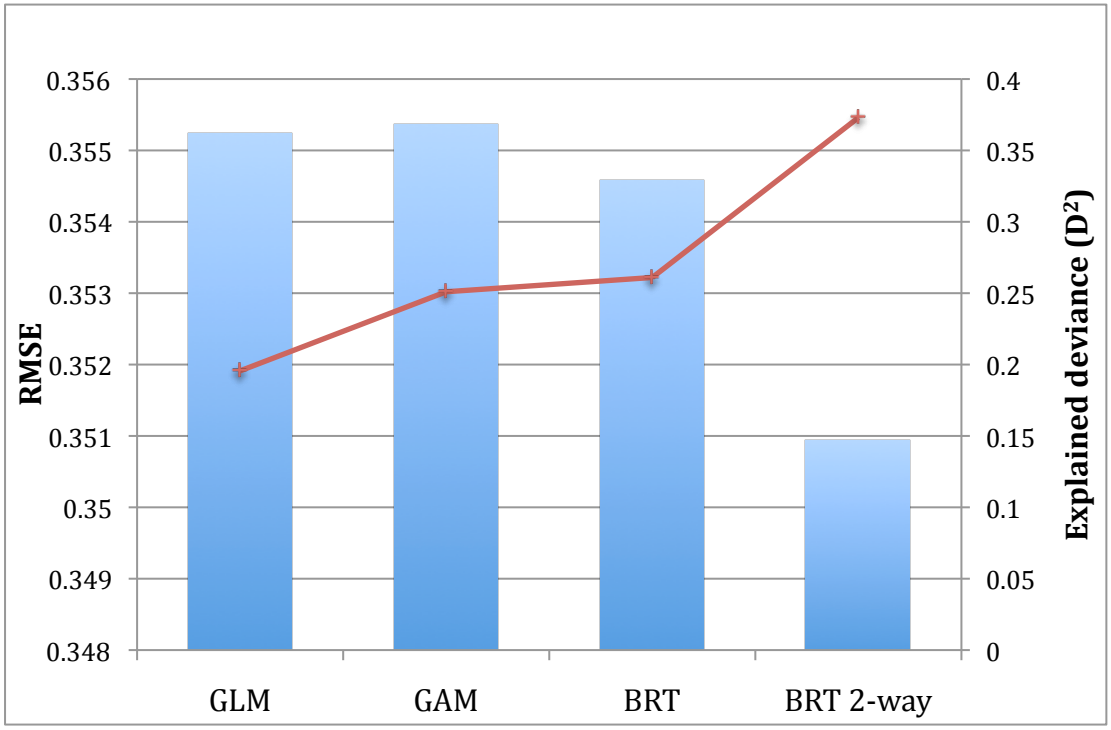


Figure 27. Calculations of RMSE (bars), using the test dataset, and percentage of explained deviance, D^2, (line) of the four trained models; note y-axis scale changed to emphasize differences

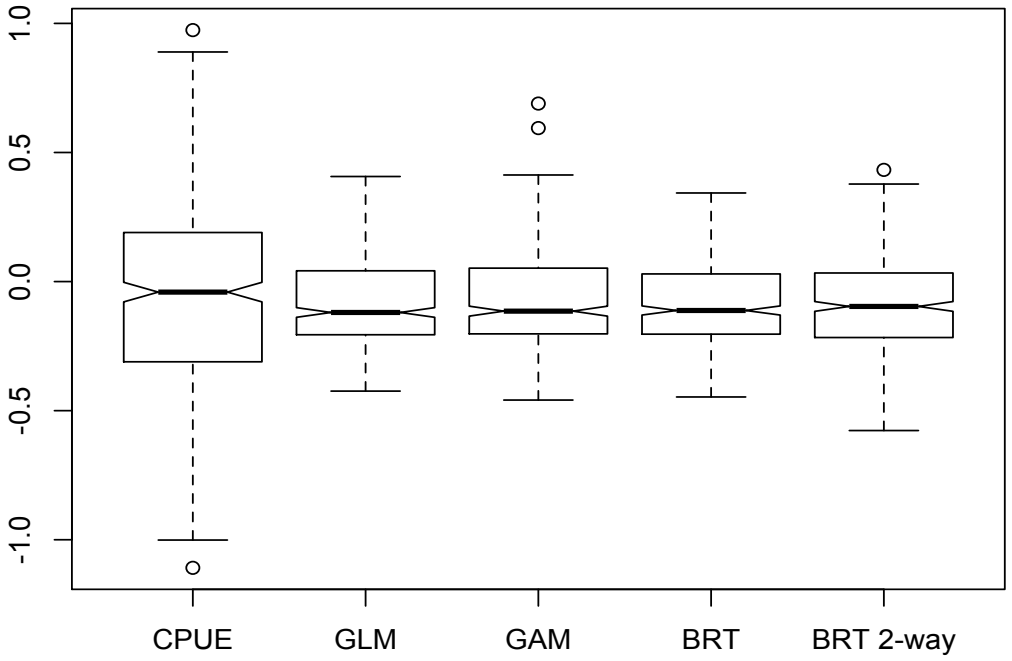


Figure 28. Boxplots of actual log(CPUE) data *versus* model predictions, using the test dataset; notches indicate robust estimates of the medians

Table 8. Summary statistics of actual log(CPUE) from test dataset and model predictions thereof: mean, variance, 25%, 50% and 75% quartiles

	mean	variance	25% Q	50% Q	75% Q
log(CPUE)	-0.075	0.151	-0.310	-0.041	0.190
GLM	-0.076	0.029	-0.210	-0.120	0.042
GAM	-0.074	0.035	-0.200	-0.115	0.052
BRT	-0.081	0.025	-0.200	-0.112	0.030
BRT 2-way	-0.084	0.031	-0.220	-0.096	0.033

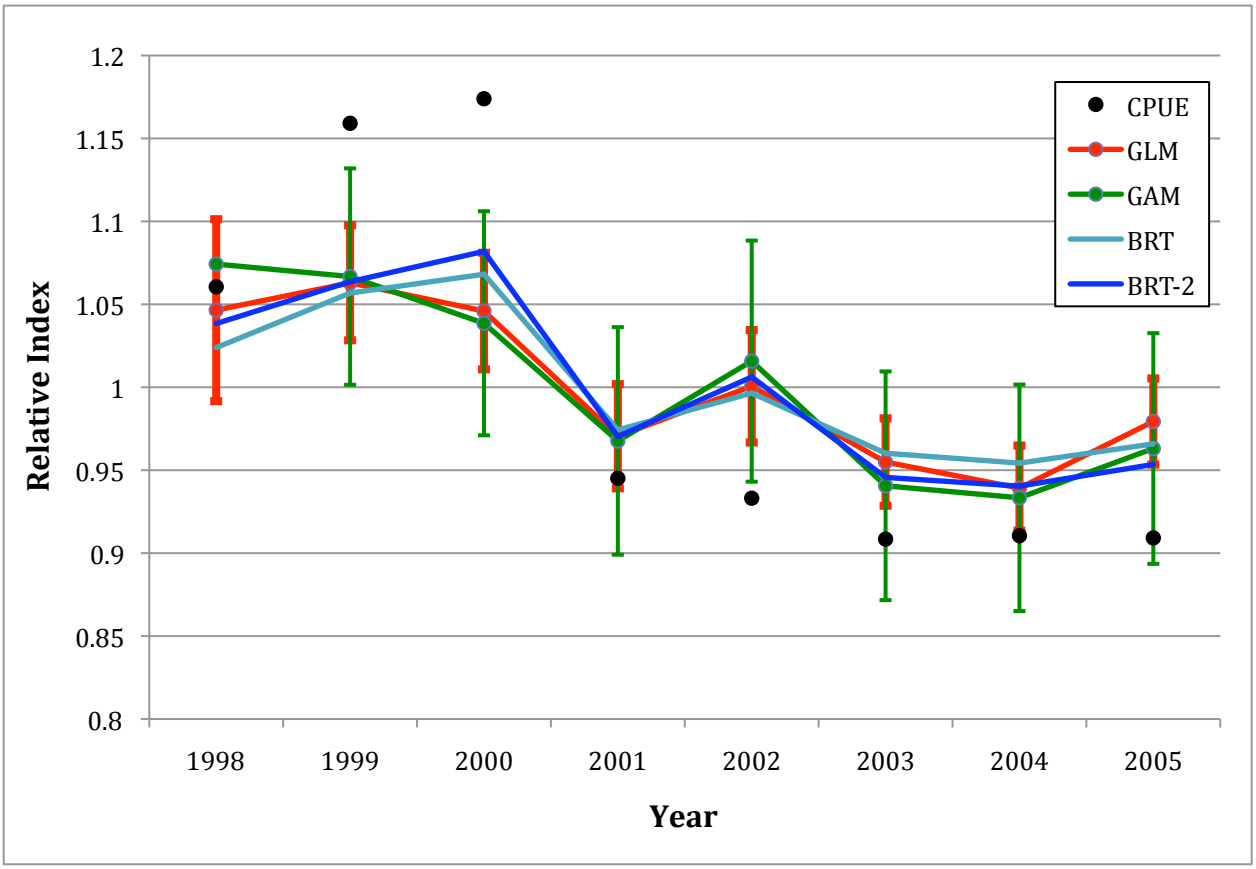
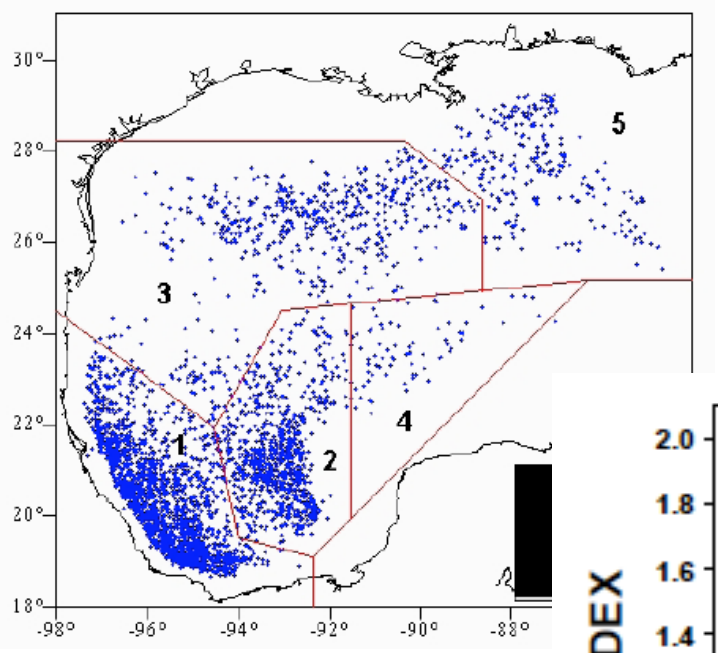


Figure 29. Comparisons of the standardized CPUE indices from the four trained models: GLM (+- SE), GAM (+- SE), BRT and BRT 2-way; estimates are centered to their mean



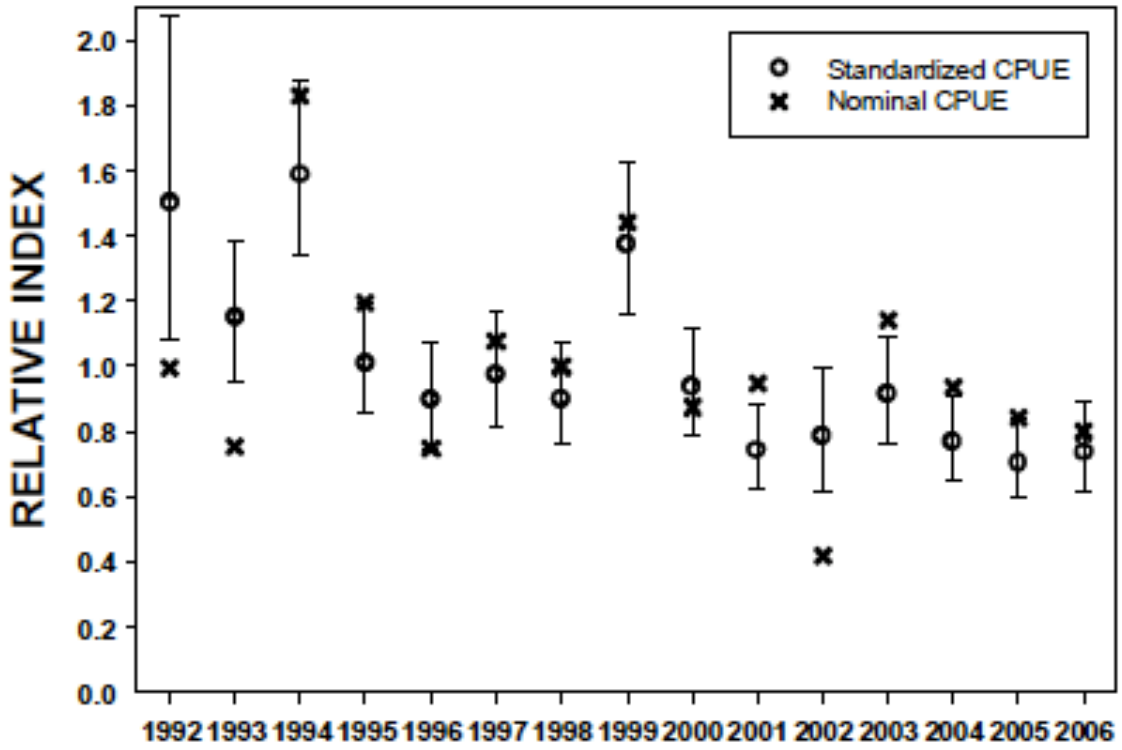


Figure 30. GOM longline observer data used (above) in the 2008 ICCAT Atlantic yellowfin tuna stock assessment; relative indices abundance with 95% confidence intervals (right), as determined by the GLM: Catch/set= year + mean temp. + setstart + quarter + zone + mean temp.*quarter, error distribution= Poisson, log-link; figures adapted from Brown & Ramirez-Lopez (2009)

Discussion

The present study adds to the growing body of literature investigating applications of newly developed techniques to the catch-rate standardization process. Many of these recent studies applied or developed novel methods aimed at resolving one of the many statistical issues encountered by fisheries scientists, such as zero-inflation, overdispersion or spatial autocorrelation. Here, performance of the boosted regression tree approach was evaluated, where advantages of this method include: 1) robust parameter estimation, using the stochastic gradient boosting algorithm to minimize variance and bias, 2) reduced risk of misspecification, as the model is learned form the data and 3) all the normal benefits of tree-based models in handling model predictors (*e.g.* automatic fitting of complex interactions and unaffected by: multicollinearity, missing predictor values and outliers).

Criteria used to evaluate performance depend upon the final objective, or use of the standardization model. If the statistical model will be used for prediction purposes, minimization of variance should be given priority. Usually, however, standardizations are performed to produce estimates, annual estimates of abundance, as input for a population model, therefore emphasis should be placed on minimizing bias (Bishop, 2006). In the present study, models were evaluated comparatively, with the newer BRT-based methods compared to traditional GLM and GAM approaches. A boxplot was used to visualize the resultant variance and bias of predictions made by the trained models on test data (Figure 28). Additionally, comparisons between the trained models focused on influential predictors, the role of two-way interactions and analyses of model residuals. All the analyses presented here, were performed in the R statistical software, and the script for these analyses is provided in appendix (Appendix II).

In modeling, the optimization of prediction accuracy, or minimization of prediction error, involves a theoretical trade-off between variance and bias. As model complexity increases,

model predictions become less biased, however, prediction variance increases (see Figure 31). Mathematically, the relationship between the prediction error, measured as mean squared error (MSE), of an estimator, δ, and variance/bias follows: prediction error = MSE(δ) = var(δ) + bias(δ)^2 . As can be seen from this relationship, MSE, or as calculated here, RMSE, alone is not a good indicator of the bias present in an estimate because RMSE could change due to a change in variance, with no change in bias. For this reason, both variance and RMSE were used for model comparisons.

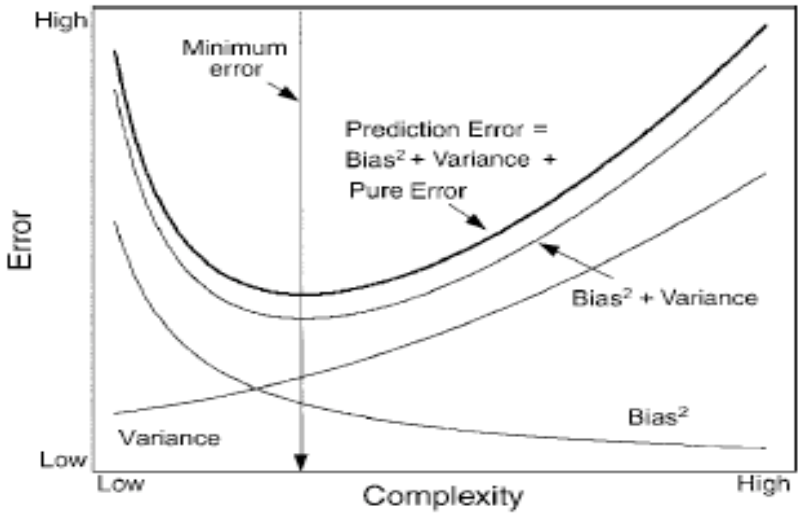


Figure 31. Theoretical relationship between prediction error (RMSE) and model complexity; figure taken from De'ath, 2007

Comparisons of predictions of the main-effects models indicated that the BRT method provided the most robust predictions, where the reduced RMSE, as compared to the GLM and GAM predictions, was representative of decreases in both variance and bias. The introduction of all two-way interactions, BRT 2-way model, resulted in a further decrease in RMSE. However, here, the decrease in RMSE resulted from a large enough decrease in bias to offset the increased variance (See Figure 28 and Table 8), demonstrating the bias-variance trade-off depicted in Figure 31. Nonetheless, as discussed above, greater bias reduction despite increases in variance

may be ideal if the objective of the standardization model is to produce annual estimates of abundance.

From the comparisons of relative influence of model predictors, all the main-effects models produced comparable results for the highest ranked variables. For example, the five most influential predictors in the main-effects models, in descending order, were: target, ln(area), year, bait and X-cent, for the GLM; target, ln(area), year, bait and season, for the GAM; target, bait, ln(area), year and Y-cent, for the BRT model. However, the introduction of two-way interactions causes a substantial re-ordering of predictor ranks, where year, ln(bait weight), Y-cent, bait and target are the five most influential variables. This re-arrangement of ranks reflects the importance of accounting for all interaction terms, including counter-intuitive interactions. For example, in the two-way interaction model examined here, the "bait weight" and "X-cent" terms produced the strongest interaction, which would have been difficult to predict *a priori*. The rise in ranks of variables related to space-time demonstrates the importance of considering these variables in all interactions. For stock assessment purposes, the "year" term is of particular importance. In the 2-way interaction model presented here, the year term rose to first position of relative importance accounting for 11.23% of the total explained deviance (Figure 18), as compared to fourth place in the main-effects BRT model, with 9.53% (Figure 15). Interestingly, the GIS-derived predictor, "ln(area)", proved to be an influential in the model, although interpretation may be confounded. Considering that longlines are not actively towed, the minimum convex polygon used to delimit the "area swept" by the gear may actually be acting as a proxy for placement and orientation, with respect to dominant water currents.

Often, interactions are not included in standardization models due to the complications encountered in extracting the year effect from the final model. Calculations of the year term are complicated further if interactions with categorical predictors are used, and even more so, if the

interactions of the categorical year term itself are included (Maunder & Punt, 2004). This is the case, however, for GLM and GAM methods, but not for BRTs. Here, the BRT two-way interaction model was fitted in less than a minute, and the year term was extracted just as quickly. Nonetheless, the ease with which interactions of the year term can be included in BRT models raises questions concerning the meaning of the year-effect. If the year-effect theoretically represents variations in local abundance, is it appropriate to include all possible interactions of the year term? For example, it would be reasonable to assume that the interaction of year and any measure of primary productivity (*e.g.* net primary productivity, chlorophyll concentration, *etc.*) would be meaningful, and logically, primary productivity could affect prey species abundance, thereby influencing local abundance of yellowfin tuna. However, a "year x bait weight" cross-term is related to catchability, and as such, should be excluded with the other predictors of catchabililty.

Results of residual analyses indicated that the raw data used here ($CPUE_{yft}$) exhibited a relatively weak level of spatial autocorrelation, which was passed through the main-effects models with only a slight decrease in severity, or slope, as observed in the semivariogram plots of the model residuals. For the two-way BRT model, there was a notable decrease in the level of spatial autocorrelation in the model residuals, indicating that the source was, predominantly, from two-way interactions (Figure 25). Other sources of spatial non-stationarity persisting in the BRT two-way model residuals (Figures 26) may be explored, potentially, by increasing the "interaction depth" parameter of the BRT model, or maybe more efficiently, in using a local analysis technique, such as geographically weighted regression (GWR) (*e.g.* Windle *et al*., 2009).

In general, there was consensus amongst the standardized indices of abundance from the four models studied here, which suggest a gradual decline in yellowfin tuna CPUE for the Gulf

of Mexico longline fishery. These results are consistent with the relative indices of abundance from the Gulf of Mexico (Brown & Ramirez-Lopez, 2009) used by the International Commission for the Conservation of Atlantic Tunas (ICAAT) in the 2008 Atlantic yellowfin tuna stock assessment (ICCAT, 2008). Formal comparisons, however, were not made between the present study and the ICAAT indices due to differences in data availability (*i.e.* Mexican observer data not analyzed, here) (Figure 30). Nonetheless, small differences did exist amongst the compared models, particularly estimates of catch-rate represented by the fully-parameterized GLM and GAM models *versus* the BRT-based models. Differences between these two model types were most apparent for years 2000 and 2004-2005. The year 2000 estimates show a dichotomy in how the fully-parameterized and the BRT-based models fitted the data. Estimates for the final years, 2004-2005, appear slightly more optimistic for the fully-parameterized models than for BRT-based indices, where slopes of the former group were markedly steeper (See Figure 29). However, these modest differences are unlikely to have practical significance on population model projections. Furthermore, as the BRT method does not provide variance estimates for the model parameters, the statistical significance of these differences in annual estimates was not determined.

Overall, the performance indices for the main-effects models indicate that the BRT method provided the most robust model predictions, as discussed above. Nevertheless, due to only slight differences amongst these indices, results are not conclusive; differences observed in the results for the main-effects models could be due to chance alone. Future comparisons should be performed using different datasets to train and test the models, from which multiple, random training and test sets could be drawn. Although model comparisons were inconclusive, the BRT method provides other conveniences, such as flexibility in handling predictor variables and the ease with which interaction effects can be incorporated, or just explored. On the other hand, the

principal shortcoming of the BRT method is the lack of error measurement for the extracted year term, which is the basis of the weighting system used when ranking indices of abundance from multiple sources. However, error estimates could be obtained in implementing a bootstrapping routine.

Here, I provided a first comparison of BRT performance with traditional GLM and GAM methods, and found several statistical and computational advantages to using the newer method. Future research could build upon these trials in incorporating novel datasets, or in extending comparisons to include zero-count data in comparing delta-GLM *versus* a delta-BRT model. A recently developed R package, "mboost" (Hothorn *et al.*, 2009), applies the boosting methodology to GLMs and GAMs, whose application to catch-rate standardization also merits research attention. Until a single, integrated statistical modeling approach is developed that addresses all the statistical concerns embodied in catch-rate data (*i.e*. zero-inflation, over-dispersion, autocorrelation, *etc*.), a multi-component, spatially-aware, non-linear mixture model will probably be the next to best approach. Such a model could take the form of a delta-BRT-GWR.

Recent advances in statistical and machine learning technology have allowed fisheries scientists to begin addressing the numerous assumptions previously required in standardizing fisheries data. Given the sheer abundance of catch-rate data, and the varying quality and statistical properties thereof, the importance of refining methods used in the standardization process cannot be underestimated. As annual estimates of abundance produced by standardization models are a fundamental input into many stock assessments, the quality of the input will affect the quality of the output. Thus, in order to avoid becoming "part of a fairy tale about sustainable stocks" (Bishop, 2006) continual progress must be made in refining every step of the fisheries management process.

Literature Cited

Anderson, C.N.K., Hsieh, C., Sandin, S.A., Hewitt, R., Hollowed, A., Beddington, J., May, R.M. and Sugihara, G. 2008. Why fishing magnifies fluctuations in fish abundance. *Nature*. 452: 835-839.

Arreguin-Sanchez, F. 1996. Catchability: a key parameter for fish stock assessment. *Reviews in Fish Biology and Fisheries* 6: 221-242.

Baum, J.K., Myers, R.A., Kehler, D.G., Worm, B., Harley, S.J. and Doherty, P.A. 2003. Collapse and conservation of shark populations in the northwest Atlantic. *Science*. 299: 389-392.

Battaile, B.C. and Quinn, T.J. 2004. Catch per unit effort standardization of the eastern Bering Sea walleye Pollock (Theragra chalcogramma) fleet. *Fisheries Research*. 70:161-177.

Behrenfeld, M.J. and Falkowski, P.G. 1997. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography*. 42: 1-20.

Beyer, H. L. 2004. Hawth's Analysis Tools for ArcGIS. Available at http://www.spatialecology.com/htools.

Bigelow, K.A, Boggs, C.H. and He, X. 1999. Environmental effects on swordfish and blue shark catch rates in the US North Pacific longline fishery. *Fisheries Oceanography*. 8: 178-198.

Bishop, J. 2006. Standardizing fishery-dependent catch and effort data in complex fisheries with technology change. *Rev. Fish Biol. Fisheries*. 16: 21-38.

Bigelow, K.A. and Maunder, M.N. 2007. Does habitat or depth influence catch rates of pelagic species? *Can. J. Fish. Aquat. Sci.* 64: 1581-1594.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.G. 1984. *Classification and regression trees*. Wadsworth International Group, Belmont, California, USA.

Brown, C.A. and Ramirez-Lopez, K. 2009. Standardized catch rates for yellowfin tuna (*Thunnus albacares*) in the Gulf of Mexico longline fishery for 1992-2007 based upon observer programs from Mexico and the United States. *Col. Vol. Sci. Pap. ICCAT*. 64: 1068-1079.

Cappo, M., De'ath, G., Boyle, S., Aumend, J., Olbrich, R., Hoedt, F., Colton, C., Perna, P. and Brunskill, G. 2005. Development of a robust classifier of freshwater residence in barramundi (Lates calcarifer) life histories using elemental ratios in scales and boosted regression trees. *Marine and Freshwater Research*. 56: 1-11.

Cheng, Y.W. and Gallinat, M.P. 2004. Statistical analysis of the relationship among environmental variables, inter-annual variability and smolt trap efficiency of salmonids in the Tucannon River. *Fisheries Research*. 70: 229-238.

Ciannelli, L., Fauchald, P., Chan, K.S., Agostini, V.N. and Dingsor, G.E. 2008. Spatial fisheries ecology: Recent progress and future prospects. *Journal of Marine Systems*. 71: 223-236.

Damalas, D., Meglofonou, P. and Apostolopoulou, M. 2007. Environmental, spatial, temporal and operational effects on swordfish (Xiphias gladius) catch rates of eastern Mediterranean Sea longline fisheries. *Fisheries Research*. 84: 233-246.

De'ath, G. 2002. Multivariate regression trees: a new technique for constrained classification analysis. *Ecology*. 83: 1103-1117.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. *Ecology*. 88:243-251.

De'ath, G. and Fabricius, K.E. 2000. Classification and regression trees: a powerful yet simple technique for the analysis of complex ecological data. *Ecology*. 81: 3178-3192.

Dick, E.J. 2004. Beyond 'lognormal versus gamma': discrimination among error distributions for generalized linear models. *Fisheries Research*. 70: 351-366.

ESRI, 2005. ArcGIS. Environmental Systems Research Institute, Redlands, CA.

Elith, J., Leathwick, J.R., Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*. 77: 802-813.

Gaertner, D. and Dreyfus-Leon, M. 2004. Analysis of non-linear relationships between catch per unit effort and abundance in a tuna purse-seine fishery simulated with artificial neural networks. *ICES Journal of Marine Science*. 61: 812-820.

Glazer, J.P. and Butterworth, D.S. 2002. GLM-based standardization of the catch per unit effort series for South African west coast hake, focusing on adjustments for targeting other species. *South African Journal of Marine Science*. 24:323-339.

Goni, R., Quetglas, A. and Renones, O. 2003. Differential catchability of male and female European spiny lobster Palinurus elephas (Fabricius, 1787) in traps and t rammelnets. *Fisheries Research*. 65: 295-307.

Gordoa, A., Maso, M. and Voges, L. 2000. Monthly variability in the catchability of Namibian hake and its relationship with environmental seasonality. *Fisheries Research* 48: 185-195.

Guisan, A., Edwards, T.C and Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*. 157: 89-100.

Haggarty, D.R. and King, J.R. 2006. CPUE as an index of relative abundance for nearshore reef fishes. *Fisheries Research* 81: 89-93.

Harley, S.J., Myers, R.A. and Dunn A. 2001. Is catch-per-unit-effort proportional to abundance? *Can. J. Fish. Aquat. Sci.* 58: 1760-1772.

Hastie, T. and Tibshirani, R. 1987. Generalized additive models: some applications. *J. Am. Stat. Assoc.* 82: 371-386.

Hastie, T.J., Tibshirani, R.J. and Friedman, J.H. 2001. *The elements of statistical learning*. Springer-Verlag, New York, New York, USA.

Hazin, H. and Erzini, K. 2008. Assessing swordfish distributions in the South Atlantic from spatial predictions. *Fisheries Research*. 90: 45-55.

Howell, E.A. and Kobayashi, D.R. 2006. El Nino effects in the Palmyra Atoll region: oceanographic changes and bigeye tuna (Thunnus obesus) catch rate variability. *Fisheries Oceanography*. 15: 477-489.

Hothorn, T., Buhlmann, P., Kneib, T., Schmid, M. and Hofner, B. 2009. Package 'mboost'. *R-News*. 16:56:33.

ICCAT, 2008. Report of the 2008 ICCAT yellowfin and skipjack stock assessment meeting (Florianopolis, Brazil—July 21 to 29, 2008). SCRS/2008/016.

ICCAT, 2007. Report of the 2006 Atlantic swordfish stock assessment session (Madrid, September 4 to 8, 2006). *Col. Vol. Sci. Pap. ICAAT*. 60: 1787-1896.

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T. and Taylor, P. 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*. 321: 267-281.

Lo, N. C., Jacobson, L.D. and Squire, J.L. 1992. Indices of relative abundance from fish spotter data based on del-lognormal models. *Can. J. Fish. Aquat. Sci*. 49: 2515-2526.

Maunder, M.N. and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent approaches. *Fisheries Research*. 70: 141-159.

Minami, M. Lennert-Cody, C.E., Gao, W. and Roman-Verdesoto, M. 2007. Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fisheries Research*. 84: 210-221.

Nelder, J.A. and Wedderburn, R.W.M. 1972. Generalized linear models. *J. R. Stat. Soc. A*. 135: 370-384.

Nishida, T. and Chen, D-G. 2004. Incorporating spatial autocorrelation into the general linear model with an application to the yellowfin tuna (Thunnus albacares) longline CPUE data. *Fisheries Research*. 70: 265-274.

Ortiz, M. and Arocha, F. 2004. Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: billfish in the Venezuelan tuna longline fishery. *Fisheries Research*. 70: 275-297.

Paulo, J.R. and Diggle, P.J. 2001. geoR: a package for geostatistical analysis. *R-News*. 1: 15-18.

Perry, R.I., Cury, P., Brander, K., Jennings, S., Mollmann, C. and Planque, B. 2009. Sensitivity of marine systems to climate and fishing: Concepts, issues and management responses. *Journal of Marine Systems*. In press.

Planque, B., Fromentin, J.M., Cury, P., Drinkwater, K., Jennings, S., Perry, R.I., Kifani, S. 2009. How does fishing alter marine populations and ecosystems sensitivity to climate? *Journal of Marine Systems*. In press.

Quirijns, F.J., Poos, J.J. and Rijnsdorp, A.D. 2008. Standardizing commercial CPUE data in monitoring stock dynamics: Accounting for targeting behaviour in mixed fisheries. *Fisheries Research*. 89: 1-8.

Ridgeway, G. 2009. Package 'gbm'. *R-News*. 08:05:15.

Roberts, J.J., B.D. Best, D.C. Dunn, E.A. Treml, and P.N. Halpin (in review) Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modeling & Software*. Available online: http://mgel.env.duke.edu/tools.

Rodriguez-Marin, E., Arrizabalaga, H., Ortiz, M., Rodriguez-Cabello, C., Moreno, G. and Kell, L.T. 2003. Standardization of bluefin tuna, Thunnus thynnus, catch per unit effort in the baitboat fishery of the Bay of Biscay (Eastern Atlantic). *ICES Journal of Marine Science*. 60: 1216-1231.

Rouyer, T., Fromentin, J.-M., Mendard, F., Cazelles, B., Briand, K., Pianet, R., Planque, R. and Stenseth, N.C. 2008. Complex interplays among population dynamics, environmental forcing, and exploitation in fisheries. *Proceedings of the National Academy of Sciences*. 105: 5420-5425.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch in CPUE analysis. *Fisheries Research*. 93: 154-162.

Simpfendorfer, C.A., Hueter, R.E., Bergman, U., Connett, S.M.H. 2002. Results of a fishery-independent survey for pelagic sharks in the western North Atlantic, 1977-1994. *Fisheries Research*. 55: 175-192.

Solmundsson, J., Karlsson, H. and Palsson, J. 2003. Sexual differences in spawning behaviour and catchability of plaice (Pleuronectes platessa) west of Iceland. *Fisheries Research*. 61: 57-71.

Sousa, P., Lemos, R.T., Gomes, M.C. and Azevedo, M. 2007. Analysis of horse mackerel, blue whiting, and hake catch data from Portuguese surveys (1989-1999) using an integrated GLM approach. *Aquatic Living Resources*. 20: 105-116.

Terceiro, M. 2003. The statistical properties of recreational catch rate data for some fish stocks off the northeast U.S. coast. *Fishery Bulletin*. 101: 653-672.

Tsuboi, J. and Endou, S. 2008. Relationships between catch per unit effort, catchability, and abundance based on actual measurements of Salmonids in a mountain stream. *Transactions of the American Fisheries Society*. 137: 496-502.

Vayssieres, M.P., Plant, R.E. and Allen-Diaz, B.H. 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*. 11: 679-694.

Venables, W.N. and Dichmont, C.M. 2004. GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research*. 70: 319-337.

Walsh, W.A. and Kleiber, P. 2001. Generalized additive model and regression tree analyses of blue shark (Prionace glauca) catch rates by the Hawaii-based commercial longline fishery. *Fisheries Research*. 53: 115-131.

Walsh, W.A., Kleiber, P. and McCracken, M. 2002. Comparison of logbook reports of incidental blue shark catch rates by Hawaii-based longline vessels to fishery observer data by application of a generalized additive model. *Fisheries Research*. 58: 79-94.

Windle, M.J.S., Rose, G.A., Devillers, R. and Fortin, M-J. 2010. Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. *ICES Journal of Marine Science*. 67:1-10.

Wood, S.N. 2001. MGCV: GAMs and generalized ridge regression for R. *R News*. 1(2): 20-25.

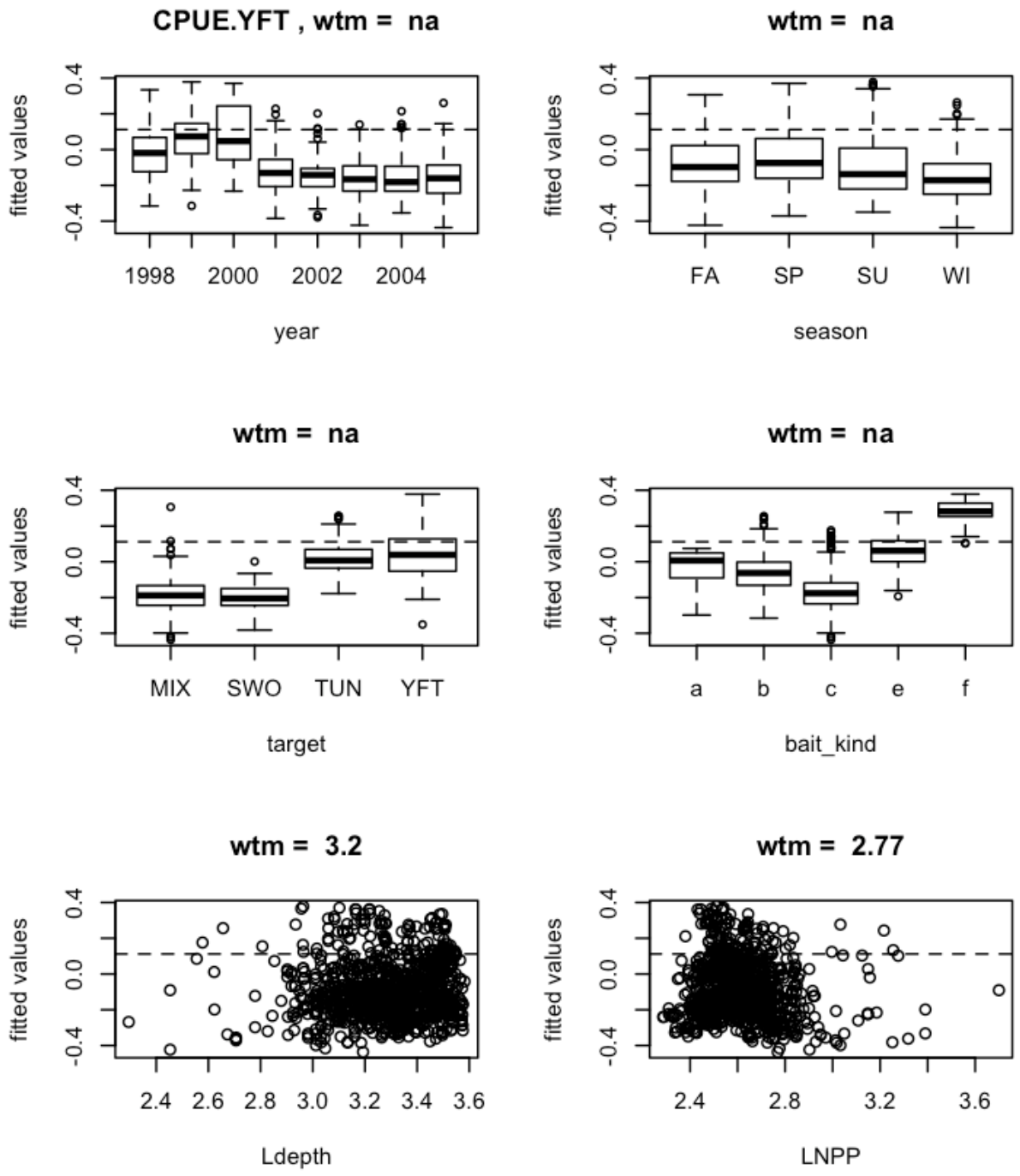
Wood, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Ass*. 70: 495-518.

Wood, S.N. 2008. Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Statsit. Soc. B.* 70: 495-518.

Zagaglia, C.R., Lorenzzetti, J.A. and Stech, J.L. 2004. Remote sensing data and longline catches of yellowfin tuna (Thunnus albacares) in the equatorial Atlantic. *Remote Sensing of Environment*. 93: 267-281.

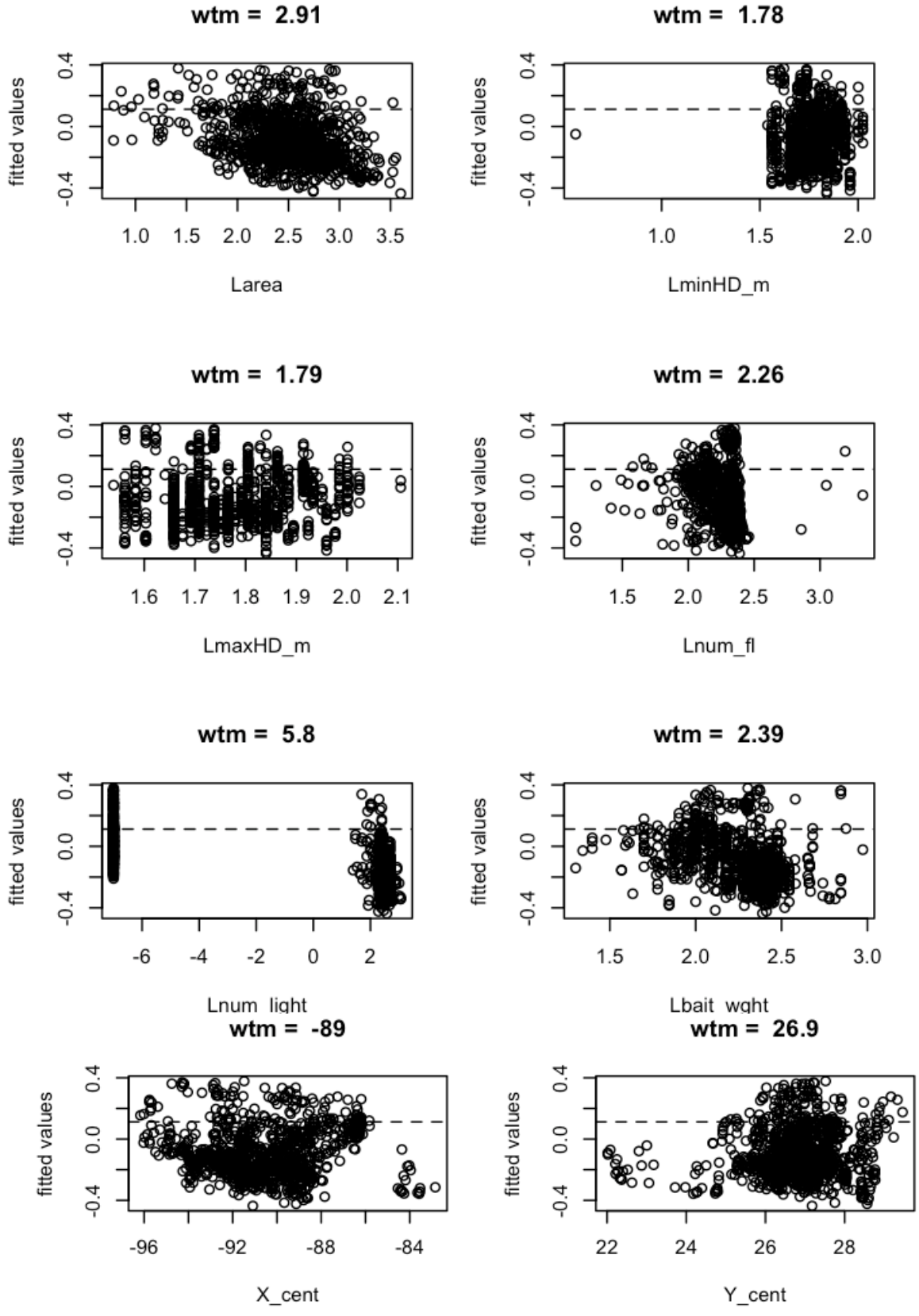
Ziegler, P.E., Frusher, S.D. and Johnson, C.R. 2003. Space-time variation in catchability of southern rock lobster Jasus edwardsii in Tasmania explained by environmental, physiological and density-dependent processes. *Fisheries Research*. 61: 107-123.

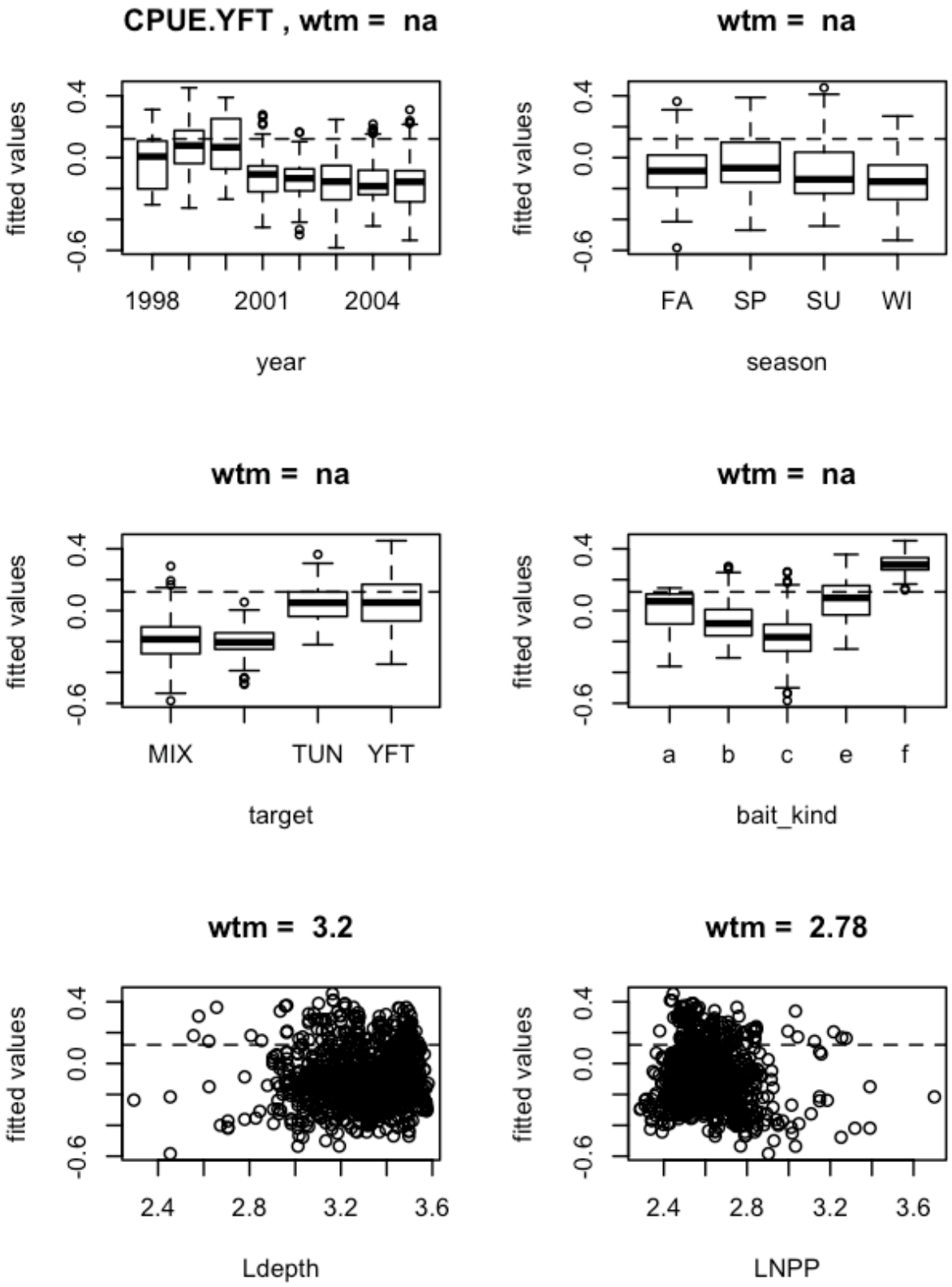
Appendix I: Fitted Versus Predicted Value Plots of BRT-based Model Predictors



A1.1 BRT: Fitted versus observed values for the BRT main effects model, with the weighted mean (wtm) of continuous variables indicated, continued below

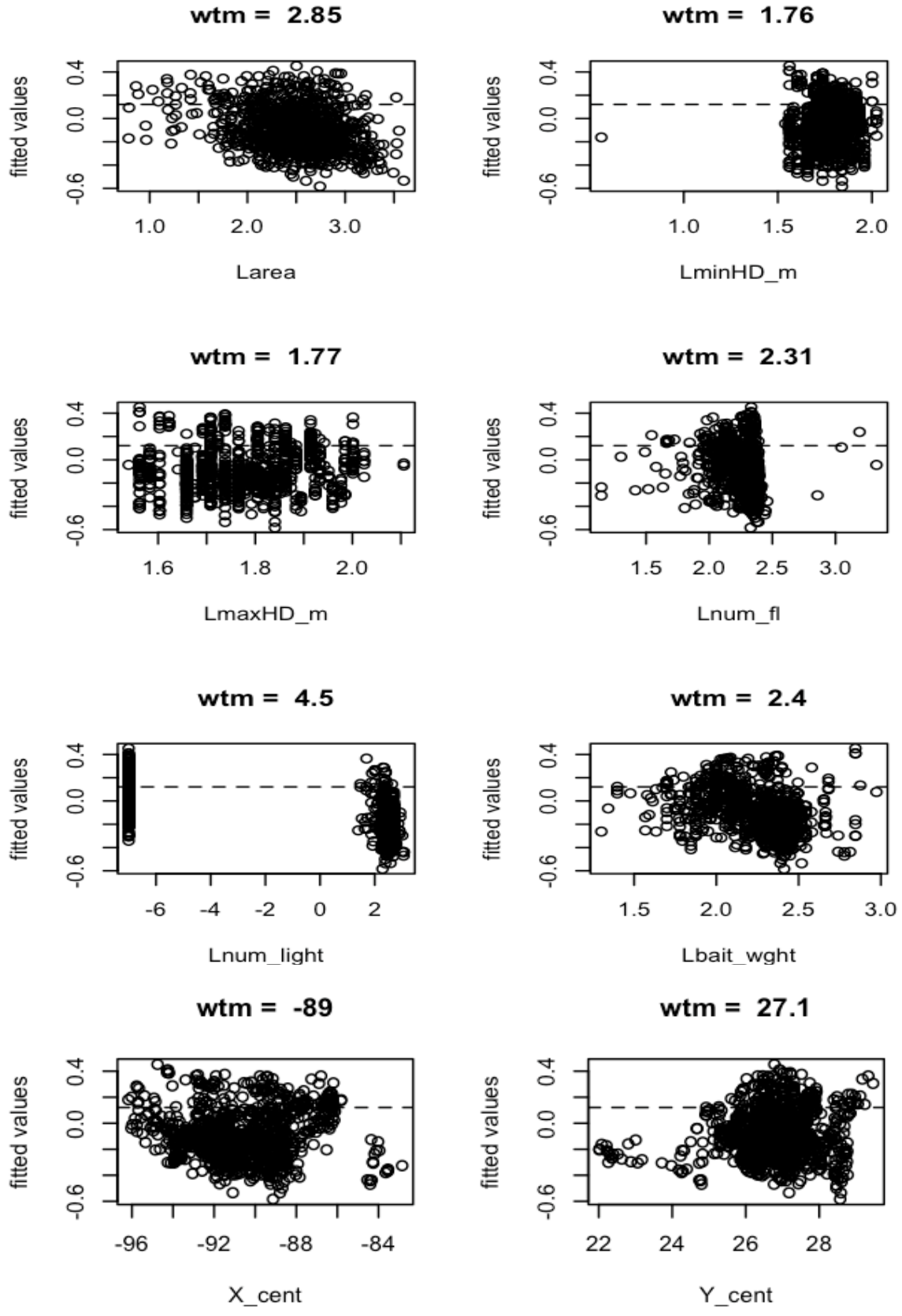
figure continued





A1.2 BRT 2-way: Fitted versus observed values for the BRT-2way model, with the weighted mean (wtm) of continuous variables indicated, continued below

figure continued



Appendix II: R Code Used for Analyses and Figures

```
rm(list=ls())
catch<-read.csv("F:\\data_analysis\\final_analyses\\stats_catch_mod3.csv",header=T)
attach(catch)
names(catch)
dim(catch)

#Random stratified sampling without replacement (test.set=30%, train.set=70%)
library(sampling)
test<-strata(catch,c("year"),size=c(21,55,53,59,48,80,79,85),method="srswor")
test.set<-getdata(catch,test.set)
row.names<-as.vector(test.set$ID_unit)
train.set<-catch[-row.names,]

> table(test.set$year)
1998 1999 2000 2001 2002 2003 2004 2005
  21   55   53   59   48   80   79   85

> table(train.set$year)
1998 1999 2000 2001 2002 2003 2004 2005
  50  128  124  137  112  188  185  199

> table(catch$year)
1998 1999 2000 2001 2002 2003 2004 2005
  71  183  177  196  160  268  264  284

write.csv(test.set,file="F:\\data_analysis\\final_analyses\\Analyses\\test_set.csv")
write.csv(train.set,file="F:\\data_analysis\\final_analyses\\Analyses\\train_set.csv")

# Process training dataset
catch.train<-
read.csv("/Users/shaneabeare/Desktop/data_analysis/final_analyses/Analyses/Training/train_set.
csv",header=T)
attach(catch.train)
names(catch.train)
dim(catch.train)
# 1123 61

# Remove NAs from data table
catch.names<-
c("unique","CPUE.YFT","year","season","month2","target","bait_kind","Ldepth","LNPP","Lare
a","LminHD_m","LmaxHD_m",
"soak_t","Lnum_fl","Lnum_light","Lbait_wght","f_hooks","X_cent","Y_cent")
yft.train<-catch.train[catch.names]
dim(yft.train)
# 1123 19
yft.train2<-na.omit(yft.train)
```

```
dim(yft.train2)
#1016 19
attach(yft.train2)
names(yft.train2)
yft.train2$year<-as.factor(yft.train2$year)
yft.train2$season<-as.factor(yft.train2$season)
offset<-log(yft.train2$soak_t*yft.train2$f_hooks/1000)

# Process test dataset
catch.test<-
read.csv("/Users/shaneabeare/Desktop/data_analysis/final_analyses/Analyses/Test/test_set.csv",h
eader=T)
attach(catch.test)
names(catch.test)
dim(catch.test)
#480 64
yft.test<-catch.test[catch.names]
dim(yft.test)
# 480 19
yft.test2<-na.omit(yft.test)
dim(yft.test2)
#436 19
attach(yft.test2)
names(yft.test2)
yft.test2$year<-as.factor(yft.test2$year)
yft.test2$season<-as.factor(yft.test2$season)
#########################################################################

#Descriptive statistics
#Figure1# Pie chart of abundant catch
pie.chart<-
c(sum(BLK),sum(BUM),sum(DOL),sum(FAL),sum(GEM),sum(LAX),sum(SAI),sum(SRX),su
m(SWO),sum(WAH),sum(WHM),sum(YFT))
pie.names=c("BLK","BUM","DOL","FAL","GEM","LAX","SAI","SRX","SWO","WAH","WH
M","YFT")
pie(pie.chart,labels=pie.names,col=topo.colors(15),radius=1)

#Figure2 & 3# Empirical distributions of Catch datalibrary(vcd)
yft.fit<-goodfit(YFT,type=c("poisson"))
yft2.fit<-goodfit(YFT,type=c("nbinomial"))
rootogram(yft.fit,main="YFT Catch\n\(Poisson)",xlab="Catch per
set",xlim=c(0,30),scale=c("raw"),type=c("standing"),points_gp=gpar(col="red",cex=0.5),pch=19
)
rootogram(yft2.fit,main="YFT Catch\n\(Negative binomial)",xlab="Catch per
set",xlim=c(0,30),scale=c("raw"),type=c("standing"),points_gp=gpar(col="red",cex=0.5),pch=19
)

#Figure4# Log(CPUE) histogram and Q-Q plot
```

```
par(mfrow=c(2,1))
hist(CPUE.YFT2,col="gray",main="YFT CPUE",xlab="CPUE [ln(Catch per
1000hook*hours)]",ylab="frequency",xlim=c(-1.5,1.5))
lines(yft.x,yft.norm,col="red",lwd=2)
qqnorm(CPUE.YFT2)
qqline(CPUE.YFT2,col="red",lwd=2)

#Figure5# Bar graph: Number of sets/year, target and bait kind summary
library(colorRamps)
target.year<-table(target,year)
barplot(target.year,ylim=c(0,350),ylab="Number of
sets",xlab="Year",font.lab=2,col=blue2green(4))
legend("topleft",title="Targets",c("MIX","SWO","TUN","YFT"),
fill=blue2green(4),horiz=TRUE,inset=0.025)
bait_kind.year<-table(bait_kind,year)
barplot(bait_kind.year,ylim=c(0,300),ylab="Number of
sets",xlab="Year",font.lab=2,col=green2red(5))
legend("topleft",title="Bait kind",c("a","b","c","e","f"),
fill=green2red(5),horiz=TRUE,inset=0.025)
##############################################################################

# Statistical Models
################################Zero-truncated log-normal GLM##################
yft.glm.noff<-glm(CPUE.YFT~year + season + target + bait_kind + Ldepth + LNPP + Larea +
LminHD_m + LmaxHD_m + Lnum_fl + Lnum_light + Lbait_wght + X_cent +
Y_cent,family=gaussian,data=yft.train2)
summary(yft.glm.noff)
yft.glm.noff.sum<-summary(yft.glm.noff)
delim.table(yft.glm.noff.sum,filename="/Users/shaneabeare/Desktop/data_analysis/final_analyse
s/Analyses/Training/YFT_training/yft.glm.noff.sum.csv")

#Figure8# GLM partial residual plots
library(car)
par(mfrow=c(2,2),mar=c(4,3,4,4))
cr.plots(yft.glm.noff,main=NULL,line=TRUE,smooth=FALSE,col="black",ylab="")

#Figure9# GLM diagnostic plots
par(mfrow=c(2,2))
plot(yft.glm.noff)

#########################Zero-truncated lognormal GAM#########################
library(mgcv)
yft.gam.noff<-gam(CPUE.YFT~year + season + target + bait_kind + s(Ldepth) + s(LNPP) +
s(Larea) + s(LminHD_m) + s(LmaxHD_m) + s(Lnum_fl) + s(Lnum_light) + s(Lbait_wght) +
s(X_cent) + s(Y_cent),family=gaussian,data=yft.train2)
summary(yft.gam.noff)

yft.gam.null<-gam(CPUE.YFT~1,family=gaussian,data=yft.train2)
```

```
anova(yft.gam.null,yft.gam.noff,test="F")

#Figure11# GAM partial residual plots
par(mfrow=c(2,2),mar=c(4,4,1,1))
plot(yft.gam.noff,residuals=TRUE,rug=TRUE,se=TRUE,all.terms=TRUE,shade=TRUE,ylab=""
)

#Figure12# GAM diagnostic plots
gam.check(yft.gam.noff)

#Figure13# Cooks distance for gam
plot(cooks.distance(yft.gam.noff),ylab="",main="GAM-YFT: Cook's Distance")
identify(cooks.distance(yft.gam.noff),tolerance=0.1)

########################## Boosted regression tree models##########################
source("brt.functions.R")
library(gbm)

# Boosted regression tree: main effects #Figure14#
set.seed(123)
yft.gbm2<-
gbm.step(data=yft.train2,gbm.x=3:16,gbm.y=2,family="gaussian",tree.complexity=1,learning.rat
e=0.01,bag.fraction=0.5,plot.folds=FALSE)

# Boosted regression tree: 2-way interaction model #Figure17#
set.seed(123)
yft.gbm2b<-
gbm.step(data=yft.train3,gbm.x=3:16,gbm.y=2,family="gaussian",tree.complexity=2,learning.rat
e=0.01,bag.fraction=0.5)

#Figures15 & 18# bar plot of predictor influence
par(mar=c(4,6,4,3))
summary(yft.gbm2,cBars=length(yft.gbm2$var.names),n.trees=yft.gbm2$n.trees,plotit=TRUE,o
rder=TRUE,normalize=TRUE,cex.axis=1,las=2,main=NULL)
summary(yft.gbm2b,cBars=length(yft.gbm2b$var.names),n.trees=yft.gbm2b$n.trees,plotit=TRU
E,order=TRUE,normalize=TRUE,cex.axis=1,las=2,main=NULL)

#Figures16 & 20# Fitted function plots
gbm.plot(yft.gbm2,smooth=TRUE,rug=TRUE,n.plots=14,write.title=F,rug.side=1,rug.lwd=1,rug
.tick=0.05,plot.layout=c(2,2))
gbm.plot(yft.gbm2b,smooth=TRUE,rug=TRUE,n.plots=14,write.title=F,rug.side=1,rug.lwd=1,ru
g.tick=0.05,plot.layout=c(2,2))

#Appendix I# Fits plus resids
gbm.plot.fits(yft.gbm2,plot.layout=c(2,2))
gbm.plot.fits(yft.gbm2b,plot.layout=c(2,2))
```

```
# Ranked interactions
yft.gbm2b.int<-gbm.interactions(yft.gbm2b)
yft.gbm2b.int

#Figure19# interaction plot
par(mfrow=c(2,2))
gbm.perspec(yft.gbm2b,13,12,x.range=c(-100,-
70),y.range=c(1,3),z.range=c(0,0.4),phi=20,col="light
blue",cex.axis=0.8,cex.lab=1,ticktype="detailed",xlim=c(-100,-60),ylim=c(0.5,3),main="(X-
cent) x (Lbait wght.)")
##########################################################################

# Residual Analyses
resid.plot<-data.frame(yft.train2$year,glm.resid,gam.resid,gbm2.resid,gbm2b.resid)
library(doBy)
glm.year<-summaryBy(formula=glm.resid~yft.train2.year,data=resid.plot,FUN=c(mean,sd))
gam.year<-summaryBy(formula=gam.resid~yft.train2.year,data=resid.plot,FUN=c(mean,sd))
gbm2.year<-summaryBy(formula=gbm2.resid~yft.train2.year,data=resid.plot,FUN=c(mean,sd))
gbm2b.year<-
summaryBy(formula=gbm2b.resid~yft.train2.year,data=resid.plot,FUN=c(mean,sd))
model.resid<-data.frame(glm.year,gam.year,gbm2.year,gbm2b.year)
write.csv(model.resid,"/Users/shaneabeare/Desktop/data_analysis/final_analyses/Analyses/Traini
ng/YFT_training/model.resid.csv")

#Figure21#
plot(density(glm.resid),col="red",lwd=2,ylim=c(0,1.5),main="")
lines(density(gam.resid),col="darkgreen",lwd=2)
lines(density(gbm2.resid),col="goldenrod3",lwd=2)
lines(density(gbm2b.resid),col="blue",lwd=2)
abline(v=0)
legend("topright",inset=0.025,title="Model residuals",legend=c("GLM","GAM","BRT","BRT 2-
way"),col=c("red","dark green","goldenrod3","blue"),lty=1,lwd=2)

#Figure22#
boxplot(glm.resid,gam.resid,gbm2.resid,gbm2b.resid,data=resid.plot,names=c("GLM","GAM","
BRT","BRT 2-way"),notch=TRUE)

#Figure23#
pairs(formula=~glm.resid+gam.resid+gbm2.resid+gbm2b.resid,data=resid.plot,labels=c("GLM",
"GAM","BRT","BRT 2-way"))

# Empirical semivariograms
library(geoR)
#breaks
dists<-dist(yft.train2[,18:19])
summary(dists)
breaks=seq(0,1,l=14)
```

```
# Omnidirectional semivariogram
yft.vario1<-variog(coords=yft.train2[,18:19],data=yft.train2[,2],option=c("bin"),breaks=breaks)
# Directional semivariograms
yft.vario2<-
variog(coords=yft.train2[,18:19],data=yft.train2[,2],option=c("bin"),breaks=breaks,unit.angle="d
egrees",direction=0,tolerance=22.5)
yft.vario3<-
variog(coords=yft.train2[,18:19],data=yft.train2[,2],option=c("bin"),breaks=breaks,unit.angle="d
egrees",direction=45,tolerance=22.5)
yft.vario4<-
variog(coords=yft.train2[,18:19],data=yft.train2[,2],option=c("bin"),breaks=breaks,unit.angle="d
egrees",direction=90,tolerance=22.5)
yft.vario5<-
variog(coords=yft.train2[,18:19],data=yft.train2[,2],option=c("bin"),breaks=breaks,unit.angle="d
egrees",direction=135,tolerance=22.5)

#Figure24# All directional variograms
plot(yft.vario1,type="b",lwd=3,main=NULL,pts.range=c(1),scaled=FALSE,ylim=c(0,0.2),var.lin
es=TRUE,xlab="distance (degrees)")
lines(yft.vario2$u,yft.vario2$v,col="red")
lines(yft.vario3$u,yft.vario3$v,col="dark green")
lines(yft.vario4$u,yft.vario4$v,col="dark blue")
lines(yft.vario5$u,yft.vario5$v,col="orange")
legend("bottomright",inset=0.025,title="Variogram Directions",legend=c("Omnidirect","0
deg","45 deg","90 deg","135 deg"),
col=c("black","red","dark green","dark blue","orange"),lty=1,lwd=2)

#Calculate model residual variograms
yft.vario.resid1<-
variog(coords=yft.train2[,18:19],data=residuals(yft.glm.noff),option=c("bin"),breaks=breaks)
yft.vario.resid2<-
variog(coords=yft.train2[,18:19],data=residuals(yft.gam.noff),option=c("bin"),breaks=breaks)
yft.vario.resid3<-
variog(coords=yft.train2[,18:19],data=residuals(yft.gbm2),option=c("bin"),breaks=breaks)
yft.vario.resid4<-
variog(coords=yft.train2[,18:19],data=residuals(yft.gbm2b),option=c("bin"),breaks=breaks)
yft.vario.resid5<-
variog(coords=yft.train2[,18:19],data=residuals(yft.gbm2c),option=c("bin"),breaks=breaks)

#Figure25# Residual variogram plots
plot(yft.vario1,type="b",main=NULL,pts.range=c(1),scaled=FALSE,ylim=c(0.05,0.15),var.lines
=TRUE,xlab="distance (degrees)")
lines(yft.vario.resid1$u,yft.vario.resid1$v,type="l",lwd=2,col="cornflowerblue")
lines(yft.vario.resid2$u,yft.vario.resid2$v,type="l",lwd=2,col="purple")
lines(yft.vario.resid3$u,yft.vario.resid3$v,type="l",lwd=2,col="dark green")
lines(yft.vario.resid4$u,yft.vario.resid4$v,type="l",lwd=2,col="darkgoldenrod2")
lines(yft.vario.resid5$u,yft.vario.resid5$v,type="l",lwd=2,col="blue")
```

```
#Figure26# GLM bubble plot
glm.resid.p<-residuals(yft.glm.noff,type="pearson")
mydata<-data.frame(glm.resid.p,yft.train2$X_cent, yft.train2$Y_cent)
coordinates(mydata)<-c("yft.train2.X_cent","yft.train2.Y_cent")
bubble(mydata,"glm.resid.p",col=c("black","darkgrey"),pch=16,maxsize=2,main="GLM",identif
y=FALSE,xlab="Longitude",ylab="Latitude",scales=list(draw=TRUE))

#Figure26# GAM bubble plot
gam.resid.p<-residuals(yft.gam.noff,type="pearson")
mydata2<-data.frame(gam.resid.p,yft.train2$X_cent,yft.train2$Y_cent)
coordinates(mydata2)<-c("yft.train2.X_cent","yft.train2.Y_cent")
bubble(mydata2,"gam.resid.p",col=c("black","darkgrey"),pch=16,maxsize=2,main="GAM",iden
tify=FALSE,xlab="Longitude",ylab="Latitude",scales=list(draw=TRUE))

#Figure26# BRT bubble plot
gbm2.resid.p<-residuals(yft.gbm2,type="pearson")
mydata3<-data.frame(gbm2.resid.p,yft.train2$X_cent,yft.train2$Y_cent)
coordinates(mydata3)<-c("yft.train2.X_cent","yft.train2.Y_cent")
bubble(mydata3,"gbm2.resid.p",col=c("black","darkgrey"),pch=16,maxsize=2,main="BRT",iden
tify=FALSE,xlab="Longitude",ylab="Latitude",scales=list(draw=TRUE))

#Figure26# BRT 2-way bubble plot
gbm2b.resid.p<-residuals(yft.gbm2b,type="pearson")
mydata4<-data.frame(gbm2b.resid.p,yft.train2$X_cent,yft.train2$Y_cent)
coordinates(mydata4)<-c("yft.train2.X_cent","yft.train2.Y_cent")
bubble(mydata4,"gbm2b.resid.p",col=c("black","darkgrey"),pch=16,maxsize=2,main="BRT 2-
way",identify=FALSE,xlab="Longitude",ylab="Latitude",scales=list(draw=TRUE))
#####################################################################

#Model predictions for year-effect
year.glm<-
predict(yft.glm.noff,newdata=NULL,type="term",se.fit=TRUE,terms="year",dispersion=NULL,
na.action=na.pass)
year.gam<-
predict(yft.gam.noff,type="terms",se.fit=TRUE,terms="year",block.size=1000,na.action=na.pass
)
year.gbm2<-plot(yft.gbm2,i.var=1,return.grid=TRUE)
year.gbm2b<-plot(yft.gbm2b,i.var=1,return.grid=TRUE)

#nominal catch
library(doBy)
nom.mean2<-summaryBy(formula=CPUE.YFT~year,data=yft.train2,FUN=c(mean,sd))
#compile
year.f<-as.factor(c(1998,1999,2000,2001,2002,2003,2004,2005))
glm.pred<-cbind(as.vector(unique(year.glm$fit)),as.vector(unique(year.glm$se.fit)))
gam.pred<-cbind(as.vector(unique(year.gam$fit)),as.vector(unique(year.gam$se.fit)))

year.effect<-data.frame(nom.mean2,glm.pred,gam.pred,year.gbm2,year.gbm2b)
```

```
write.csv(year.effect,"/Users/shaneabeare/Desktop/data_analysis/final_analyses/Analyses/Test/year.effect2.csv")

##Model predictions with test data
glm.test<-
predict(yft.glm.noff,newdata=yft.test2,type="response",se.fit=TRUE,dispersion=NULL,na.action=na.pass)
glm.fit<-as.vector(glm.test$fit)
glm.se<-as.vector(glm.test$se.fit)

gam.test<-predict(yft.gam.noff,newdata=yft.test2,type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)

gbm2.test<-
predict(yft.gbm2,yft.test2,n.trees=yft.gbm2$gbm.call$best.trees,type="response",se.fit=TRUE)
gbm2.fit<-as.vector(gbm2.test)

gbm2b.test<-
predict(yft.gbm2b,yft.test2,n.trees=yft.gbm2$gbm.call$best.trees,type="response",se.fit=TRUE)
gbm2b.fit<-as.vector(gbm2b.test)

year<-as.vector(yft.test2$year)
CPUE<-as.vector(yft.test2$CPUE.YFT)

model.perform<-data.frame(year,CPUE,glm.fit,glm.se,gam.fit,gam.se,gbm2.fit,gbm2b.fit)
write.csv(model.perform,"/Users/shaneabeare/Desktop/data_analysis/final_analyses/Analyses/Test/model.perform3.csv")

#Figure28# Boxplots of model predictions
attach(model.perform)
boxplot(CPUE,glm.fit,gam.fit,gbm2.fit,gbm2b.fit,notch=TRUE,names=c("CPUE","GLM","GAM","BRT","BRT 2-way"))
```

Vita

Shane M. Abeare, born December 1976 in Flint, Michigan, spent the early years of his life in Goodrich, Michigan, with his parents and three brothers. He graduated high school in 1995, and later, from the University of Michigan-Flint (UM) in December 1999. As an undergraduate of biology and chemistry at UM, Shane eagerly sought opportunities to gain experience in lab and field research. During the last couple of years as an undergraduate, he was invited to assist on two field research projects: a project studying the fruit bats of Papua, New Guinea and Queensland, Australia, headed by Dr. Eugene Studier; and another project investigating mycorrhizal fungi in the Peruvian Amazon rainforest, headed by Tracy Wacker.

After graduation, he quickly moved from the cold climate of Michigan to live in Davis, California, where he worked for a year as a research assistant at the Center for Neuroscience, University of California-Davis. After realizing that lab work was not his calling in life, he joined U.S Peace Corps and shipped-off to spend the next couple of years living in the village of Babadé in Togo, West Africa. In Togo, Shane worked as an environmentalist, promoting sustainable agriculture techniques, and later, worked as the interim-Director of the last remaining intact national park in Togo—Fazao-Malfakassa National Park.

After leaving the Peace Corps in 2002, he decided to spend some time travelling alone around West Africa, then to South Africa, back to Michigan and, finally, stopped to work another year in California. In southern California, Shane worked with the government to curb a viral outbreak affecting birds, as a member of the Exotic Newcastle Disease Taskforce at the California Animal Health and Food Safety Laboratory. After completing the mission in southern California, he enrolled in the African Mammalogy master's program at the University of Pretoria, South Africa. For his master's research, Shane designed a statistical model to predict

the movements of African buffalo in Kruger National Park, as part of a larger bovine tuberculosis research project.

After graduating from the University of Pretoria, in 2004, Shane assisted on a Great White Shark research project in Mosselbai, South Africa, and shortly thereafter, accepted a job with the Wildlife Conservation Society in the Republic of Congo. Working at Conkouati-Douli National Park in the Congo, he began to understand the need that exists in the conservation world for fisheries-trained marine conservationists, so after leaving the Congo, he moved to Paris, France, and from there applied to the oceanography and coastal sciences master's program at Louisiana State University (LSU).

Since acceptance at LSU, Shane has worked under the direction of Dr. Joseph Powers on statistical modeling techniques used to model fisheries catch-rate data, and will graduate with a Master's of Science in oceanography and coastal sciences and a minor in experimental statistics, in December 2009. After graduation, Shane will continue planning his doctoral research project being pursued through the University of New Orleans, which will address fisheries management and conservation issues in East Africa.