NOAA Technical Memorandum NMFS-AFSC-269

A Comparison of Statistical Methods to Standardize Catch-Per-Unit-Effort of the Alaska Longline Sablefish Fishery

by
I. Mateo and D. H. Hanselman

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Alaska Fisheries Science Center

February 2014

NOAA Technical Memorandum NMFS

The National Marine Fisheries Service's Alaska Fisheries Science Center uses the NOAA Technical Memorandum series to issue informal scientific and technical publications when complete formal review and editorial processing are not appropriate or feasible. Documents within this series reflect sound professional work and may be referenced in the formal scientific and technical literature.

The NMFS-AFSC Technical Memorandum series of the Alaska Fisheries Science Center continues the NMFS-F/NWC series established in 1970 by the Northwest Fisheries Center. The NMFS-NWFSC series is currently used by the Northwest Fisheries Science Center.

This document should be cited as follows:

> Mateo, I., and D. H. Hanselman. 2014. A comparison of statistical methods to standardize catch-per-unit-effort of the Alaska longline sablefish. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-269, 71 p.
>
> Document available: http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-269.pdf

Reference in this document to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.

NOAA Technical Memorandum NMFS-AFSC-269

A Comparison of Statistical Methods to Standardize Catch-Per-Unit-Effort of the Alaska Longline Sablefish Fishery

by
I. Mateo[1] and D. H. Hanselman[2*]

[1] National Research Council
Research Associateship Programs
500 Fifth Street NW (Keck 568)
Washington DC 20001

[2]Auke Bay Laboratories
Alaska Fisheries Science Center
17109 Point Lena Loop Rd
Juneau AK 99801-8344

*primary contact

www.afsc.noaa.gov

U.S. DEPARTMENT OF COMMERCE
Penny. S. Pritzker, Secretary
National Oceanic and Atmospheric Administration
Kathryn D. Sullivan, Under Secretary and Administrator
National Marine Fisheries Service
Eileen Sobeck, Assistant Administrator for Fisheries

February 2014

This document is available to the public through:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, VA 22161

www.ntis.gov

ABSTRACT

Improving existing catch per unit effort (CPUE) models for construction of a fishery abundance index is important to the Alaska sablefish (*Anoplopoma fimbria)* stock assessment. Performance of statistical methods including Generalized Linear Models (GLM), Generalized Additive Models (GAM), and Boosted Regression Trees (BRT) were evaluated using CPUE data collected by observers- from the sablefish longline fishery in the Gulf of Alaska, the Bering Sea, and the Aleutian Islands during 1995-2011. Due to the nonlinearity of several important covariates found during the diagnostics, GLM was dismissed as a potential method to standardize CPUE. Fitted GAM models for the Gulf of Alaska subregions: West Yakutat, Western Gulf, Central Gulf, and Southeast accounted for 42%, 29%, 30%, and 45% of total model deviance explained, respectively. BRT models accounted for 47%, 31%, 30%, and 46 %, respectively. For the Bering Sea and Aleutian Islands subregions, fitted GAM models accounted for 58% and 54% of total model deviance explained, respectively. BRT models accounted for 63% and 60% for the Bering Sea and the Aleutian Islands subregions, respectively. Predictive performance metrics (Root Mean Square Error) and 5-fold cross-validation results showed GAM and BRT models had similar predictive power. However, variance was significantly higher in GAM model predictions. In general, the BRT model performance was superior or equally robust to traditional methods such as GLM and GAM and should be considered as a potential statistical method for CPUE standardization.

CONTENTS

INTRODUCTION

Catch per unit effort (CPUE) is used widely in fisheries management and marine conservation efforts as direct proxy of abundance (Hilborn and Walters 1992, Harley et al. 2001, Erisman et al. 2011). CPUE is commonly obtained from commercial fishery-dependent data because it is readily available and less resource-intensive than conducting a statistically designed fishery-independent survey (Hilborn and Walters 1992, Harley et al. 2001, Erisman et al. 2011). CPUE is often assumed to have a linear relationship with abundance. This assumption has been challenged for many years as CPUE rates can be influenced by many factors such as fleet dynamics, schooling behavior, gear selection, and seasonal and spatial allocation of fishing effort in a way that interpretation of CPUE can be misleading if these confounding factors are not taken into account (Hilborn and Walters 1992, Harley et al. 2001). Hilborn and Walters (1992) identified two major situations that affect interpretation of mean CPUE as an index of stock abundance: 1) hyperstability (when abundance declines faster than CPUE declines), and 2) hyperdepletion (when abundance declines slower than CPUE declines). In general, a fishery with a situation of hyperstability could be attributed to schooling behavior or changes in the species spatial distribution rather than changes in abundance (van der Lee 2012). Hyperdepletion, on the other hand, can occur when a more vulnerable portion of the population is easily caught, followed by a more cryptic portion that avoids fishing mortality (van der Lee 2012). Hyperstability can cause overfishing to go undetected, while hyperdepletion can result in foregone yields when adopted management strategies such as catch limits are based on inaccurate estimates of abundance (Erisman et al. 2011, van der Lee 2012).

Sablefish (*Anoplopoma fimbria*) is a commercially important fish species in the North Pacific Ocean. Almost 90% of the Alaska sablefish catch is obtained using bottom longline gear

(Hanselman et al. 2008, 2010, 2012). One source of information on abundance trends for sablefish stock assessment is the index of abundance derived from the fishery longline catch-per-unit-effort (CPUE) time series. Recently, Hanselman et al. (2010) expressed concerns regarding whether the current CPUE index accurately represented sablefish stock abundance in Alaskan waters due to hyperstability of the index. Since the sablefish fishery moved to individual fishing quotas in 1995, fisherman could fish most of the year at their convenience, allowing vessels to target the best habitat rather than the fishing grounds closest to port (Sigler and Lunsford 2001). This diffusion of effort in time and space has made changes in catch rates difficult to detect. There have been spatial shifts in fishing in some of the subregions of the Gulf of Alaska and Bering Sea/Aleutian Islands regions. For example, centroids (yearly averages of latitude and longitude weighted by CPUE) of CPUE-weighted sablefish catch location from 1995 to2011 in West Yakutat show an easterly shift in CPUE distribution (Fig. 1) while the centroids of CPUE in the Bering Sea have moved northwest (Fig. 2).

The Alaska sablefish assessment authors acknowledged the difficulty of fully understanding and quantifying changes in the fishery that would explain the patterns observed in the fishery index of abundance (Hanselman et al. 2010). Because of these concerns, the 2009 Center for Independent Experts sablefish assessment review and a 2010 sablefish modeling workshop recommended the development of statistical models to standardize sablefish fishery CPUE in a way that provides reliable abundance indices for subsequent assessments (Hanselman et al. 2010). However, selecting a statistical model to use in predicting sablefish abundance is not easy given that the longline fishery data are often noisy, over-dispersed, or zero- inflated (Hanselman et al. 2010).

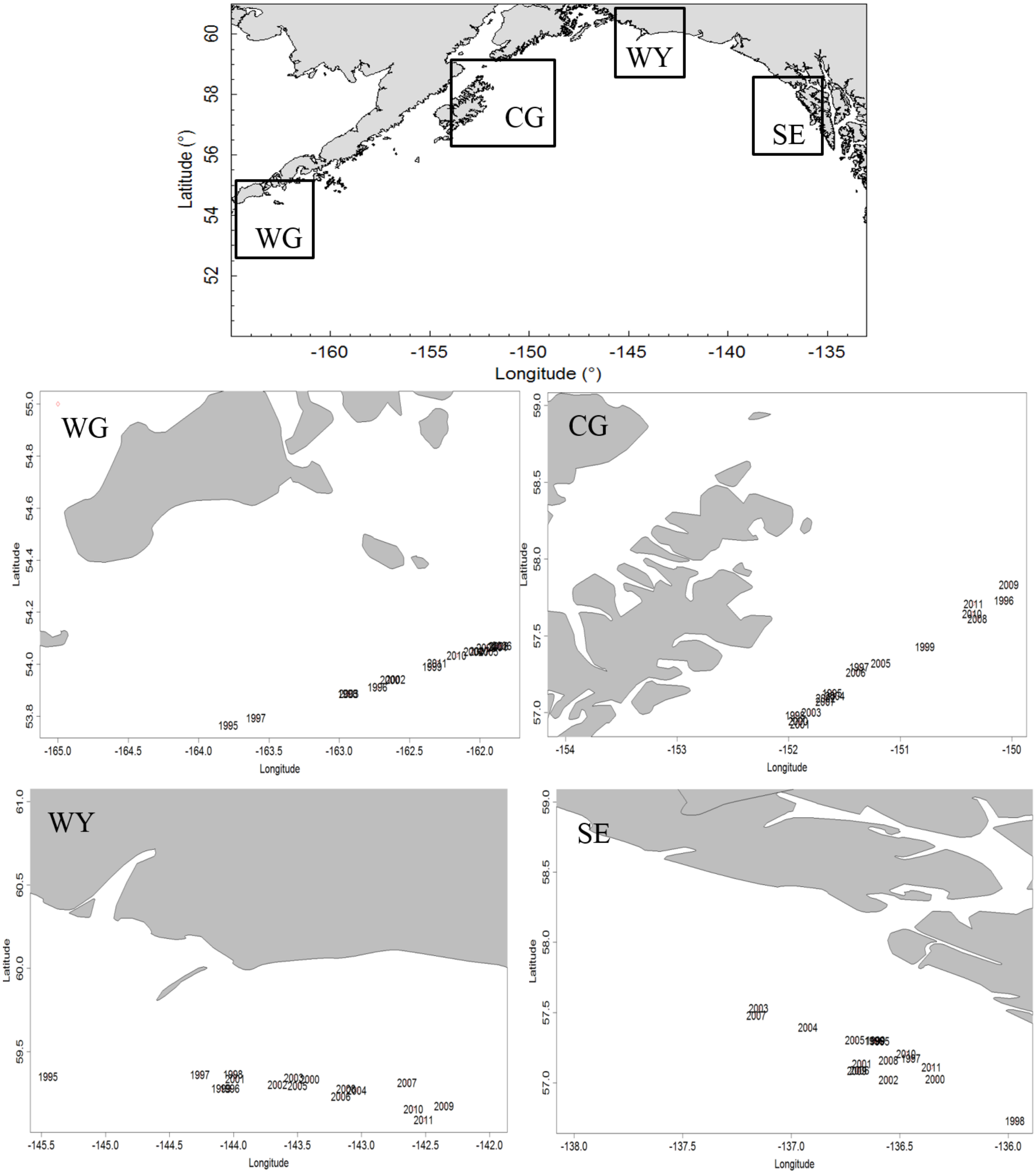


Figure 1. -- Centroids of sablefish CPUE spatial and temporal distribution for years 1995-2011 in the Gulf of Alaska subregions: West Yakutat (WY), Western Gulf (WG), Central Gulf (CG), Southeast (SE).

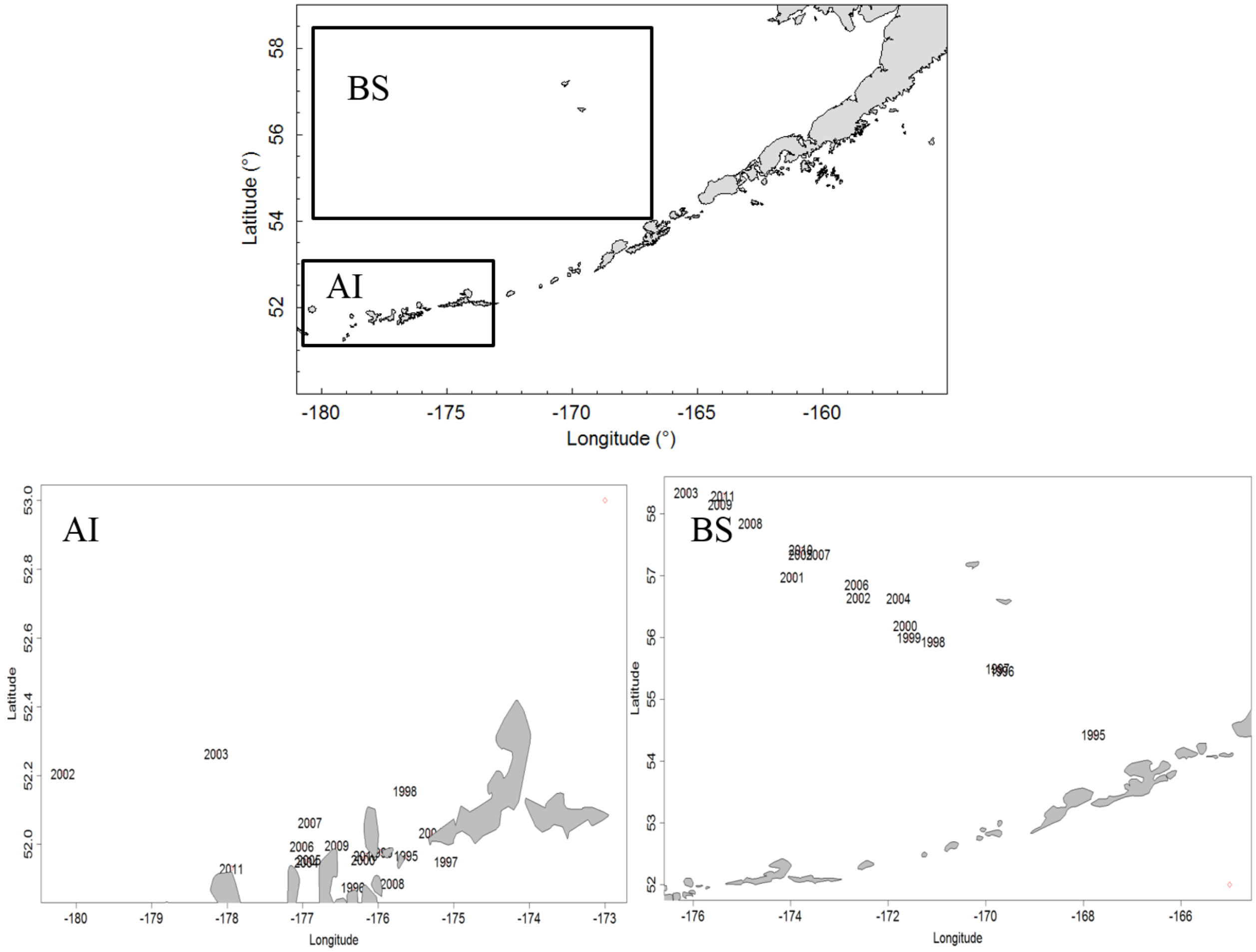


Figure 2. -- Centroids of sablefish CPUE spatial and temporal distribution for years 1995-2011 in the Bering Sea (BS) and Aleutian Islands (AI) regions.

There has been extensive research on CPUE standardization techniques (Quinn and Deriso 1999, Hinton and Maunder 2004, Venables and Ripley 2002, Maunder and Punt 2004). Among all these methods, generalized linear models (GLM, Venables and Ripley 2002) and generalized additive models (GAM, Venables and Ripley 2002, Wood 2006) are the most commonly used to standardize catch rates (Maunder and Punt 2004, Hinton and Maunder 2004, Venables and Dichmont 2004). The GLM differs from ordinary linear models by allowing fitting of categorical variables, variables that are not continuous, and Poisson data such as counts

(Venables and Ripley 2002, Maunder and Punt 2004). The GAM offers an extension from GLM that is more flexible for dealing with non-linear relationships of CPUE to spatial (e.g., latitude or longitude) and environmental variables (e.g., temperature, salinity, dissolved oxygen) (Venables and Ripley 2002, Maunder and Punt 2004, Wood 2006) due to the inclusion of smoothing functions for fitting model parameters.

A more recent approach in ecology is the utilization of boosted regression trees (BRT) to explain species distributions based on ecological and environmental characteristics (De'ath 2007, Elith et al. 2008, Abeare 2009, Pittman et al. 2009, Froeschke et al. 2010). This method has been successfully used to standardize CPUE of the yellowfin tuna (*Thunnus albacares*) longline fishery from the Gulf of Mexico (Abeare 2009). BRT methods were also used to model the spatial distribution of wahoo (*Acanthocybium solandri*) CPUE in the Gulf of Mexico and it outperformed other statistical methods such as GAMs (Martínez-Rincón et al. 2012).

The BRT model is a flexible regression modeling approach that is a robust extension of regression trees (Elith et al. 2008). Regression trees relate the response to the predictors by recursive binary splits. The boosting is a way to combine many simple regression trees into an overall model that has improved predictive performance. This boosting creates this group of simple regression trees by sequentially modeling the residuals from each subset of data during the model fit. This forward stagewise fitting and model averaging approach reduces bias and variance (Elith et al. 2008). Among some of the advantages of this technique are that it handles interactions between variables more efficiently than traditional methods such as GLMs and GAMs, and it more efficiently addresses issues like missing data and outliers. Disadvantages over traditional methods are that it is a more complex, and results are not as easy to interpret under standard frequentist theory (e.g., P-values of coefficients) (De’ath 2007, Elith et al. 2008, Abeare 2009).

The primary objective of this study is to evaluate the use of GLMs, GAMs, and BRTs to standardize abundance indices of the sablefish longline fisheries in Alaskan waters by comparing model performance (model fitting and prediction error) among all models. This study documents analyses of data from Gulf of Alaska (GOA) and the Bering Sea and Aleutian Islands (BSAI) longline fisheries.

MATERIALS AND METHODS

Data Sources

The longline fishery CPUE data used were collected from the National Marine Fisheries Service Fisheries Monitoring and Analysis Program, and provided by the Alaska Fisheries Information Network (AKFIN). Observer sampling coverage in Alaskan waters depends on the size of the vessel, the gear utilized, and the fishery that the vessel is operating in (Cahalan 2010). There are three sampling strata based on vessel overall length: less than 60 feet, between 60 and 125 feet, and greater than 125feet. The observer program (through 2011) only deployed observers on vessel size classes of 60-125 and >125 feet. Longline vessels in the 60-125 feet size class are required by mandate to have observer coverage on 30% of their trips (Cahalan 2010). During typical longline fishing activities, an observer samples the species composition and the number caught for about one-third of a set. These sampling periods are distributed systematically throughout the entire set. Total size of the set is calculated by the product of the mean number of hooks per gear segment (skate) and the total number of skates (Cahalan 2010). Records of longline sets available in the observer program extend from 1991 to 2011.

There were a number of potential explanatory variables in the database. Location information available was latitude and longitude at the start and end of each set. Depth information was available for the bottom and the gear. Performance of the gear/set was documented. Vessel information was only individual vessel ID codes and vessel lengths. The date of each haul was also available. Another field indicates if the target was Individual Fishing Quotas (IFQ), which would mean it was either a sablefish or a halibut target. There were also CPUE data for all other species caught on the set. We filtered the database as follows:

Missing Values -- Variables that had limited amount of observations were eliminated from the database. These included average hook spacing, latitude at beginning of fish sets (58% of observations with values), longitude at beginning of fishing sets (64% of observations with values), and gear depth (it was usually the same as bottom depth, or missing).

Vessel ID **--** Different levels of fishing experience can potentially affect CPUE (Hinton and Maunder 2004). Therefore, analyses performed here limited the database to information from vessels that fished continuously from 1995 to 2011. Fifty vessels were used from 100 vessels in the original database. Data were only modeled from 1995 to2011 because the fishery changed from open-access to IFQs in 1995, which marked a major change in fishery behavior (Sigler and Lunsford 2001).

Bottom Depth -- Sets that contained bottom depths shallower than 183 m (100 fathoms) were discarded because adult sablefish are rarely caught in depths shallower than 100 fathoms (Hanselman et al. 2008).

Performance Description **--** The performance description variable is information that describes any factors that may have affected catch rates. These descriptions are subjective and were not a large part of the database. Therefore, only records of sets with "No problem" were selected for subsequent analysis.

***IFQ Flag* --** Only records that were designated as IFQ sets (meaning sablefish or halibut targets) were used.

***CPUE of Dominant Species in the Longline Fishery* --** Dominant species CPUEs to be considered as potential explanatory variables were selected based on K- means clustering techniques using percentage of species relative abundance and frequency of occurrence as selection criteria (He et al. 1997, Hazin and Erzini 2008). For the GOA region, cluster analysis showed six clusters based on the dominance of three species categories (Pacific halibut, grenadier, and sablefish). For the BSAI region, cluster analysis showed six clusters based on the dominance of five species categories (Pacific halibut, grenadier, Pacific cod, Greenland turbot, and sablefish). After examining collinearity between continuous variables, a strong negative correlation (r = -0.84) was found only between Pacific cod and depth. Therefore, the Pacific cod CPUE variable was omitted and depth was retained for subsequent analyses.

Final Data Set -- After processing the data, the GOA data had < 1% of records with CPUE = 0 while the BSAI data had ~15% records with CPUE = 0, implying zero inflation was not an issue in the GOA and was only a moderate situation in the BSAI region. Therefore, we did not consider zero-inflated models for either region. Because CPUE data were positively skewed and continuous, CPUE values were defined as the logarithm of catch in kg +1 per 1,000 hooks. The explanatory variables that remained for fitting were latitude and longitude at the end of the set, Julian date (day of year), year, vessel size, bottom depth, and CPUE of grenadier, halibut, and Greenland turbot. Due to differences in the magnitudes of abundance among subregions and potential differences between predictors among areas (Hanselman et al. 2010), standardization of CPUE and calculation of abundance indices were conducted for individual subregions. Subregions within the GOA are West Yakutat, Eastern Gulf and Southeast Outside, Central Gulf, and Western Gulf; subregions within BSAI are Bering Sea and Aleutian Islands.

Statistical Models

Selection of useful explanatory variables for subregion models was done by Akaike Information Criteria (AIC) using the “dredge” function in the “MuMIn” package using R (R Development Core Team 2012). This function compared all possible GAM models of the explanatory variables in the filtered database (interactions were not explored). After completing this procedure, all explanatory variables that were selected by the dredge function within each subregion were used in all models to maintain comparability among methods except for WY where AIC suggested excluding latitude. The GLM, GAM, and BRT full models were fitted with Gaussian distribution errors. Parameters of the GLM and GAM models were obtained by optimizing maximum likelihood estimates from resulting iterations of penalized least squares (Wood 2006). For the GAM model, default settings were used to fit the data. For the BRT model, the tree complexity was set at 1 to analyze main effects, learning rate was 0.01, and bag fraction was 0.5 (Elith et al. 2008). Models and subsequent analyses were developed in R (R Development Core Team 2012) using the packages “MASS”, “mgcv”, and “gbm” with code modified from Abeare (2009) and Elith et al. (2008).

The full models for subsequent analysis by subregion are as follow. All continuous predictors are italicized.

GOA Subregion West Yakutat

ln(sablefish CPUE+1) = Year + *Julian date* + *Longitude* + *Bottom Depth* + *Vessel Size* + ln(*Grenadier* CPUE+1) + ln(*Halibut* CPUE+1).

GOA Subregions Western Gulf, Central Gulf, and Southeast

ln (sablefish CPUE+1) = Year + *Julian date* + *Latitude* + *Longitude* + *Bottom Depth* + *Vessel Size* + ln (*Grenadier* CPUE+1) + ln (*Halibut* CPUE+1).

BSAI Subregions Bering Sea and Aleutian Islands

ln (sablefish CPUE+1) = Year + *Julian date* + *Latitude* + *Longitude* (end of fishing trip) + *Bottom Depth* + *Vessel Size* + ln (*Grenadier* CPUE+1) + ln (*Halibut* CPUE+1) + ln(*Turbot* CPUE+1).

Model Selection Strategy

The principal goal of the study was to select the best model that could be used for subsequent prediction of sablefish abundance. In principle, the vast majority of studies using GLM and GAM approaches have utilized information criteria theory based on maximum likelihood (Akaike 1974) for model selection. However, given that boosted regression trees methods do not use maximum likelihood approaches (Dea'th 2007, Elith et al. 2008), a different strategy was used to infer the best model framework.

Our model selection strategy was divided into two parts. The first part involved model training and cross-validation of the models. The criteria used were based on how the models met assumptions of distributional errors and spatial independence, examination of marginal effects on the response variable, percentage of deviance explained, and measures of predictive performance (Root Mean Square Error (RMSE)).

Once the best model(s) were selected using these criteria, the second part of the strategy was to compare the standardized model's CPUE to the nominal CPUE values and infer trends among subregions using the selected models in order to see if the models are realistically describing CPUE trends and what factors are causing discrepancies among the models and the observed CPUE. Most CPUE standardization studies have concentrated on the removal of the effects of predictors in order to obtain to an unbiased index of abundance. However, there are few studies focusing on understanding differences in standardized and unstandardized CPUEs (e.g., Bentley et al., 2012). For communicating the value of standardization to stakeholders, it is important to determine which variables prevent unstandardized CPUEs from being a reliable measure of abundance. We examined the residual differences between the modeled CPUE predictions and nominal CPUE to determine when the selected models were following nominal CPUE trends and when they differed.

Changes in patterns of yearly abundance indices by adding each explanatory variables one at a time (step plots) were used to visualize how each explanatory variable contributed to the differences between the standardized and unstandardized nominal CPUE (Bentley et al. 2012). By examining the CPUE changes of adding each predictor, inferences can be made on which variables have the most influence on the CPUE trend by looking at the model CPUE trend from a particular predictor that is the furthest away from the trend that only contains year as the predictor variable. If the trend of a particular predictor is close to the one that has year as a predictor variable it means that the variable has little effect on explaining differences between standardized and unstandardized CPUE.

Influence index plots were also used to quantify how much a predictor variable can contribute to differences in CPUE patterns of standardized and unstandardized values. This method is commonly used in GLM and GAM models only (Bentley et al. 2012). If the influence index of a variable is >1 it means that the inclusion of that variable increased the estimate of CPUE in that year. If the influence index is < 1, the inclusion of that variable decreased the estimate of CPUE in that year. If the influence is one, this means that the variable had no influence in that particular year.

Model Comparisons

Quantile-quantile (Q-Q) plots and residual distribution of the model fits were inspected to ensure that the residuals met the assumption of normality. We constructed semivariogram plots of the residuals of each model to examine how well spatial autocorrelation was accounted for using the "GeoR" package in R (Ribeiro and Diggle 2001). To assess how well the models fitted the data, percent of deviance explained was used. Percent deviance explained (pseudo-R^2) was calculated with the formula 1-(residual deviance/total deviance). To assess the relative importance of different explanatory variables in GLM and GAM models, relative variable importance was calculated by examining changes in improvement of AIC by examining the addition of one variable at a time (CPUE~Year+ βX_1). We used this approach because the relative importance of explanatory variables to the model fit can change based on the order the variables are entered into the model. In BRT models, the relative influence of explanatory variables was calculated by how many times that variable was selected for splitting and averaged over all trees compared to the squared improvement to the model fit (Elith et al. 2008). Then, the contribution of each variable is scaled to a total of 100.

A 5-fold cross-validation procedure (Shono 2008, Carvallho et al. 2011, Li et al. 2011, Hazin et al. 2011) was used to evaluate and validate the predictive power of the models. A stratified random sampling approach using years as strata was used in the 5-fold cross-validation procedure to retain the basic structure of the data set because the year effect was the variable of interest. We examined mean values of pseudo-R^2 (percent of deviance explained) and root mean square error (RMSE) from different combinations of training and test sets from the 5-fold cross-validation procedure (for training data n = 5; for test set data n = 20). To investigate differences between nominal mean CPUE and predicted values of each cross-validation model, paired t-tests were used. To examine significant differences in variance among cross-validation models, F-tests were used. These analyses were done on the predicted values of the combined fits of the 5-fold cross-validation models.

In order to calculate yearly CPUE indices across subregions, we made model CPUE predictions using a new data set that contained all the explanatory continuous variables held constant at their means and only varying the year. Error estimates of the abundance indices for the extracted year term are lacking in the BRT, which is a significant shortcoming of this modeling method (Abeare 2009). Thus, bootstrapping was used to obtain error estimates for the models selected (Efron and Tibshirani 1993). Data were randomly sample with replacement by year and the indices were recalculated 100 times for both GAMs and BRTs.

RESULTS

Model Assumptions

Inspections of model assumptions and diagnostics were conducted for all subregions. Since the results were similar among regions, we use the Central Gulf subregion as a representative example because of its importance to the fishery. Quantile-quantile plots of the GLM fit clearly showed that the residuals were non-normal. Inspections of major effects in partial plots on the GAM and BRT models showed that important explanatory variables such as depth were consistently nonlinear on the Central Gulf subregion. Serious violations of the assumptions of linearity and normality can invalidate the results of GLM models (Appendix Figs. 1-5). Therefore, it was clear GLM was inadequate for modeling the fishery CPUE data, and we focus on results from GAM and BRT for the rest of the analysis.

Box plots showed that the distribution of residuals was similar across models (Appendix Fig. 6). The semivariogram plots revealed notable autocorrelation on a spatial scale around 0.2-0.5 degrees of longitude in Central Gulf, which is approximately 25-50 km; this scale was consistent among all subregions (Appendix Fig.7). Spatial autocorrelation was nearly eliminated by the GAM and BRT models, as evidenced by the flattening of the curves in the semivariogram plot. Variance of the predictions from the GAM and BRT models was much lower than the variance of nominal CPUE (Appendix Fig. 7).

Percentage of Deviance Explained (pseudo-R^2) and Variable Relative Importance

Overall, percentage of deviance explained (pseudo-R^2) ranged from 31 to 47% for the BRT model and were slightly higher than those from the GAM models in the WY, WG, and SE subregions. This indicates that the BRT model fitted the data better than GAM. In the CG, the pseudo-R^2 was similar among models (Table 1). For the WY, CG, and SE subregions, the most influential variables that were consistent across models were bottom depth and Pacific halibut CPUE, based on the rankings of change in AIC improvement for the GAM model and the calculated relative variable influence for the BRT model (Appendix Table 1, Appendix Fig. 8).

For the BSAI subregions, the BRT model pseudo-R^2 were slightly higher than the GAM models in the AI, and BS subregions ranging from 53 to 68% of the percentage of deviance explained (Table 1), indicating that the BRT was superior to GAM in fitting the data. For the BS and AI subregions, the most influential variables that were consistent across models were longitude and latitude (Appendix Table 2, Appendix Fig. 9).

Table 1. -- Percentage of deviance explained by GAM and BRT models for GOA and BSAI subregions.

	Subregions		Percent of Deviance Explained
GOA	West Yakutat	GAM	42.30
		BRT	47.10
	Western Gulf	GAM	28.90
		BRT	31.40
	Central Gulf	GAM	30.30
		BRT	30.00
	Southeast	GAM	44.60
		BRT	45.56
BSAI	Bering Sea	GAM	60.50
		BRT	63.10
	Aleutian Islands	GAM	53.60
		BRT	58.40

Results from the cross-validation procedure where the data were split into five equal parts (5-fold) showed that GAM had a slightly higher mean percentage of deviance explained than the BRT model in the GOA subregions (Table 2). However, both had comparable predictive power (similar RMSE errors). For the BSAI subregions, the BRT model had higher mean percentage of deviance explained and slightly lower RMSE, than GAM. There was no significant differences ($P > 0.05$; paired t-test)) between nominal CPUE and predicted CPUE for GAM, and BRT within subregions (Table 3). However, the variance for GAM models was significantly higher ($P < 0.001$; F-tests) than BRT (Fig. 3, Table 4).

Table 2. -- Model performance results for the GAM, and BRT methods across subregions for the GOA and BSAI regions using 5-fold cross-validation on Mean Percent of Deviance explained and Root Mean Square Error (RMSE). Data are averages of each parameter from different settings of training and test sets from the 5-fold procedure. For training data n = 5, for test set data n = 20. S.E. stands for standard error.

	Subregions		Mean Percent of Deviance Explained	S.E.	Mean RMSE	S.E.
GOA	West Yakutat	GAM	50.98	4.44	0.27	0.01
		BRT	43.20	9.77	0.26	0.01
	Western Gulf	GAM	30.94	4.22	0.43	0.01
		BRT	24.69	6.11	0.43	0.01
	Central Gulf	GAM	33.34	6.70	0.44	0.01
		BRT	30.84	5.84	0.43	0.01
	Southeast	GAM	53.02	10.77	0.28	0.02
		BRT	36.52	22.21	0.27	0.01
BSAI	Bering Sea	GAM	60.02	2.01	0.53	0.05
		BRT	61.20	3.44	0.49	0.01
	Aleutian Islands	GAM	55.86	2.67	0.54	0.03
		BRT	58.68	4.74	0.52	0.02

Table 3. -- Comparisons of mean predicted values of GAM and BRT models against mean nominal CPUE using paired t-tests for each subregion.

		Subregion	df	t Stat	P(T<=t) one-tail
GOA	West Yakutat	GAM	9,159	-0.13	0.45
		BRT	9,159	-0.61	0.27
	Western Gulf	GAM	18,895	0.14	0.44
		BRT	18,895	0.39	0.35
	Central Gulf	GAM	19,847	-0.89	0.19
		BRT	19,847	-0.26	0.40
	Southeast	GAM	5,535	-0.47	0.32
		BRT	5,535	0.77	0.22
BSAI	Bering Sea	GAM	11,835	0.01	0.50
		BRT	11,835	-1.02	0.15
	Aleutian Islands	GAM	15,791	-0.25	0.40
		BRT	15,791	-0.54	0.30

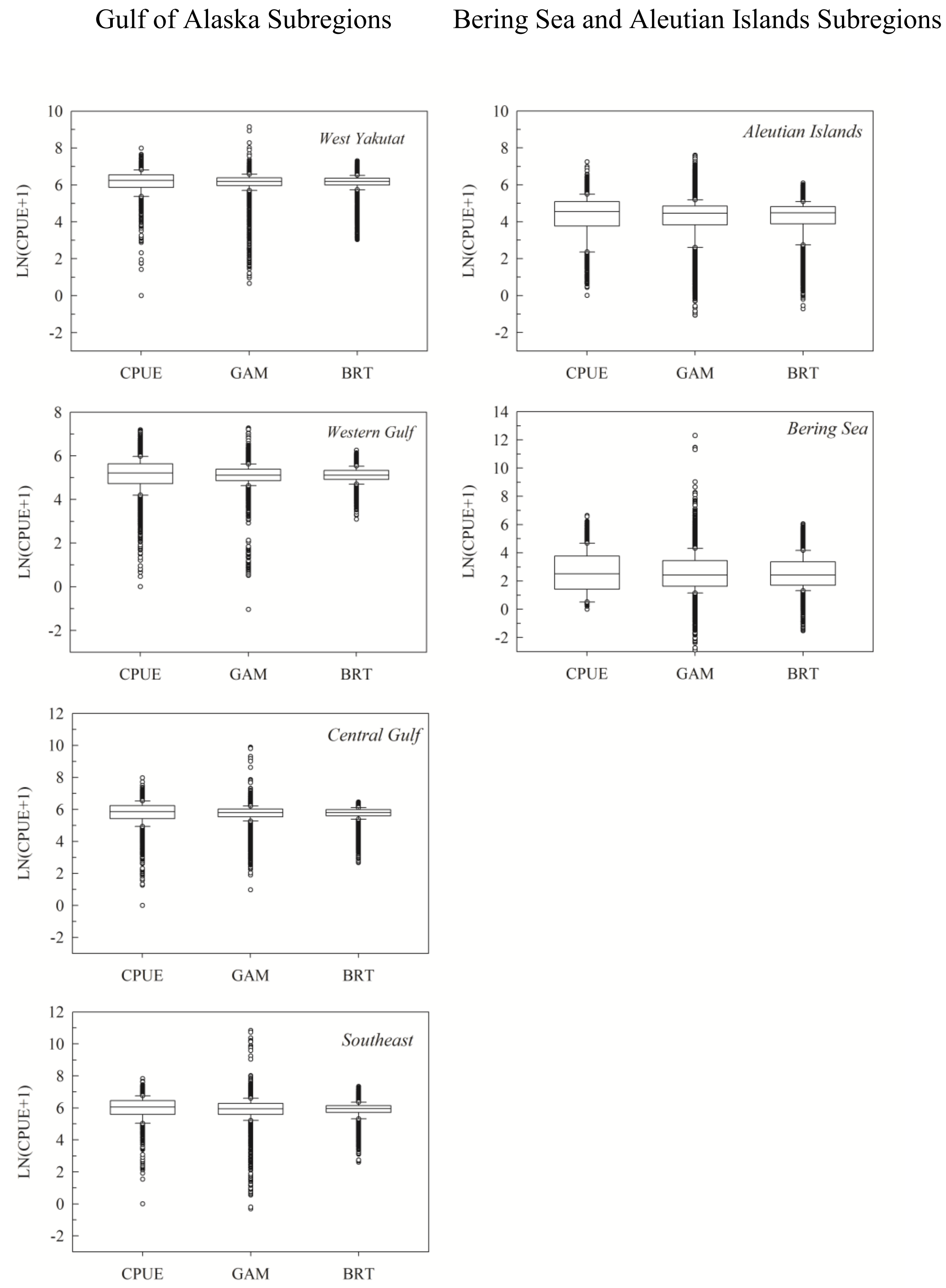


Figure 3. -- Boxplots of observed LN(CPUE +1) values from 5-fold cross-validation sets against predicted values from the models for the Gulf of Alaska, and the Bering Sea and Aleutian Islands subregions. Predicted values were obtained from pooling all the predicted values from all the test sets (n = 20) within each region.

Table 4. -- Comparisons of variances obtained from GAM and BRT models using F-tests for variances for each subregion.

	Subregion	df	Ft	P(F<=f) one-tail
GOA	West Yakutat	9,159	1.39	<0.001
	Western Gulf	18,895	1.73	<0.001
	Central Gulf	19,847	1.40	<0.001
	Southeast	5,535	2.16	<0.001
BSAI	Bering Sea	11,835	1.35	<0.001
	Aleutian Islands	15,791	1.21	<0.001

Comparisons of Standardized Models with Nominal CPUE

Overall, BRT and GAM CPUE estimates had similar trends in residual differences from nominal CPUE but differed in some areas in the magnitude of the estimates (Appendix Fig.10). In all years, both nominal and model estimated CPUE were generally highest in the West Yakutat and Southeast subregions, and were lowest in the Bering Sea and Aleutian Islands (Fig. 4). Model estimated CPUE was similar to the annual nominal CPUE in the GOA subregions of West Yakutat and Western Gulf. The largest difference in trends among models were in Central Gulf, where GAM and BRT models showed similar trends relative to the nominal CPUE but with different magnitude. Southeast GAM and BRT models also had similar trends but there were differences in magnitude. When the GAM and BRT models were compared across subregions in the Bering Sea/Aleutian Islands region, it shows a slower decline in abundance compared to the nominal CPUE indices in the Bering Sea and decreased variability in the Aleutian Islands.

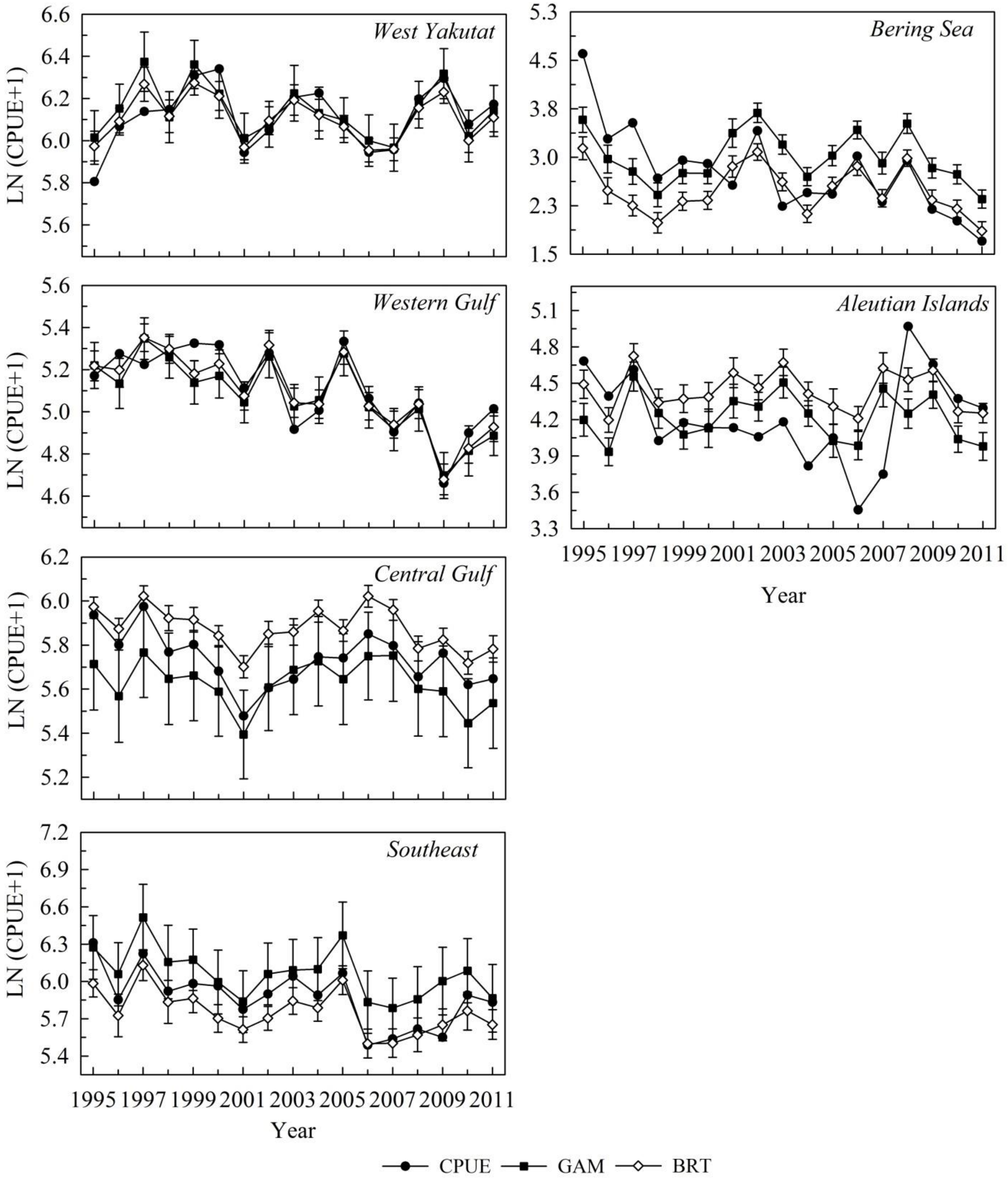


Figure 4. -- Comparisons of the unstandardized CPUE to standardized CPUE indices from GAM and BRT models for the Gulf of Alaska, and Bering Sea and Aleutian Islands subregions.

Exploration of the Standardization Effects of Predictors Influencing CPUE

Predictors that were shown to be important for contributing to differences between the nominal CPUE and model CPUE estimates differed by area and model. We examined estimated CPUE changes after adding each predictor by using step plots (adding one predictor at a time) (Appendix, Fig. 11). In the West Yakutat GAM model, longitude and grenadier CPUE were the most important predictors (Appendix Fig. 11). For the same area using the BRT model, longitude and Julian date were the most important predictors. For Western Gulf, in both models, Grenadier CPUE and Halibut CPUE were the most important predictors. Latitude and longitude were the most important predictors for the Central Gulf using the GAM model, while Grenadier CPUE and Halibut CPUE were the most important predictors in the BRT model. For the Southeast subregion, vessel size and halibut CPUE were the most important predictors for the GAM models, while longitude and latitude were the most important in the BRT models (Appendix Fig. 11). For the Bering Sea subregion, the step plots of the GAM model showed that latitude was important, whereas in the BRT model, Julian date, longitude, and latitude were the most influential predictors. For the Aleutian Islands, Halibut CPUE and vessel size were the most important predictor in the GAM model, whereas in the BRT model latitude, longitude, and vessel size were the most important (Appendix Fig. 12).

Analysis of influence plots for GAM models differed notably from the step plots within all subregions regarding what predictors were the most important in differences in standardized and unstandardized CPUE. In general, the influence plots placed more emphasis on Julian day, depth, and latitude; the step plots showed placed more importance on predictors that involved other species CPUE and longitude (Appendix Figs. 13-14).

DISCUSSION

In this study, the performance of GLM, GAM, and BRT statistical methods were evaluated to determine the most robust model for standardizing longline fishery CPUE as an index of abundance for Alaska sablefish. GAM and BRT models were more suitable for fitting the data than GLM. This was because GLMs cannot fit the nonlinear relationships that exist between response and some of the explanatory variables. For example, the three most important predictors in the main-effects models across sub regions were depth, latitude, and longitude, which all had non-linear relationships with sablefish CPUE. If GLMs were still preferred, an alternative for using the GLM would be to bin the nonlinear predictors into categorical variables, where the nonlinear patterns could be fit, but this would require many more parameters.

Both model types fit the data similarly in terms of amount of deviance explained, and diagnostic fits. A limitation of the BRT model is that it the variance of the predicted CPUE values is not straightforward. Since the BRT does not operate under a maximum likelihood framework, we used a bootstrap to attempt to compare the BRT and GAM modeled CPUE estimates directly. The GAM models had significantly higher bootstrap variances than BRT. However, bootstrapping also may not be a fair comparison because the low variance seen in the BRT predictions may be a result of the boosting algorithm reducing some of the variability in the bootstrap sample; the fitting algorithm is internally resampling the data.

In our study, the GAM model explained a higher average percentage of the deviance than the BRT in the GOA, but lower in the BS/AI. Prediction error was similar between the two model frameworks. A limitation of the GAM model is that their smoothing functions cannot effectively extrapolate predictions outside of the range of the training data that was used to build the model (Frescino et al 2001). For example, values of the test sets that are outside of the range of the training data would be assigned to the closest maximum and minimum values of the training data. Thus, there is a chance for increased uncertainty associated with extrapolation of the smoothed functions in the most extreme parts of the distribution (tails), which is not reflected in the BRT models. In a study that compared similar statistical approaches to standardize CPUE for a Yellowfin Tuna longline fishery, the GAM had a larger pseudo-R^2 (percentage of deviance explained) than GLM and BRT (Abeare 2009), but was no better at predictive performance, which was similar to our results here.

In general, the standardized indices of abundance from the two models studied here, suggested a gradual declining trend in sablefish CPUE for the Western Gulf subregion and more pronounced recent declines in the Central Gulf and Bering Sea. Other areas showed little trend. These results are consistent with the relative indices of abundance from the longline survey and the stock assessment (Hanselman et al. 2012). There were notable differences in magnitudes of CPUE rates among models across some of the subregions. GAM CPUE values were relatively higher than BRT in Southeast while BRT CPUE values in Central Gulf were higher than GAM models. These two subregions were also the ones with more uncertainty compared to other regions based on the magnitudes of the standard errors (Fig. 4). Visual step plots of the two models among subregions indicated that, except for the Western Gulf and West Yakutat regions, each model within each subregion had different explanatory predictors responsible for the differences in patterns of standardized and unstandardized CPUEs. These results could be attributed to GAM and BRT models differing significantly in their statistical properties. Thus, each variable would be weighted differently in each model. However, Bentley et al. (2012) warns that interpretation of step plots to infer why differences exists between unstandardized and standardized CPUE trends should be treated with caution as they only show incremental changes in CPUE but do not estimate the relative influence on the final model. GAM influence plots, in contrast, indicated that the predictors influencing the differences among the standardized and unstandardized CPUE were consistent with the predictors selected by the one-variable AIC models.

In summary, the utilization of statistical approaches such as GAM and BRT, which deal better with nonlinearity of predictors and spatial autocorrelation than GLMs, should be considered for subsequent CPUE standardization. Furthermore, BRT model performance in some situations was superior or equally robust to GAM. A more rigorous approach such as simulations and sensitivity analysis should be used to the test the robustness of these models. Studies including simulation analyses are invaluable for comparing model performance under different potential abundance trends and violations of assumptions (Lynch et al. 2012). The underlying abundance index used in the operating model for simulations could be used to test the effects of sample size, model misspecification, data with different proportion of zeros on models predictive power, as well as to the test if the chosen model can robustly account for hyperstability, hyperdepeletion, and spatial and temporal changes in fleet dynamics.

The results of this study contribute to sablefish fishery research by recommending potential statistical approaches for standardizing CPUE when data may be affected by spatial, temporal, or environmental factors that often have nonlinear relationships with CPUE. These methods could be reproduced efficiently and used to examine the spatial/temporal dynamics of fishing activities in other marine ecosystems with different types of gear and species. Fish population dynamics models and stock assessments depend heavily on reliable estimates of abundance. For fishery CPUE to be used with confidence, an accurate CPUE standardization model is needed. Improvement of data quality and continued evaluation of model performance should be given priority in order to provide better recommendations for management and conservation.

ACKNOWLEDGMENTS

The authors like to thank AKFIN for providing us the Gulf of Alaska and Bering Sea observer longline fishery observer data. Special thanks go to Drs. F. Mueter, J. Heifetz, P. Hulson, B. Li, Y. Chen and S. Wood; and to C.R. Lunsford, C. Rodgveller, and N. Cummings for their helpful comments with this manuscript. The primary author would like to thank J. Thornhill and S. Maclean from the National Research Council for providing funds and assistance during his time at the Alaska Fisheries Science Center's Auke Bay Laboratories in Juneau, Alaska. Finally, we would like to thank all the observers and fishermen involved in the collection of the data.

CITATIONS

Abeare, S. M. 2009. Comparisons of boosted regression trees, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico longline fishery. Master's Thesis, Louisiana State University, Baton Rouge, LA 85 pages.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Auto. Cont. 19: 716-723.

Bentley, N., T. H. Kendrick, P. J Starr, and P. A. Breen. 2012. Influence plots and metrics: tools for better understanding fisheries catch-per-unit-effort standardisations. ICES J. Mar. Sci. 69: 84-88.

Callahan, J. 2010. At –sea monitoring of commercial North Pacific groundfish catches: A range of observer sampling challenges Alaska Fisheries Science Center Quarterly Report: research feature July –August-September 2010, p. 2-5.

Carvalho, F. C., D. J. Murie, F. H. V. Hazin, H. G. Hazin, B. Leite-Mourato, and G. H. Burgess. 2011. Spatial predictions of blue shark (*Prionace glauca*) catch rate and catch probability of juveniles in the Southwest Atlantic. ICES J. Mar. Sci. 68(5):890–900.

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

Efron, B. T. and R. J. Tibshirani. 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability. Chapman and Hall (Book 57) 456 p.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77:802-813.

Erisman, B. E, L. G. Allen, J. T. Claisse, D. J Pondella, E. F Miller, and J. H. Murray. 2011. The illusion of plenty: hyperstability masks collapses in two recreational fisheries that target fish spawning aggregations. Can. J. Fish. Aquat. Sci. 68:1705-1716.

Frescino, T. S., T. C Edwards, and G. G. Moisen. 2001. Modelling spatially explicit forest structural attributes using Generalized Additive Models. J. Veg. Sci. 12: 15 -26.

Froeschke, J. T., G. W Stunz , and M. W. Wildhaber. 2010. Environmental influences on the occurrence of coastal sharks in estuarine waters. Mar. Ecol. Prog. Ser. 407:279–292.

Hanselman, D. H., C. R. Lunsford, and C. Rodgveller. 2008. Assessment of the sablefish stock in Alaska. *In* Stock assessment and fishery evaluation report for the groundfish resources of the GOA and BS/AI as projected for 2009. North Pacific Fishery Management Council, 605 W. 4th Ave, Suite 306 Anchorage, AK 99501.

Hanselman, D. H., C. R. Lunsford, and C. Rodgveller. 2010. Assessment of the sablefish stock in Alaska. *In* Stock assessment and fishery evaluation report for the groundfish resources of the GOA and BS/AI as projected for 2011. North Pacific Fishery Management Council, 605 W. 4th Ave, Suite 306 Anchorage, AK 99501.

Hanselman, D. H., C. R. Lunsford, and C. Rodgveller. 2012. Assessment of the sablefish stock in Alaska. *In* Stock assessment and fishery evaluation report for the groundfish resources of the GOA and BS/AI as projected for 2013. North Pacific Fishery Management Council, 605 W 4th Ave, Suite 306 Anchorage, AK 99501.

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

Hazin, H., and K. Erzini. 2008. Assessing swordfish distributions in the South Atlantic from spatial predictions. Fish. Res. 90(1-3):45-55.

Hazin, H. G., T. Fredou, F. Hazin, and P. Travassos. 2011. Standardized CPUE series of bigeye tuna, *Thunnus obesus*, caught by Brazilian tuna longline fisheries in the Southwestern Atlantic Ocean (1980-2008). ICCAT Col. Vol. Sci. Pap. (1). 66:387-398.

He, X., K. A. Bigelow, and C. H. Boggs. 1997. Cluster analysis of longline sets and fishing strategies within the Hawaii-based fishery. Fish. Res. 31(1):147-158.

Hilborn, R., and C. J. Walters. 1992. Quantitative fisheries stock assessment: choice, dynamics, and uncertainty. Chapman and Hall, London.

Hinton, M. G., and M. N. Maunder. 2004. Methods for standardizing CPUE and how to select among them. ICCAT Col. Vol Sci. Pap. 56(1):169-177.

Li, Y., Y. Jiao, and Q. He. 2011. Decreasing uncertainty in catch rate analyses using Delta-AdaBoost: An alternative approach in catch and bycatch analyses with high percentage of zeros. Fish. Res. 107:261-271.

Lynch, P. D., K. W Shertzer and R. J. Latour. 2012. An evaluation of methods for standardizing catch rates of highly migratory species. ICCAT Col. Vol Sci. Pap. 68:1498-1509.

Martínez-Rincón, R. O., S. Ortega-García, and J. G. Vaca-Rodríguez. 2012. Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (*Acanthocybium solandri)* in the Mexican tuna purse-seine fishery. Ecol. Model. 233:20-25.

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

Pittman, S. J., M. B. Costa, and T. A. Battista. 2009. Using lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals. J. Coast. Res. S1:27–38.

Quinn, T. J., II, and R. B. Deriso. 1999. Quantitative fish dynamics. New York: Oxford University Press. 542 p.

R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. [Available from http://www.R-project.org, accessed Jul. 2012]

Ribeiro, J. R., and P. J. Diggle. 2001. GeoR: A package for geostatistical analysis. R-News Vol 1, No 2. ISSN 1609-3631.

Sigler, M. F., and C. R. Lunsford. 2001. Effects of individual quotas on catching efficiency and spawning potential in the Alaska sablefish fishery. Can. J. Fish. Aquat. Sci. 58: 1300-1312.

Shono, H. 2008. Application of the Tweedie distribution to zero-catch data in CPUE analysis. Fish. Res. 93:154-162.

van der Lee, A. 2012. Fleet dynamics around a seasonal regulatory closure on the Scotian shelf. M.S. thesis University of Manitoba, Manitoba , Canada. 1,101 p.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S-plus. Fourth edition. Springer-Verlag, New York.

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

Wood, S. N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC. 410 p.

APPENDIX

Appendix Table 1. -- Ranking importance of explanatory variables by improvement on Δ AIC calculated from GAM models on GOA subregions.

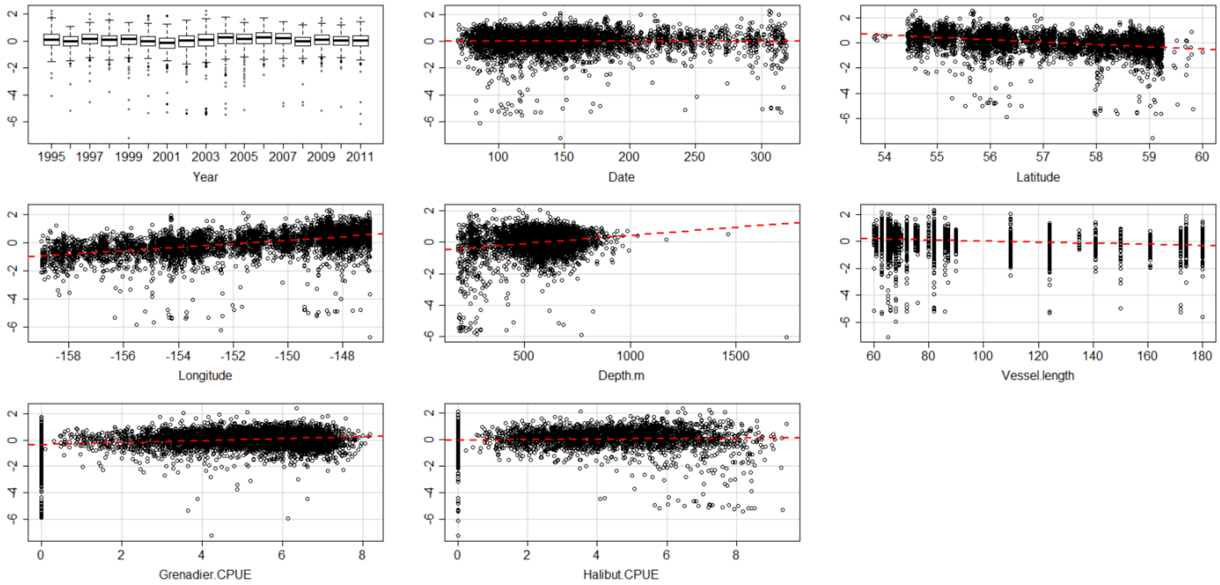
	West Yakutat			Western Gulf		
	AIC	ΔAIC	Rank	AIC	ΔAIC	Rank
Null	5,408.07			11,752.43		
Year	5,369.02	39.05	7	11,547.08	205.35	4
Julian date	5,239.96	168.11	5	11,611.84	140.59	7
Latitude			6	11,582.29	170.14	6
Longitude	5,251.40	156.68	6	11,363.83	388.60	2
Grenadier CPUE	5,152.82	255.26	3	11,553.69	198.74	5
Halibut CPUE	5,021.95	386.12	2	11,665.13	87.30	8
Vessel size	5,300.48	107.60	4	11,257.21	495.22	1
Depth	4,667.78	740.29	1	11,512.16	240.27	3

	Central Gulf			Southeast		
Central Gulf	AIC	ΔAIC	Rank	AIC	ΔAIC	Rank
Null	12,242.14			3,936.40		
Year	12,162.41	79.73	8	3,903.86	32.55	6
Julian date	12,150.96	91.18	7	3,831.29	105.11	4
Latitude	12,050.48	191.66	6	3,930.57	5.84	8
Longitude	12,043.58	198.56	5	3,925.24	11.17	7
Grenadier CPUE	11,663.87	578.27	2	3,802.14	134.27	3
Halibut CPUE	11,896.76	345.38	3	3,747.51	188.90	2
Vessel size	11,994.77	247.37	4	3,901.70	34.70	5
Depth	11,592.67	649.47	1	3,554.49	381.91	1

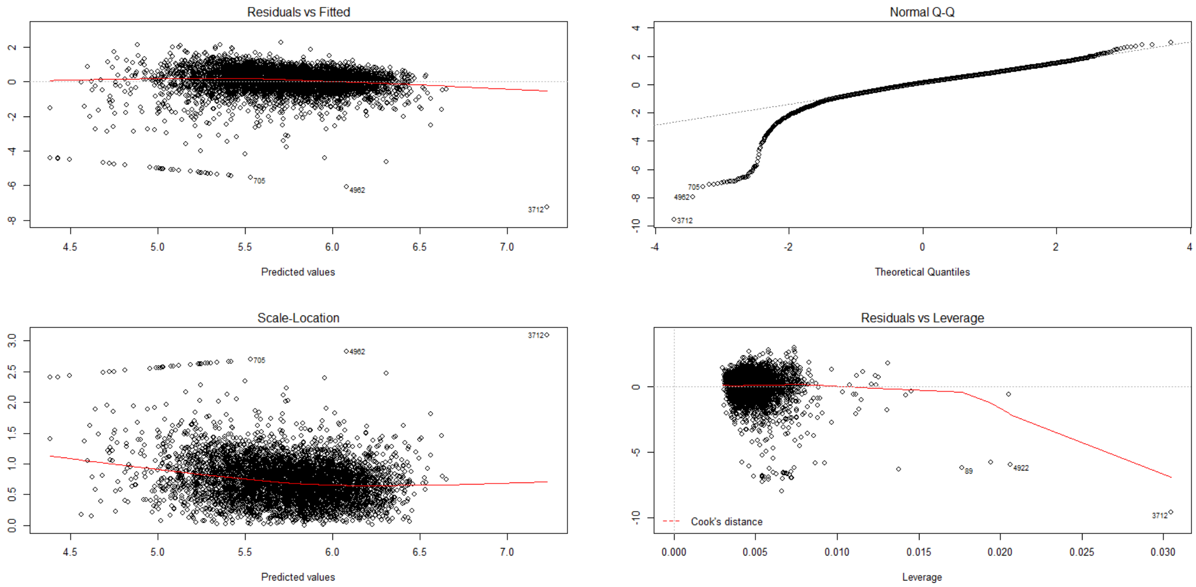
Appendix Table 2. -- Ranking importance of explanatory variables by improvement on Δ AIC calculated from GAM models on BSAI subregions.

	Aleutian Islands			Bering Sea		
	AIC	ΔAIC	Ranking	AIC	ΔAIC	Ranking
Null	13,704.59			10898.80		
Year	13,466.13	238.46	8	10507.07	391.73	3
Julian date	13,290.20	414.39	6	10662.96	235.84	6
Latitude	12,963.50	741.09	3	9669.242	1,229.558	2
Longitude	12,649.76	1054.83	1	9592.06	1,306.74	1
Grenadier CPUE	13,236.09	468.50	5	10707.09	191.71	7
Halibut CPUE	13,629.85	74.74	9	10639.10	259.70	5
Turbot CPUE	13,446.10	258.49	7	10773.52	125.28	9
Vessel size	13,005.82	698.77	4	10600.94	297.86	4
Depth	12,676.04	1028.55	2	10734.11	164.69	8

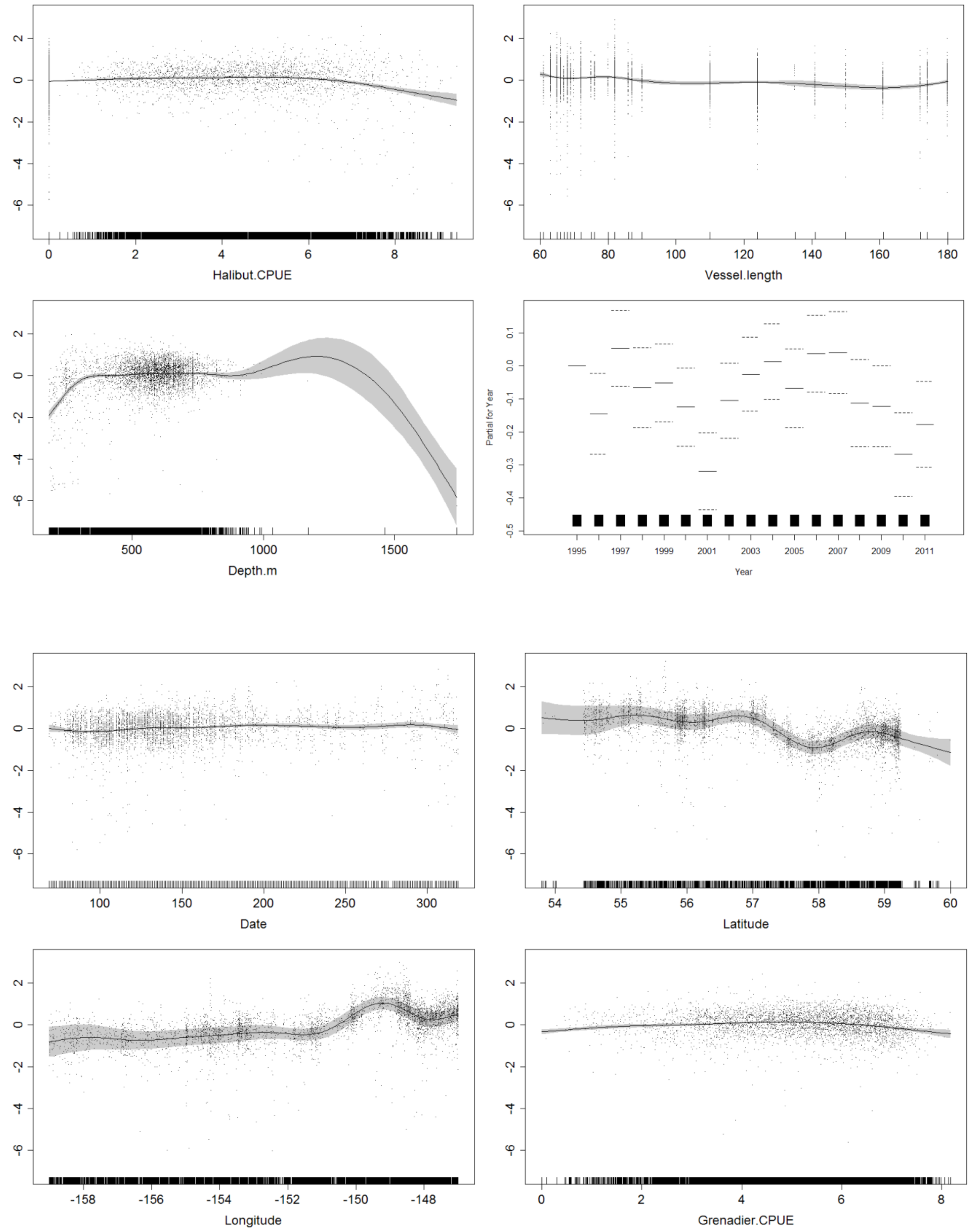
Appendix Figure 1. -- GLM Analysis of the main effects of eight predictor variables in the Central Gulf subregion.



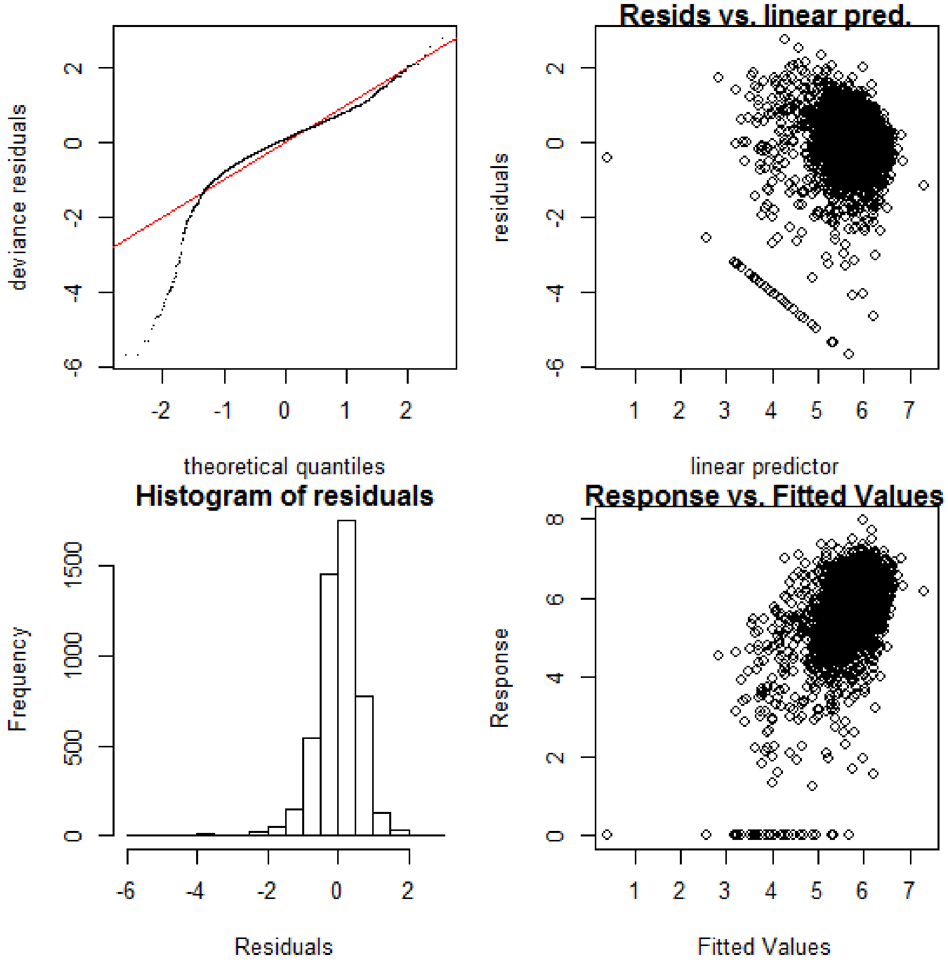
Appendix Figure 2. -- Diagnostics plots of goodness of fit of GLM model for the Central Gulf subregion.



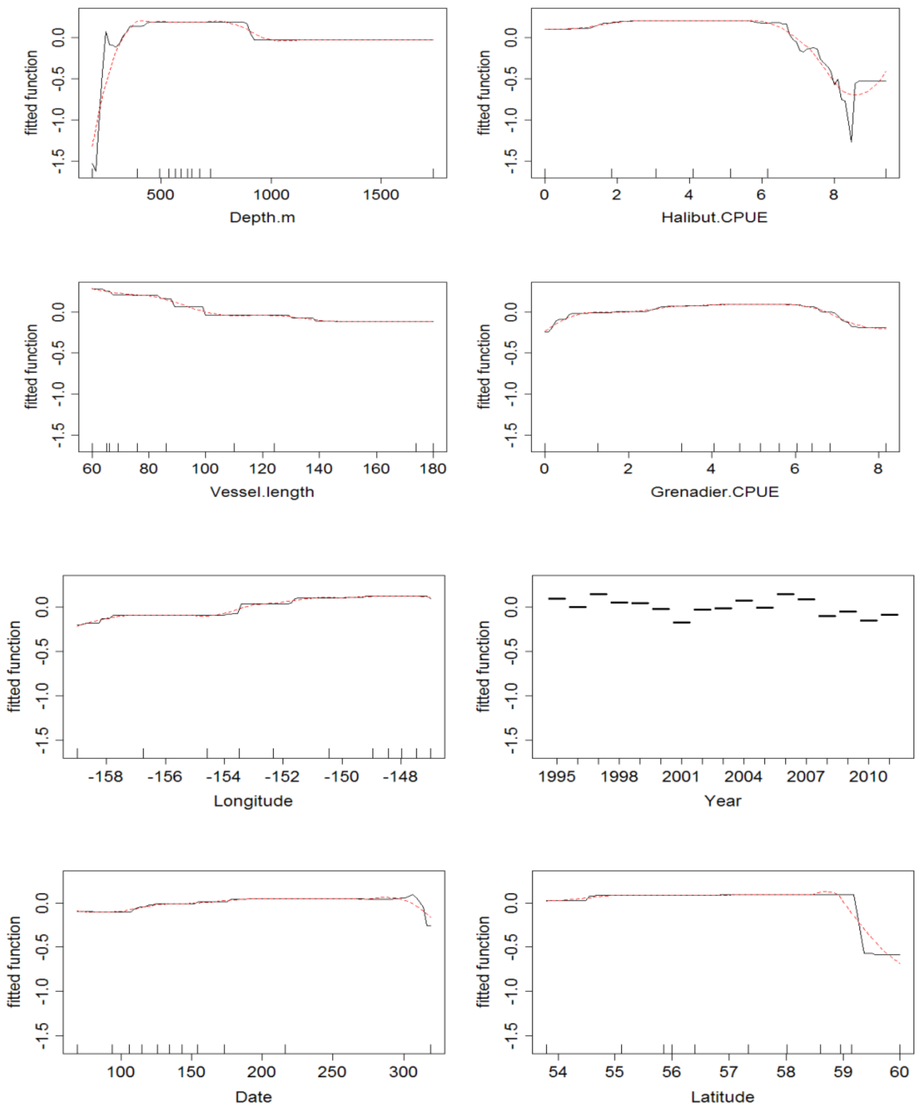
Appendix Figure 3. -- GAM Analysis of the main effects of eight predictor variables in the Central Gulf subregion.



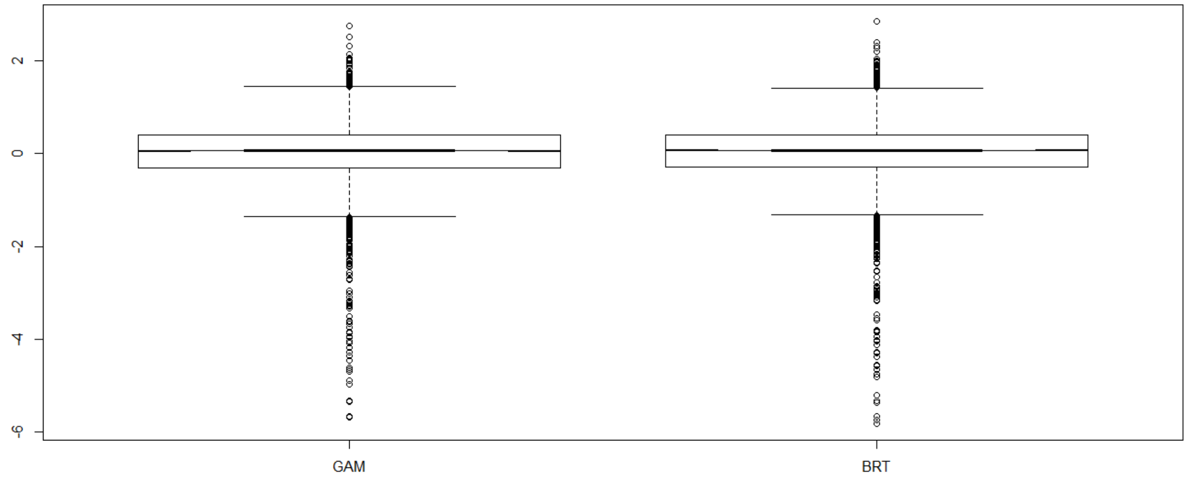
Appendix Figure 4. -- Diagnostics plots of goodness of fit of GAM model for the Central Gulf subregion.



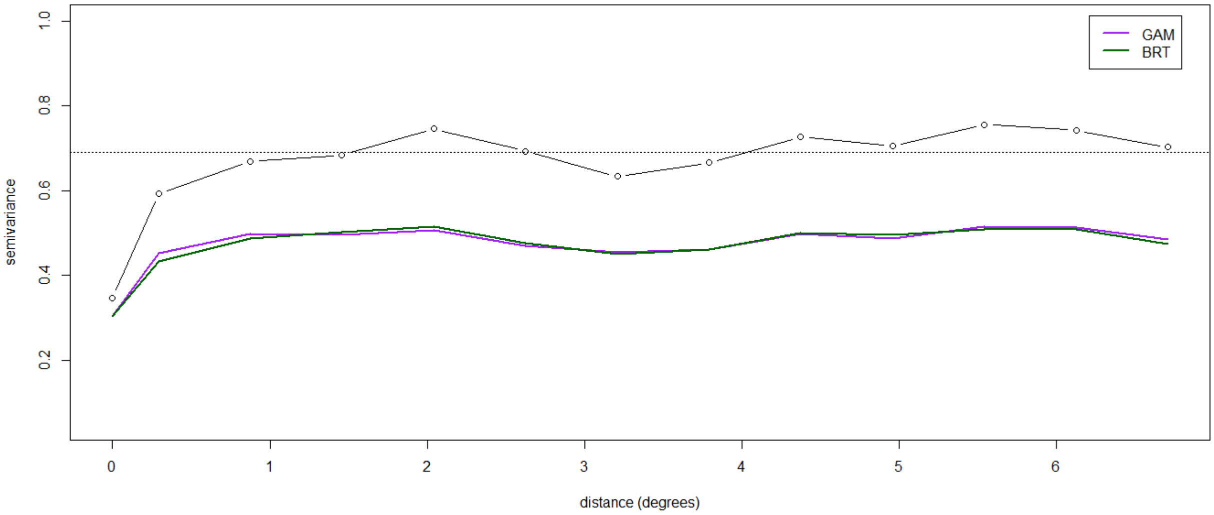
Appendix Figure 5. -- Fitted functions for main effects in the BRT model for the Central Gulf Subregion.



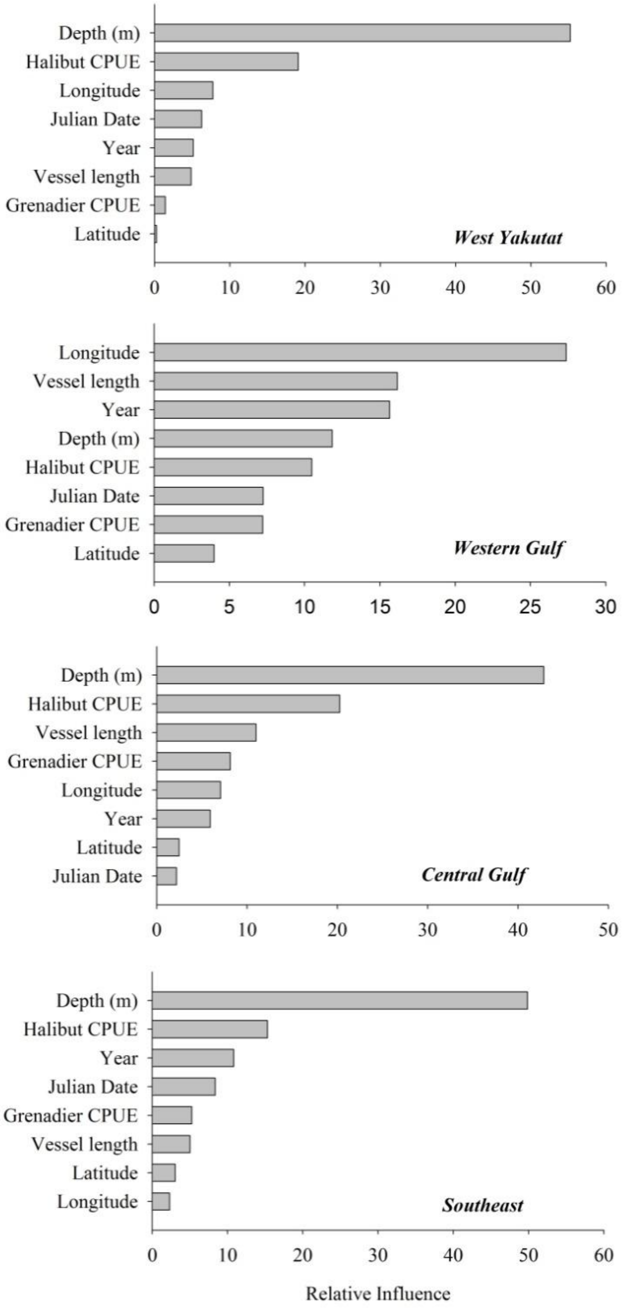
Appendix Figure 6. -- Analysis of residuals distribution among GAM, and BRT models for the central Gulf of Alaska using box plots. The notches on the box plots are medians.



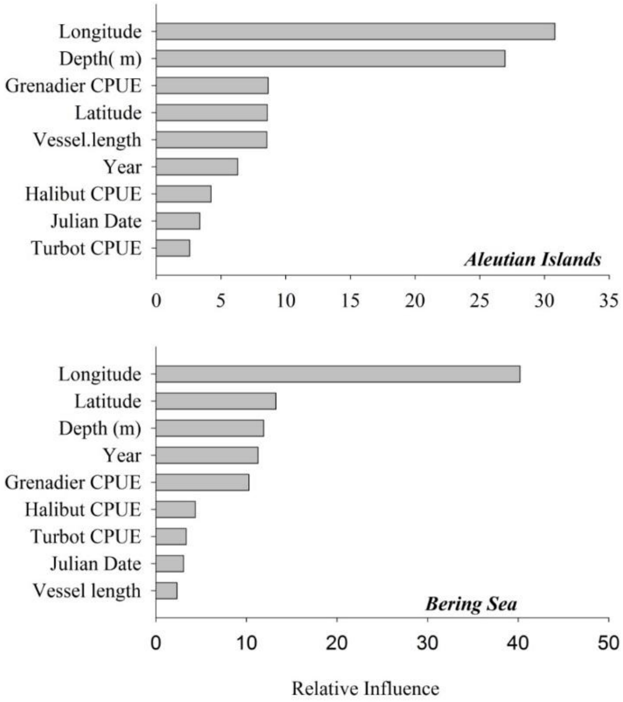
Appendix Figure 7. -- Semivariograms of residuals from GAM and BRT models for the Central Gulf of Alaska. The semivariogram sill for CG (sample variance (s^2)) is set at 0.742. Data are scaled in degrees.



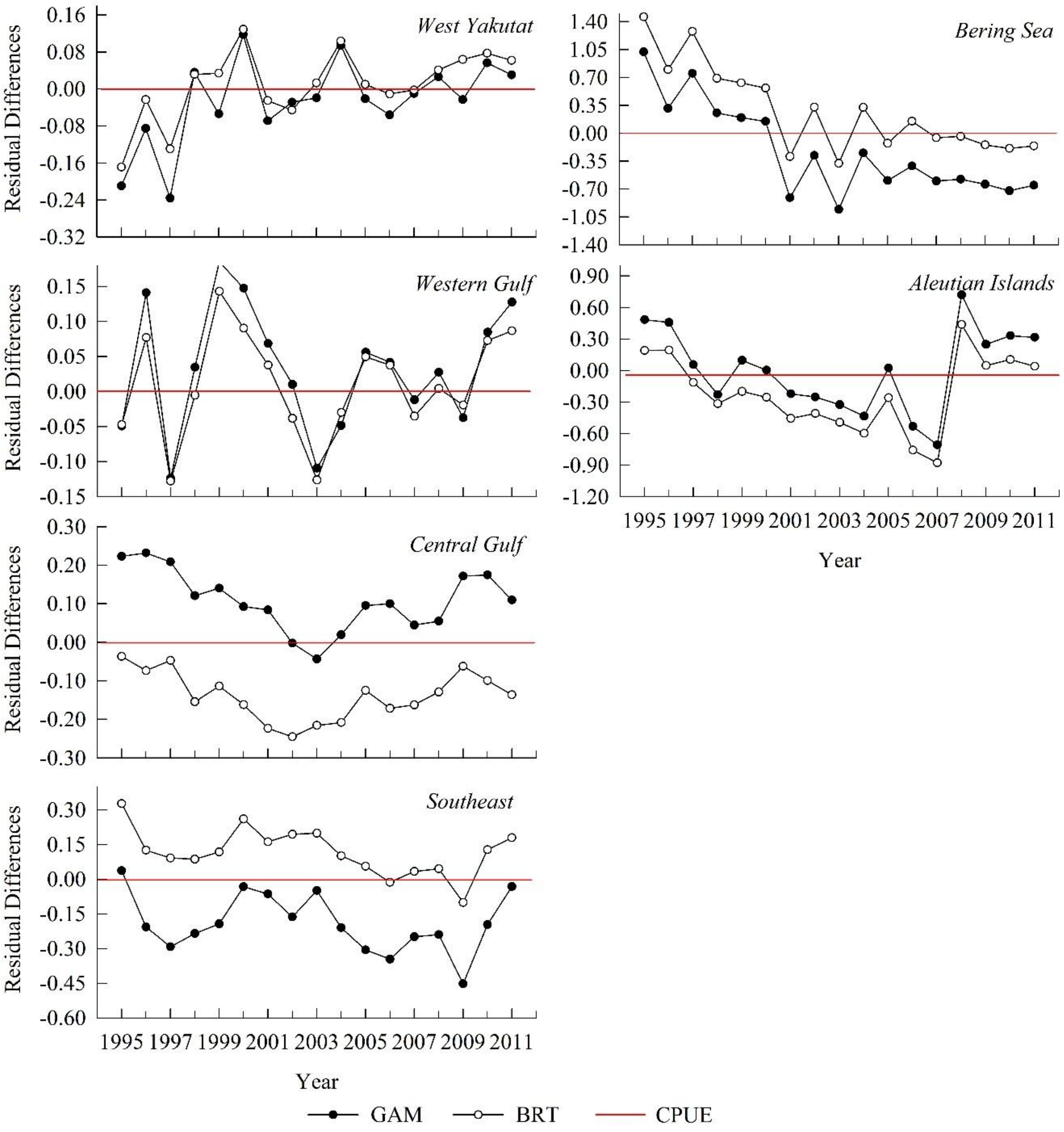
Appendix Figure 8. -- Summary of percent relative contribution of predictors on the BRT model for the Gulf of Alaska subregions.



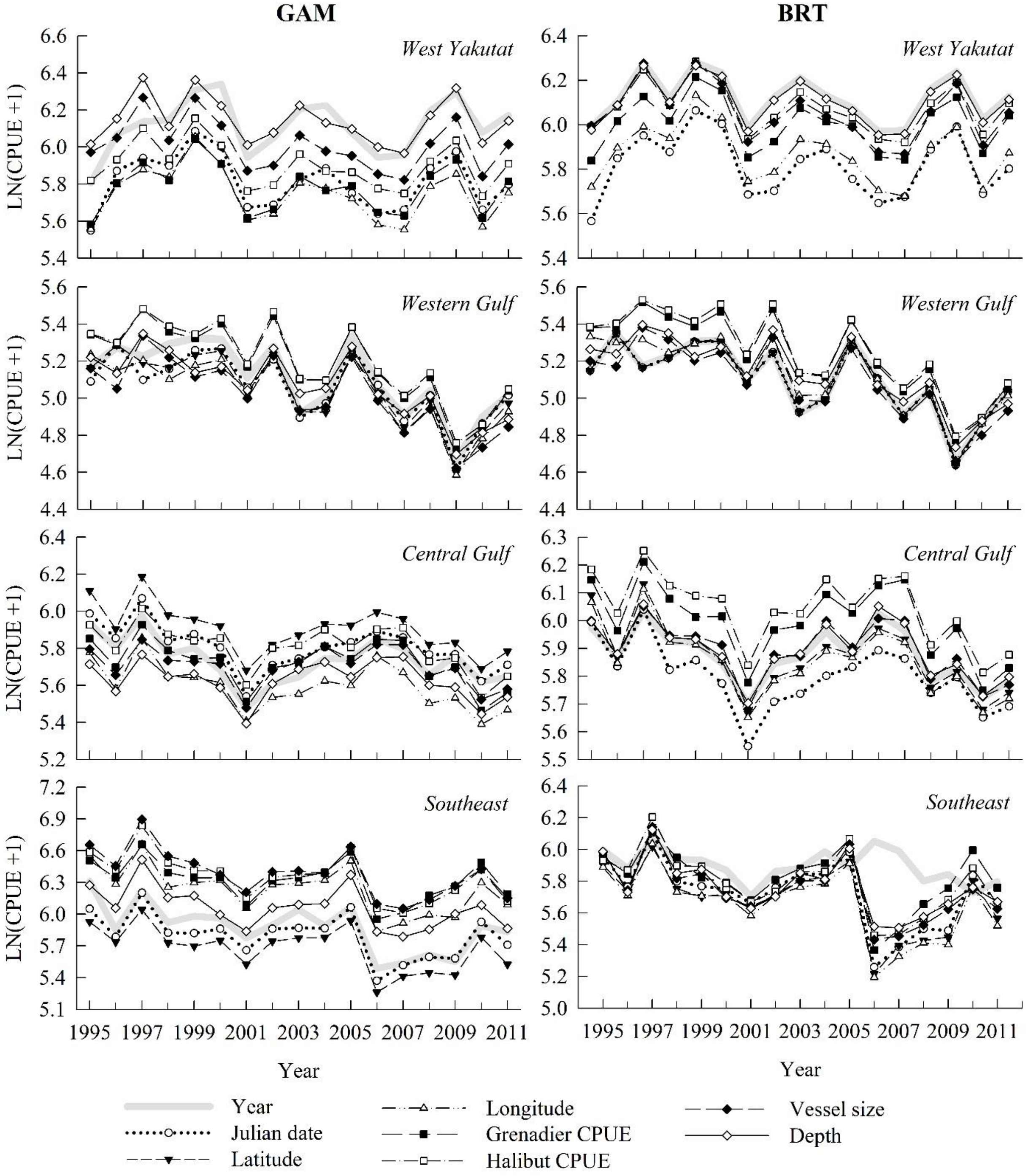
Appendix Figure 9. -- Summary of percent relative contribution of predictors on the BRT model for the Bering Sea and Aleutian Islands subregions.



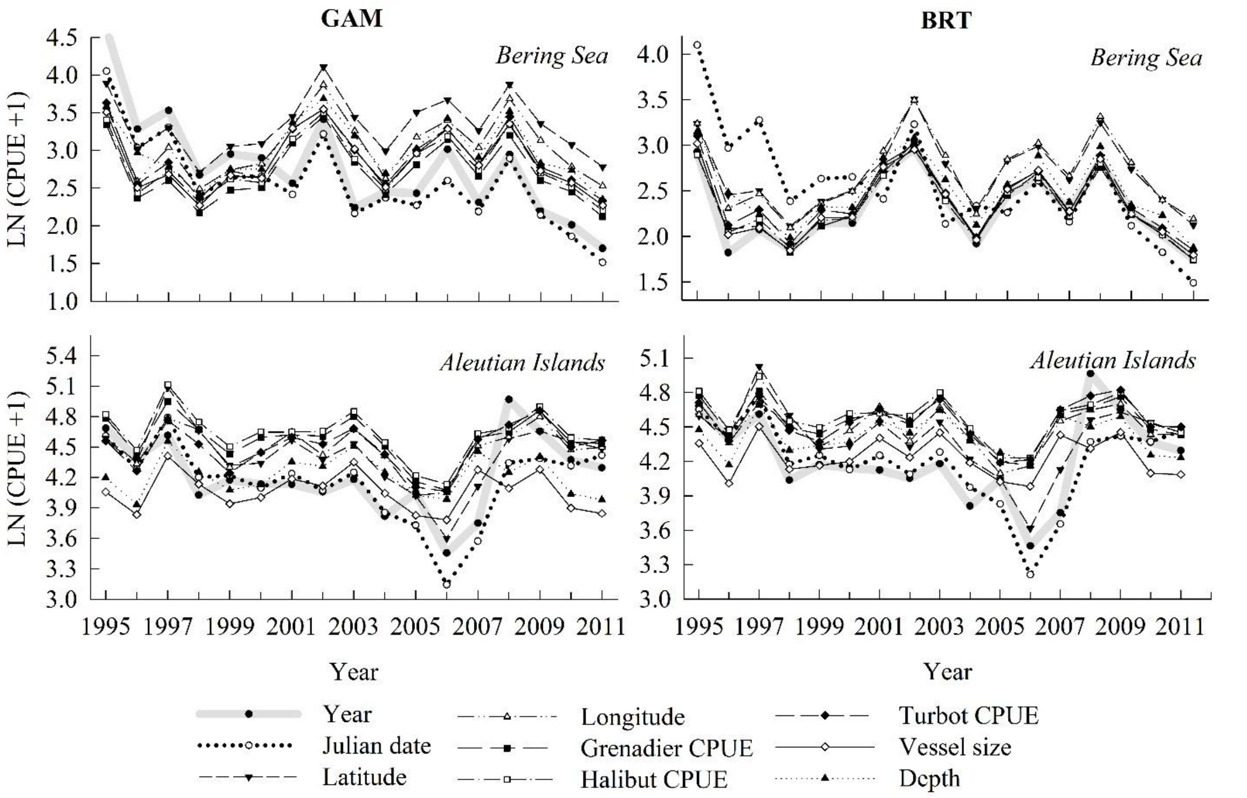
Appendix Figure 10. -- Comparisons of residual differences between nominal CPUE and estimated values from the selected models in GOA and BSAI subregions.



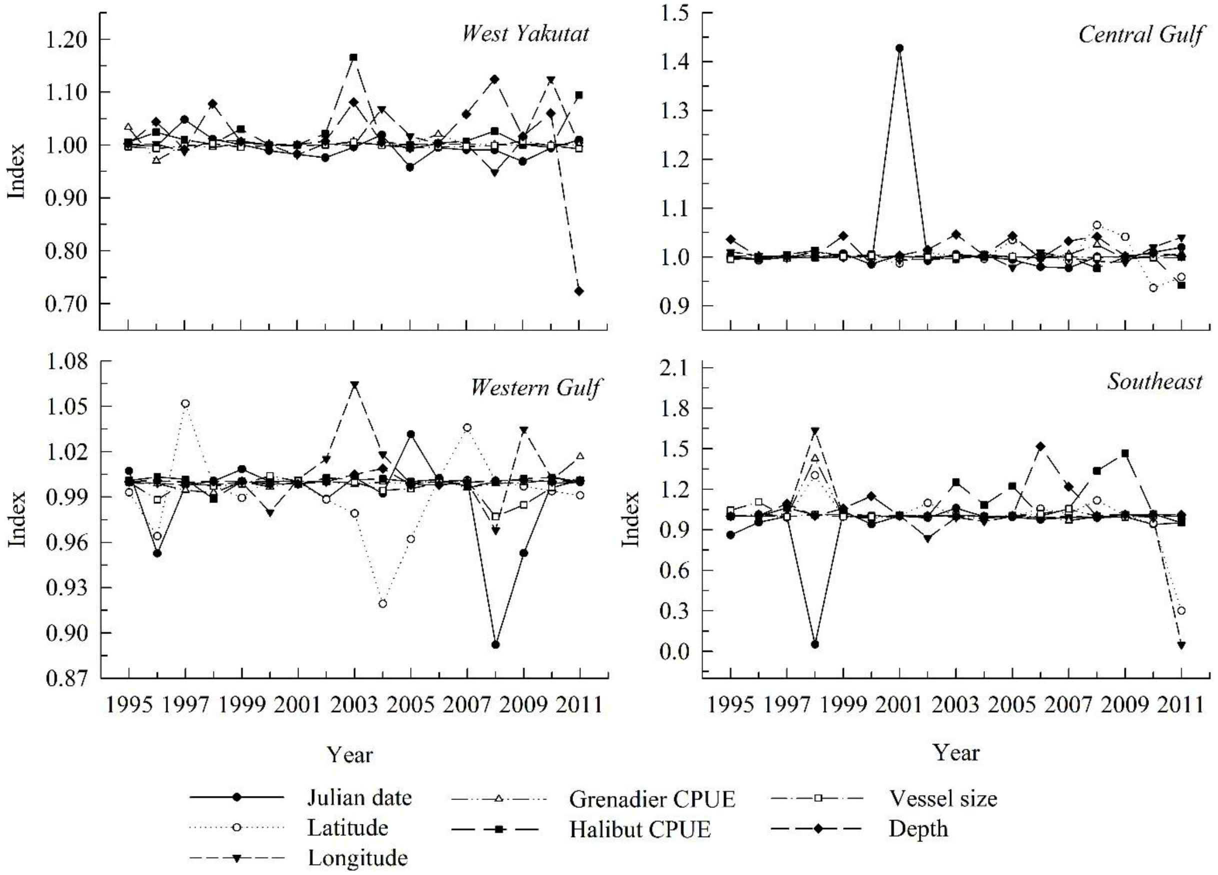
Appendix Figure 11. -- Step plots of CPUE changes by adding each predictor on the GAM and BRT models for the GOA subregions.



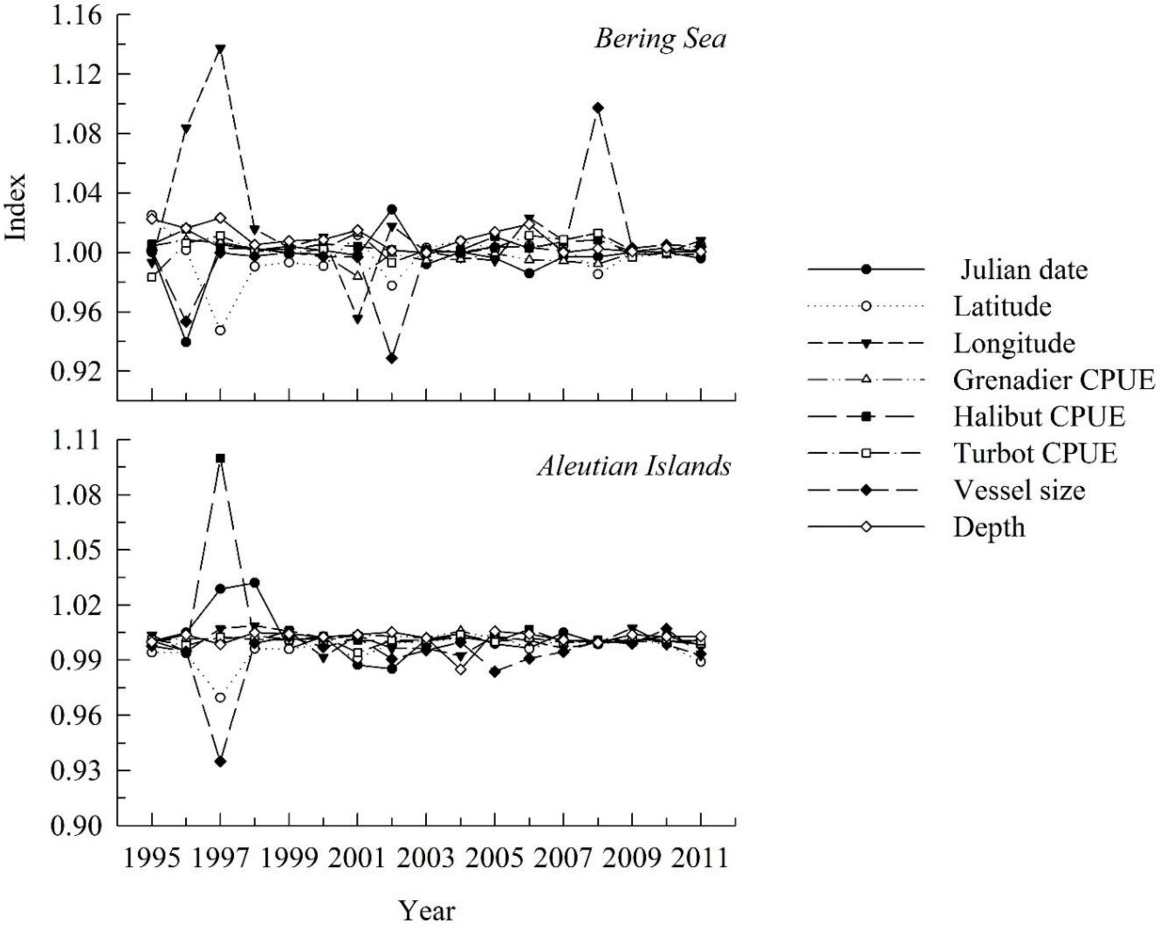
Appendix Figure 12. -- Step plots of CPUE changes by adding each predictor using GAM and BRT models for the BSAI subregions.



Appendix Figure 13. -- Influence plots on CPUE patterns by each predictor using GAM models for the GOA subregions.



Appendix Figure 14. -- Influence plots on CPUE patterns by each predictor using GAM models for the BSAI subregions.



Code to Standardize CPUE for the Sablefish Longline Fishery Using Generalized Linear Methods, General Additive Methods, and Boosted Regression Trees

I Code for accessing original database from AKFIN and to incorporate vessel length from a different source database.

Comments: This code is utilized to incorporate zeros on the fishing trips that did not catch sablefish or targeted species. Thus, you have a complete list of all the fishing trips for a particular species for that year.

```
path<-getwd()

obsall<-read.csv(paste(path,"/norpac_haul_hook_count_report.csv",sep=""),header=TRUE,skip=8,nrows=3500000) #nrows is a subset, remove to get whole thing

vess_length<-read.csv(paste(path,"/Vessel Length.csv",sep=""))

zerosable<-obsall[obsall$Species!=203,]

zerosable<-zerosable[!duplicated(zerosable$Haul.Join),]

zerosable$Species<-203

zerosable$Species.Name<-"SABLEFISH (BLACKCOD)"

zerosable$Extrapolated.Weight..kg.<-0

zerocod<-obsall[obsall$Species!=202,]

zerocod<-zerocod[!duplicated(zerocod$Haul.Join),]

zerocod$Species<-202

zerocod$Species.Name<-"PACIFIC COD"

zerocod$Extrapolated.Weight..kg.<-0

zerohalibut<-obsall[obsall$Species!=101,]

zerohalibut<-zerohalibut[!duplicated(zerohalibut$Haul.Join),]

zerohalibut$Species<-101

zerohalibut$Species.Name<-"PACIFIC HALIBUT"

zerohalibut$Extrapolated.Weight..kg.<-0

zerogrenadier<-obsall[obsall$Species!=80,]

zerogrenadier<-zerogrenadier[!duplicated(zerogrenadier$Haul.Join),]

zerogrenadier$Species<-80

zerogrenadier$Species.Name<-"GRENADIER UNIDENTIFIED"

zerogrenadier$Extrapolated.Weight..kg.<-0

zeroturbot<-obsall[obsall$Species!=102,]

zeroturbot<-zeroturbot[!duplicated(zeroturbot$Haul.Join),]

zeroturbot$Species<-102

zeroturbot$Species.Name<-"TURBOT"

zeroturbot$Extrapolated.Weight..kg.<-0

zerogiantgrenadier<-obsall[obsall$Species!=82,]
```

```
zerogiantgrenadier<-zerogiantgrenadier[!duplicated(zerogiantgrenadier$Haul.Join),]
zerogiantgrenadier$Species<-82
zerogiantgrenadier$Species.Name<-"GIANT GRENADIER"
zerogiantgrenadier$Extrapolated.Weight..kg.<-0
zeropacificgrenadier<-obsall[obsall$Species!=81,]
zeropacificgrenadier<-zeropacificgrenadier[!duplicated(zeropacificgrenadier$Haul.Join),]
zeropacificgrenadier$Species<-81
zeropacificgrenadier$Species.Name<-" PACIFIC GRENADIER"
zeropacificgrenadier$Extrapolated.Weight..kg.<-0
obsallwithzeros<-rbind(obsall,zerosable,zerocod,zerohalibut,zerogrenadier,zeroturbot,zerogiantgrenadier,zeropacificgrenadier)
names(obsallwithzeros)
obsallwithzeros <-obsallwithzeros[obsallwithzeros$Species %in% c(203,202,101, 80,102,82,81), ]
obs_vess<-as.matrix(obsallwithzeros$Vessel)
vessel_length<-matrix(nrow=length(obs_vess),ncol=1)
ves<-sort(unique(vess_length$Vessel_ID))
v_length<-matrix(nrow=length(ves),ncol=1)
for(i in 1:length(ves)){
vl<-subset(vess_length,vess_length$Vessel_ID==ves[i])
v_length[i,1]<-as.numeric(sort(unique(vl$vessel_length)))}
for(i in 1:length(ves)){
r<-which(obs_vess==as.character(ves[i]))
vessel_length[r,1]<-v_length[i]}

obsallwithzeros<-cbind(obsallwithzeros,vessel_length)

obsallwithzeros<-subset(obsallwithzeros,obsallwithzeros$vessel_length!="NA")
write.csv(obsallwithzeros, ("c:/ivan/zeros10.csv"))
```

II Code to prepare database for analysis using distinct statistical methods

#Comments. In this part the code is written to clean the database of unwanted variables by using the subsets command. In this part new variables are calculated such as CPUE

Preparation to create database for analysis of statistical methods for CPUE Standardization

```
zeros10spec <- read.csv("c:/ivan/zeros10spec.csv")
```

Code to eliminate unused variables on the database

## Comments: This code was used to delete variables that for different reasons were not considered for subsequent analyses

```
zeros10spec$X<-NULL
zeros10spec$count.segments<-NULL
zeros10spec$ Cruise<-NULL
zeros10spec$ Fishing.Start.Date<-NULL
zeros10spec$ Fishing.Depth<-NULL
zeros10spec$ Gear<-NULL
zeros10spec $Gear.Description<-NULL
zeros10spec$ Performance<-NULL
zeros10spec$ Lat.DD.Start<-NULL
zeros10spec$ Lon.DD.Start<-NULL
zeros10spec$ Avg.Number.of.Hooks<-NULL
zeros10spec$ Received.from.NORPAC<-NULL
zeros10spec$ Loaded.to.Repository<-NULL
zeros10spec$Species.Name<-NULL
zeros10spec$ Avg.Hook.Spacing..cm.<-NULL
zeros10spec$X.1<-NULL
zeros10spec$X.2<-NULL
zeros10spec$Count.Segments<-NULL
```

Code to eliminate missing values (NA) from a vector.

Comments: a decision was made to eliminate records with missing variables.

```
zeros10spec<-subset(zeros10spec,zeros10spec$CPUE.80!="NA")
zeros10spec<-subset(zeros10spec,zeros10spec$CPUE.203!="NA")
zeros10spec<-subset(zeros10spec,zeros10spec$ Skates.in.Set !="NA")
zeros10spec<-subset(zeros10spec,zeros10spec$ Hooks.per.Skate !="NA")
zeros10spec<-subset(zeros10spec,zeros10spec$Bottom.Depth!="NA")
```

```
zeros10spec<-subset(zeros10spec,zeros10spec$ Extrapolated.Weight..kg.
!="NA")
```

Code to calculate new variables

#Comments:In these lines CPUE is transformed in natural logarithm and depth (fathoms) is converted to meters. This code is also used to add months as a potential variable.

```
zeros10spec$logCPUE.203<-log(zeros10spec$CPUE.203+1)
zeros10spec$logCPUE.202<-log(zeros10spec$CPUE.202+1)
zeros10spec$logCPUE.101<-log(zeros10spec$CPUE.101+1)
zeros10spec$logCPUE.102<-log(zeros10spec$CPUE.102+1)
zeros10spec$logCPUE.80<-log(zeros10spec$CPUE.80+1)
zeros10spec$Depthmeters <-zeros10spec$ Bottom.Depth*1.8288
a<-as.Date(as.character(zeros10spec$Haul.Date),"%d-%b-%y")
b<-months(a)
zeros10spec$Months<-as.character(as.factor(b))
```

Code to eliminate subset depth above 100

```
zeros10spec<-subset(zeros10spec$Bottom.Depth >100)
```

Code to change to vector classes to factors

```
(zeros10spec$Year<-as.factor(zeros10spec$Year))
(zeros10spec$Months<-as.factor(zeros10spec$Months))
```

#Code to eliminate vessels with less than 15 years or dont fish in all years till 2011

#Comments this code was used to reduce the numbers vessels that had few trips over the years and may potentially bias the CPUE of sablefish.

```
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A062")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A088")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A159")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A165")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A176")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A181")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A182")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A187")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A192")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A196")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A197")
```

```
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A199")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A201")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A226")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A231")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A237")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A239")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A243")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A258")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A260")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A312")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A398")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A399")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A401")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A412")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A515")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A521")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A522")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A528")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A544")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A617")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A628")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A629")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A632")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A650")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A658")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A659")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A660")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A675")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A691")
zeros10spec<-subset( zeros10spec,zeros10spec$Vessel!="A692")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A697")
zeros10spec<-subset( zeros10spec,zeros10spec$Vessel!="A698")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A699")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A700")
zeros10spec<-subset(zeros10spec,zeros10spec$Vessel!="A707")
```

Code to subset by subarea

#Comments: In the analysis it was decided to run analysis by subregions and not by regions

```
AI<-subset(GOAN, GOAN$FMP.Subarea=="CG")
```

III Code to do model fitting and CPUE standardization for GLM GAM and BRT

#Comments: Here an example is used for the subregion Central Gulf. The first lines are to get rid of years 1991-1994 and to use Julian dates instead of Months. To Standarize CPUE a new database is created with all the continuous variables held against their means. Before the analysis was started the function dredge in the package MuMln was used to infer what variables to keep in the models.

```
CG <- read.csv("c:/ivan/CG.csv")
```

Code to use AKAIKE Information criteria to evauate the best model and keep the best variables influencing the AIC index.

```
xx<-dredge(CG.gam,extra=alist(AIC,BIC,ICOMP,Cp))

write.csv(xx,"CGdredge.csv")


##To eliminate Years 1991-1994

CG<-subset(CG, CG$Year >="1995")

#### Turn dates into time of year #####

names(CG)[3]<-"Date"

CG$Date<-strptime("30.12.1899 00:00:00","%d.%m.%Y %H:%M:%S")+(CG$Date*60*60*24)

CG$Date<-as.numeric(format(strptime(CG$Date,"%Y-%m-%d %H:%M:%S"),"%j"))

CG$Year<-as.factor(CG$Year)

CG$Month<-as.factor(CG$Month)

### Make marginal mean data frame

CGFULL <- CG

for(j in 1:length(CG[,1])) {

  CGFULL[j,3:9]<-apply(CG[,3:9],FUN=mean,MARGIN=2)

}
```

#GLM

```
CGglm<-glm(Sablefish.CPUE~Year+ Date + Latitude+ Longitude+ Depth.m+ Vessel.length +Grenadier.CPUE+Halibut.CPUE,family=gaussian,data=CG)

summary(CGglm)
```

#GAM

```
library(mgcv)

CGgam<- gam(Sablefish.CPUE~Year+ s(Date) + s(Latitude)+ s(Longitude)+ s(Grenadier.CPUE)+s(Halibut.CPUE) +s(Vessel.length)+ s(Depth.m),family=gaussian,data=CG)

summary(CGgam)
```

```
#BRT
source("brt.functions.R")
library(gbm)
CGgbm  <- gbm.step(data=CG,
   gbm.x = 2:9,
   gbm.y = 1,
   family = "gaussian",
   tree.complexity = 1,
   learning.rate = 0.01,
   bag.fraction = 0.5)
## To predict standardized CPUE with marginal means
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
gbm.test<-
predict(CGgbm, CGFULL ,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se,gbm.fit)
```

IV Code to create model diagnostics from GAM and BRT models modified from (Abeare 2009)

```
###GLM partial residual plots
library(car)
par(mfrow=c(2,2),mar=c(4,3,4,4))
crPlots(CGglm,main=NULL,line=TRUE,                                   cex.lab=1.5,
cex.axis=1.5,smooth=FALSE,col="black",ylab="")
# GLM diagnostic plots
par(mfrow=c(2,2))
plot(CGglm )

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

```
# GAM diagnostic plots
gam.check(GAMCG)
# Cooks distance for gam
plot(cooks.distance(GAMCG),ylab="",main="GAM-SABLEFISH CG: Cook's Distance")
identify(cooks.distance(GAMCG),tolerance=0.1)
######################### Boosted regression tree models##########################


# bar plot of predictor influence
par(mar=c(4,11,4,3))
summary(GBMCG,cBars=length(GBMCG$var.names),n.trees=GBMCG$n.trees,plotit=TRUE,order=TRUE,normalize=TRUE,cex.axis=1,las=2,main=NULL)


# Fitted function plots
gbm.plot(CGgbm,smooth=TRUE,rug=TRUE,n.plots=14,write.title=F,rug.side=1,rug.lwd=1,rug.tick=0.05, cex.lab=1.5, cex.axis=1.5,show.contrib = F,plot.layout=c(2,2))


# Residual Analyses
residgam<-resid(GAMCG)
 residgbm<-GBMCG$residuals
gam.resid<-residgam
gbm.resid<-residgbm
resid.plot<-data.frame(CG$Year,gam.resid,gbm.resid)
library(doBy)
gam.year<-summaryBy(formula=gam.resid~CG.Year,data=resid.plot,FUN=c(mean,sd))
gbm.year<-summaryBy(formula=gbm.resid~CG.Year,data=resid.plot,FUN=c(mean,sd))
model.resid<-data.frame(gam.year,gbm.year)
write.csv(model.resid, ("c:/model.resid.csv")
#Box Plots residuals
boxplot(gam.resid,gbm.resid,data=resid.plot,names=c("GAM","BRT"),notch=TRUE)
```

Code to create estimate semivariograms and bubble plots to infer autocorrelation among reponse and explanatory variables

```
# Semivariograms Plot
library(geoR)
#breaks
dists<-dist(CG[,[4:5])
```

```
summary(dists)
breaks=seq(0,6,l=13)
# Omnidirectional semivariogram
CG.vario1<-variog(coords=CG[,4:5],data=CG[,1],option=c("bin"),breaks=breaks)
#Model residual variograms
CG.vario.resid2<-
variog(coords=CG[,4:5],data=residuals(GAMCG),option=c("bin"),breaks=breaks)
CG.vario.resid3<-
variog(coords=CG[,4:5],data=residuals(GBMCG),option=c("bin"),breaks=breaks)
#Residual variogram plots
plot(CG.vario1,type="b",main=NULL,pts.range=c(1),scaled=FALSE,ylim=c(0.05,5),var.lines=TRUE,xlab="distance (degrees)")
lines(CG.vario.resid2$u,CG.vario.resid2$v,type="l",lwd=2,col="purple")
lines(CG.vario.resid3$u,CG.vario.resid3$v,type="l",lwd=2,col="dark green")
legend("bottomright",inset=0.025,legend=c("GAM","BRT"),
col=c("purple","dark green"),lty=1,lwd=2)
```

V Code to include error estimates in GAM Models

```
CG <- read.csv("c:/ivan/CG.csv")
##To eliminate Years 1991-1994
CG<-subset(CG, CG$Year >="1995")

#### Turn dates into time of year #####
names(CG)[3]<-"Date"
CG$Date<-strptime("30.12.1899 00:00:00","%d.%m.%Y %H:%M:%S")+(CG$Date*60*60*24)
CG$Date<-as.numeric(format(strptime(CG$Date,"%Y-%m-%d %H:%M:%S"),"%j"))

CG$Year<-as.factor(CG$Year)
### Make marginal mean data frame

CGFULL <- CG
for(j in 1:length(CG[,1])) {
 CGFULL[j,3:9]<-apply(CG[,3:9],FUN=mean,MARGIN=2)
}
```

```
library(mgcv)
library(gbm)
library(doBy)
library(sampling) # must be done after doBy because of the conflict with the survival package called by doBy

#initialize the data frame
#nits is number of iterations
styr<-1995
endyr<-2011
nits<-100
bootframe<-data.frame(matrix(1,nits,endyr-styr+1)) # Don't use c as a variable because of its use as concatentate
for(i in 1:nits) {
#Random stratified sampling with replacement
nbyyear<-as.numeric(table(CG$Year))
nbyyear
test<-strata(CG,c("Year"),size=nbyyear,method="srswr")
test.set<-getdata(CG,test)
row.names<-as.vector(test.set$ID_unit)
train.set<-CG[-row.names,]
table(test.set$Year)
table(train.set$Year)

test.set$Year<-as.factor(test.set$Year)
#GAM it up
gam.test<-gam(Sablefish.CPUE~Year+ s(Date) + s(Latitude)+ s(Longitude)+ s(Grenadier.CPUE)+s(Halibut.CPUE)+ s(Vessel.length) + s(Depth.m),family=gaussian,data=test.set)
gam.pred<-predict(gam.test,newdata=CGFULL, type="response",se.fit=TRUE) # This is a more normal bootstrap of the data
ys<-as.vector(CGFULL$Year)
gamfit<-as.vector(gam.pred$fit)
MYearCPUE<-data.frame(ys,gamfit)
GAMSE<-MYearCPUE
GAMby<-summaryBy(formula=gamfit ~ys,data=GAMSE ,FUN=mean)
bootframe[i,]<-GAMby[,2];
}
## 95% percentile Confidence intervals
cis<-apply(bootframe, 2, quantile, proCG = c(0.025,0.975))
means<-apply(bootframe,2,mean)
sd(bootframe)
### Make some CI plots
plot(1:17,exp(cis[1,]),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))),pch="")
lines(1:17,exp(means),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))))
lines(1:17,exp(cis[1,]),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))),lty=2)
lines(1:17,exp(cis[2,]),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))),lty=2)
```

V Code to include error estimates in BRT Models

```
CG <- read.csv("c:/ivan/CG.csv")
##To eliminate Years 1991-1994
CG<-subset(CG, CG$Year >="1995")
#### Turn dates into time of year #####
names(CG)[3]<-"Date"
CG$Date<-strptime("30.12.1899 00:00:00","%d.%m.%Y %H:%M:%S")+(CG$Date*60*60*24)
CG$Date<-as.numeric(format(strptime(CG$Date,"%Y-%m-%d %H:%M:%S"),"%j"))
CG$Year<-as.factor(CG$Year)
### Make marginal mean data frame
CGFULL <- CG
for(j in 1:length(CG[,1])) {
  CGFULL[j,3:9]<-apply(CG[,3:9],FUN=mean,MARGIN=2)
}
library(mgcv)
library(gbm)
library(doBy)
library(sampling) # must be done after doBy because of the conflict with the survival package called by doBy
#initialize the data frame
#nits is number of iterations
styr<-1995
endyr<-2011
```

```
nits<-100
bootframe<-data.frame(matrix(1,nits,endyr-styr+1)) # Don't use c as a variable because of its use as
concatentate
for(i in 1:nits) {
 #Random stratified sampling with replacement
nbyyear<-as.numeric(table(CG$Year))
nbyyear
test<-strata(CG,c("Year"),size=nbyyear,method="srswr")
test.set<-getdata(CG,test)
row.names<-as.vector(test.set$ID_unit)
train.set<-CG[-row.names,]
table(test.set$Year)
table(train.set$Year)
test.set$Year<-as.factor(test.set$Year)
 test.set <-test.set[,c(1,33,2:36)]
test.set$row.names<-NULL

#BRT it up
        CGgbm  <- gbm.step(data=test.set,
           gbm.x = 2:9,
           gbm.y = 1,
           family = "gaussian",
           tree.complexity = 1,
           learning.rate = 0.01,
           bag.fraction = 0.5)
gbm.test<- predict(CGgbm, CGFULL
,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
gbm.fit<-as.vector(gbm.test)

ys<-as.vector(CGFULL$Year)
gbmfit<-as.vector(gbm.test)
MYearCPUE<-data.frame(ys,gbmfit)
GAMSE<-MYearCPUE
GBMby<-summaryBy(formula=gbmfit ~ys,data=GAMSE ,FUN=mean)
bootframe[i,]<-GBMby[,2];
}
## 95% percentile Confidence intervals
cis<-apply(bootframe, 2, quantile, proCG = c(0.025,0.975))
means<-apply(bootframe,2,mean)
sd(bootframe)
### Make some CI plots
plot(1:17,exp(cis[1,]),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))),pch="")
lines(1:17,exp(means),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))))
lines(1:17,exp(cis[1,]),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))),lty=2)
lines(1:17,exp(cis[2,]),ylim=c(0.9*min(exp(cis)),1.1*max(exp(cis))),lty=2)
sd(bootframe)
```

VII Code to create 5-fold crossvalidation procedure to infer model performance of GAM and BRT

Code for K-Folds 5-fold

```
CG <- read.csv("c:/ivan/CG.csv")
##To eliminate Years 1991-1994
CG<-subset(CG, CG$Year !="1991")
CG<-subset(CG, CG$ Year !="1992")
CG<-subset(CG, CG$Year !="1993")
CG<-subset(CG, CG$Year !="1994")

(CG$Year<-as.factor(CG$Year))
(CG$Months<-as.factor(CG$Month))
table(CG$Year)

#Random stratified sampling without replacement
library(sampling)
test<-strata(CG[order(CG$Year),],c("Year"),
size=c(65,48,65,55,61,63,61,69,76,65,63,65,50,42,54,49,41), "srswor", TRUE)
test.set<-getdata(CG,test)
row.names<-as.vector(test.set$ID_unit)
train.set<-CG[-row.names,]
test1<-strata(train.set[order(train.set$Year),],c("Year"),
size=c(65,48,65,55,61,63,61,69,76,65,63,65,50,42,54,49,41), "srswor", TRUE)
fold1<-getdata(train.set,test1)
```

```
row.names<-as.vector(fold1$ID_unit)
train.set2<-train.set[-row.names,]
test2<-strata(train.set2[order(train.set2$Year),],c("Year"),
size=c(65,48,65,55,61,63,61,69,76,65,63,65,50,42,54,49,41), "srswor", TRUE)
fold2<-getdata(train.set2,test2)
row.names<-as.vector(fold2$ID_unit)
train.set3<-train.set2[-row.names,]
test3<-
strata(train.set3[order(train.set3$Year),],c("Year"),size=c(65,48,65,55,61,63,61,69,76,65,63,65,50,42,54,4
9,41), "srswor", TRUE)
fold3<-getdata(train.set3,test3)
row.names<-as.vector(fold3$ID_unit)
fold4<-train.set3[-row.names,]
write.csv(fold4, ("c:/ivan/KFold4CG.csv"))
write.csv(fold3, ("c:/ivan/KFold3CG.csv"))
write.csv(fold2, ("c:/ivan/KFold2CG.csv"))
write.csv(fold1, ("c:/ivan/KFold1CG.csv"))
write.csv(test.set, ("c:/ivan/FoldtestCG.csv"))
```

VIII Code to combined all folds in to one file for subsequent analysis

```
CG <- read.csv("c:/ivan/CG.csv")
model.performKfold1CG <- read.csv("c:/ivan/model.performKfold1CG.csv")
model.performKfold2CG <- read.csv("c:/ivan/model.performKfold2CG.csv")
model.performKfold3CG <- read.csv("c:/ivan/model.performKfold3CG.csv")
model.performKfold4CG <- read.csv("c:/ivan/model.performKfold4CG.csv")
model.performKfold5CG <- read.csv("c:/ivan/model.performKfold5CG.csv")
model.performKfold6CG <- read.csv("c:/ivan/model.performKfold6CG.csv")
model.performKfold7CG <- read.csv("c:/ivan/model.performKfold7CG.csv")
model.performKfold8CG <- read.csv("c:/ivan/model.performKfold8CG.csv")
model.performKfold9CG <- read.csv("c:/ivan/model.performKfold9CG.csv")
model.performKfold10CG <- read.csv("c:/ivan/model.performKfold10CG.csv")
model.performKfold11CG <- read.csv("c:/ivan/model.performKfold11CG.csv")
model.performKfold12CG <- read.csv("c:/ivan/model.performKfold12CG.csv")
model.performKfold13CG <- read.csv("c:/ivan/model.performKfold13CG.csv")
model.performKfold14CG <- read.csv("c:/ivan/model.performKfold14CG.csv")
model.performKfold15CG <- read.csv("c:/ivan/model.performKfold15CG.csv")
model.performKfold16CG <- read.csv("c:/ivan/model.performKfold16CG.csv")
model.performKfold17CG <- read.csv("c:/ivan/model.performKfold17CG.csv")
model.performKfold18CG <- read.csv("c:/ivan/model.performKfold18CG.csv")
model.performKfold19CG <- read.csv("c:/ivan/model.performKfold19CG.csv")
model.performKfold20CG <- read.csv("c:/ivan/model.performKfold20CG.csv")


CGALLFolds<-rbind(model.performKfold1CG,model.performKfold2CG,model.performKfold3CG,
model.performKfold4CG, model.performKfold5CG, model.performKfold6CG, model.performKfold7CG,
model.performKfold8CG,model.performKfold9CG,model.performKfold10CG,
model.performKfold11CG,model.performKfold12CG,model.performKfold13CG,
model.performKfold14CG,model.performKfold15CG,model.performKfold16CG,
model.performKfold17CG,model.performKfold18CG,
model.performKfold19CG,model.performKfold20CG)
write.csv(CGALLFolds, ("c:/ivan/ CGALLFolds.csv"))
CGALLFolds <- read.csv("c:/ivan/CGALLFolds.csv")
```

X Code to calculate step plots for GAM models

Comments: These plots are used to discern discrepeancies among unstandarized and standardized CPUES. The reason for this is that there could be other variables influencing more the CPUE rather than the abundance itself**.**

```
CG <- read.csv("c:/ivan/CG.csv")
 ##To eliminate Years 1991-1994
 CG<-subset(CG, CG$Year >="1995")
  #### Turn dates into time of year #####
 names(CG)[3]<-"Date"
 CG$Date<-strptime("30.12.1899 00:00:00","%d.%m.%Y %H:%M:%S")+(CG$Date*60*60*24)
 CG$Date<-as.numeric(format(strptime(CG$Date,"%Y-%m-%d %H:%M:%S"),"%j"))
  CG$Year<-as.factor(CG$Year)
 CG$Month<-as.factor(CG$Month)
### Make marginal mean data frame
  CGFULL <- CG
 for(j in 1:length(CG[,1])) {
   CGFULL[j,3:9]<-apply(CG[,3:9],FUN=mean,MARGIN=2)
 }
```

#GAM Year

```
library(mgcv)
library(doBy)
```

```
CGgam<- gam(Sablefish.CPUE~Year,family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMSE
GAMYEAR<-GAMSE
GAMYEAR
#GAM Date
library(mgcv)
CGgam<- gam(Sablefish.CPUE~Year+ s(Date), family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMDATE<-GAMSE
GAMDATE
#GAM latitude
library(mgcv)
CGgam<- gam(Sablefish.CPUE~Year+ s(Date) + s(Latitude),family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMSE
GAMLAT<-GAMSE
```

```
GAMLAT
```

#GAM Longitude

```
library(mgcv)
CGgam<- gam(Sablefish.CPUE~Year+ s(Date) + s(Latitude)+ s(Longitude),family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)

year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMLONG<-GAMSE
GAMLONG
#GAM Grenadier
CGgam<- gam(Sablefish.CPUE~Year+ s(Date) + s(Latitude)+ s(Longitude)
+s(Grenadier.CPUE),family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMGREN<-GAMSE
GAMGREN
```

#GAM Halibut

```
CGgam<-gam(Sablefish.CPUE~Year+s(Date)+s(Latitude)+
s(Longitude)+s(Grenadier.CPUE)+s(Halibut.CPUE),family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
```

```
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMSE
GAMHAL<-GAMSE
GAMHAL
```

#GAM Vessel

```
CGgam<-gam(Sablefish.CPUE~Year+s(Date)+s(Latitude)+s(Longitude)+
s(Grenadier.CPUE)+s(Halibut.CPUE)+ s(Vessel.length),family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMVESSEL<-GAMSE
GAMVESSEL
```

#GAM Depth

```
CGgam<-gam(Sablefish.CPUE~Year+s(Date)+s(Latitude)+s(Longitude)+
s(Grenadier.CPUE)+s(Halibut.CPUE)+ s(Vessel.length)+s(Depth.m),family=gaussian,data=CG)
gam.test<-predict(CGgam,newdata= CGFULL, type="response",se.fit=TRUE)
gam.fit<-as.vector(gam.test$fit)
gam.se<-as.vector(gam.test$se.fit)
year<-as.vector(CGFULL$Year)
CPUE<-as.vector(CGFULL$Sablefish.CPUE)
MYearCPUE<-data.frame(year,CPUE,gam.fit,gam.se)
GAMSE<-summaryBy(formula=gam.fit ~year,data= MYearCPUE,FUN=c(mean))
GAMDEPTH<-GAMSE

GAMDEPTH
```

XI Code to calculate step plots Boosted regression trees

```
Library(gbm)
CG <- read.csv("c:/ivan/CG.csv")
##To eliminate Years 1991-1994
CG<-subset(CG, CG$Year >="1995")
```

```
#### Turn dates into time of year #####
names(CG)[3]<-"Date"
CG$Date<-strptime("30.12.1899 00:00:00","%d.%m.%Y %H:%M:%S")+(CG$Date*60*60*24)
CG$Date<-as.numeric(format(strptime(CG$Date,"%Y-%m-%d %H:%M:%S"),"%j"))

CG$Year<-as.factor(CG$Year)
CG$Month<-as.factor(CG$Month)
### Make marginal mean data frame

CGFULL <- CG
for(j in 1:length(CG[,1])) {
  CGFULL[j,3:9]<-apply(CG[,3:9],FUN=mean,MARGIN=2)
}

#GAM Year
CGgbm  <- gbm(Sablefish.CPUE~Year ,data=CG, distribution = "gaussian", interaction.depth =
1,shrinkage= 0.01, bag.fraction = 0.5, cv.folds=10, n.trees=5000)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Year<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Year
year.gbm<-plot(CGgbm,i.var=1,return.grid=TRUE)
year<-year.gbm
year
#GAM Date
source("brt.functions.R")
library(gbm)
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:3,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Date<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Date
year.gbm<-plot(CGgbm,i.var=1,return.grid=TRUE)
date<-year.gbm
Date
#GAM latitude
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:4,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Latitude<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Latitude
year.gbm<-plot(CGgbm,i.var=1,return.grid=TRUE)
latitude<-year.gbm
```

```
latitude
#GAM Longitude
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:5,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Longitude<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Longitude

year.gbm<-plot(CGgbm,i.var=1,return.grid=TRUE)
longitude<-year.gbm
longitude
#GAM Grenadier
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:6,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Grenadier<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Grenadier

year.gbm<-plot(CGgbm,i.var=1,return.grid=TRUE)
grenadier<-year.gbm
grenadier

#GAM Halibut
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:7,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Halibut<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Halibut

year.gbm<-plot(CGgbm,i.var=1,return.grid=TRUE)
halibut<-year.gbm
halibut
#GAM Vessel
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:8,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
 Vessel<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
 Vessel
```

#GAM Depth

```
CGgbm  <- gbm.step(data=CG,
    gbm.x = 2:9,
    gbm.y = 1,
    family = "gaussian",
    tree.complexity = 1,
    learning.rate = 0.01,
    bag.fraction = 0.5)
gbm.test<-
 predict(CGgbm, CGFULL,n.trees=CGgbm$gbm.call$best.trees,type="response",se.fit=TRUE)
 gbm.fit<-as.vector(gbm.test)
year<-as.vector(CGFULL$Year)
 CPUE<-as.vector(CGFULL$Sablefish.CPUE)
 MYearCPUE<-data.frame(year,CPUE,gbm.fit)
 library(doBy)
Depth<-summaryBy(formula=gbm.fit ~year,data= MYearCPUE,FUN=c(mean))
Depth
```

XII Code to calculate variable relative influence using GAM method only from (Bentely 2012)

```
CG <- read.csv("c:/ivan/CG.csv")
 ##To eliminate Years 1991-1994
 CG<-subset(CG, CG$Year >="1995")
 #### Turn dates into time of year #####
 names(CG)[3]<-"Date"
 CG$Date<-strptime("30.12.1899 00:00:00","%d.%m.%Y %H:%M:%S")+(CG$Date*60*60*24)
 CG$Date<-as.numeric(format(strptime(CG$Date,"%Y-%m-%d %H:%M:%S"),"%j"))

 CG$Year<-as.factor(CG$Year)
 CG$Month<-as.factor(CG$Month)
 ### Make marginal mean data frame

 CGFULL <- CG
 for(j in 1:length(CG[,1])) {
   CGFULL[j,3:9]<-apply(CG[,3:9],FUN=mean,MARGIN=2)
 }
 #initialize the data frame

 styr<-1995
 endyr<-2011
 coef.list<-list(endyr-styr+2) # set up list of params
 test.set<-CG
 test.set$Year<-as.factor(test.set$Year)
 samsizes<-rep(1,endyr-styr+2)
 samsizes[1]<-dim(test.set[1])
library(mgcv)
 gam.test<-gam(Sablefish.CPUE~s(Date)+s(Latitude)+s(Longitude)+
s(Grenadier.CPUE)+s(Halibut.CPUE)+s(Vessel.length)+s(Depth.m)+ -1,family=gaussian,data=test.set)
 coef.list[[1]]<-coef(gam.test)
 for(i in styr:endyr) {
 test.set<-CG[CG$Year==i,]
 test.set$Year<-as.factor(test.set$Year)

 samsizes[i-1993]<-dim(test.set[1])
 mean.set<-CGFULL[CGFULL$Year==i,]
 mean.set$Year<-as.factor(mean.set$Year)
 #Gam it up
```

```
  gam.test<-gam(Sablefish.CPUE~s(Date)+s(Latitude)+s(Longitude)
s(Grenadier.CPUE)+s(Halibut.CPUE)+s(Vessel.length, k=4)+s(Depth.m),family=gaussian,data=test.set)

 #gam.pred<-predict(gam.test,newdata= test.set, type="response",se.fit=TRUE) # This is a more normal
bootstrap of the data

 #gam.pred<-predict(gam.test,newdata= CGFULL, type="response",se.fit=TRUE) # This is more like a
crossvalidation where you estimate with one set, and predict with the fullset, except backwards

 coef.list[[i-1993]]<-coef(gam.test)

 print(i)

 #colnames(bootframe)<-names(coef(gam.test))

 }

coefsums<-data.frame(matrix(endyr-styr+2,1,7))

 colnames(coefsums)<-c("Date","Lat","Lon", "Gren","Hbut", "Vess", "Depth")

 for(i in 1:(endyr-styr+2)) {

 coefsums[i,1]<-sum(coef.list[[i]][grep("ate",names(coef.list[[i]]))])

coefsums[i,2]<-sum(coef.list[[i]][grep("atit",names(coef.list[[i]]))])

 coefsums[i,3]<-sum(coef.list[[i]][grep("ongi",names(coef.list[[i]]))])

 coefsums[i,4]<-sum(coef.list[[i]][grep("renad",names(coef.list[[i]]))])

 coefsums[i,5]<-sum(coef.list[[i]][grep("alib",names(coef.list[[i]]))])

 coefsums[i,6]<-sum(coef.list[[i]][grep("ess",names(coef.list[[i]]))])

 coefsums[i,7]<-sum(coef.list[[i]][grep("epth",names(coef.list[[i]]))])

 }

 coefsums[1,]<-coefsums[1,]/samsizes[1]

 influ<-coefsums

 for (i in 2:(endyr-styr+2)) {

   influ[i,]<-exp((coefsums[i,]-coefsums[1,])/samsizes[i]) }

 rownames(influ)<-c("FULL",seq(styr,endyr))

  plot(influ[2:18,])

 influ
```

RECENT TECHNICAL MEMORANDUMS

Copies of this and other NOAA Technical Memorandums are available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22167 (web site: *www.ntis.gov)*. Paper and electronic (.pdf) copies vary in price.

AFSC-

268 FOWLER, C. W., R. D. REDEKOPP, V. VISSAR, and J. OPPENHEIMER. 2014. Pattern-based control rules for fisheries management, 116 p. NTIS number pending.

267 FOWLER, C. W., and S. M. LUIS. 2014. We are not asking management questions, 48 p. NTIS number pending.

266 LAUTH, R. R., and J. CONNER. 2014. Results of the 2011 Eastern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna, 176 p. NTIS number pending.

265 TRIBUZIO, C. A., J. R. GASPER, and S. K. GAICHAS. 2014. Estimation of bycatch in the unobserved Pacific halibut fishery off Alaska, 506 p. NTIS No. PB2014-101866.

264 STONE, R. P., K. W. CONWAY, D. J. CSEPP, and J. V. BARRIE. 2014. The boundary reefs: glass sponge (Porifera: Hexactinellida) reefs on the international border between Canada and the United States, 31 p. NTIS number pending.

263 SHELDEN K. E. W., D. J. RUGH, K. T. GOETZ, C. L. SIMS, L. VATE BRATTSTRÖM, J. A. MOCKLIN, B. A. MAHONEY, B. K. SMITH, and R. C. HOBBS. 2013. Aerial surveys of beluga whales, *Delphinapterus leucas*, in Cook Inlet, Alaska, June 2005 to 2012, 122 p. NTIS number pending.

262 WHITEHOUSE, G. A. 2013. A preliminary mass-balance food web model of the eastern Chukchi Sea, 164 p. NTIS number pending.

261 FERGUSON, M. C., and J. T. CLARKE. 2013. Estimates of detection probability for BWASP bowhead whale, gray whale, and beluga sightings collected from Twin Otter and Aero Commander aircraft,1989 to 2007 and 2008 to 2011, 52 p. NTIS number pending.

260 BREIWICK, J. M. 2013. North Pacific marine mammal bycatch estimation methodology and results, 2007-2011, 40 p. NTIS number pending.

259 HIMES-CORNELL, A., K. HOELTING, C. MAGUIRE, L. MUNGER-LITTLE, J. LEE, J. FISK, R. FELTHOVEN, C. GELLER, and P. LITTLE. 2013. Community profiles for North Pacific fisheries - Alaska. (Volumes 1-12). NTIS number pending.

258 HOFF, G. R. 2013. Results of the 2012 eastern Bering Sea upper continental slope survey of groundfish and invertebrate resources, 268 p. NTIS number pending.

257 TESTA, J. W. (editor). 2013. Fur seal investigations, 2012, 90 p. NTIS number pending.

256 LAUTH, R. R., and D. G. NICHOL. 2013. Results of the 2012 eastern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate resources, 162 p. NTIS No. PB2014100850.

255 BOVENG, P. L., J. L. BENGTSON, M. F. CAMERON, S. P. DAHLE, E. A. LOGERWELL, J. M. LONDON, J. E. OVERLAND, J. T. STERLING, D. E. STEVENSON, B. L. TAYLOR, and H. L. ZIEL.2013. Status review of the ribbon seal (*Histriophoca fasciata*), 174 p. NTIS No. PB2009104582.

254 ECHAVE, K. B., D. H. HANSELMAN, and N. E. MALONEY. 2013. Report to industry on the Alaska sablefish tag program, 1972 - 2012, 47 p. NTIS No. PB2013111080.

253 ECHAVE, K., C. RODGVELLER, and S.K. SHOTWELL. 2013. Calculation of the geographic area sizes used To create population indices for the Alaska Fisheries Science Center longline survey, 93 p. NTIS number pending.