Gulf of Alaska Analysis

Garrett Allen: (Render Time ~2 minutes)

Introduction

Environmental modeling is becoming increasingly important as climate change increases in severity, as businesses, scientists, and the policymakers need to be able to quickly react to changes in species' population. For my individual project, we intend to try to tackle this topic using methods in the spatiotemporal modeling literature, as well as try to find relationships between the covariates, species abundance, and key metrics of animal welfare in the Alaskan Gulf region.

The data we will be working with in this case study is the Alaskan Fishery Science Center's Groundfish and Crab Assessment Program (GAP), specifically surveys of populations on the bottom of the Alaskan Gulf. The dataset that we will be working with during this project is from the National Oceanic and Atmospheric Administration (NOAA) website, which contains survey data over the years 1993-2021. The data was collected via a process called bottom trawling, where researchers drag a large net across the bottom of the ocean floor to capture ground fish, crabs, and other crustaceans in order to assess their population. In this report, we will be specifically addressing a particular type of cod named the walleye pollock, as it is one of the most sustainably managed and economically valuable species to Alaska (Fisheries 2023).

Data Collection Process

The data is from the Gulf of Alaska Bottom Trawl Survey, conducted since 1984, and conducted biannually since 1999. A bottom trawl is a kind of fishing where fishing boats drag huge nets across the seafloor in order to catch creatures near the bottom of the ocean, such as crustaceans, groundfish, and sponge. To conduct the survey, fishing ships went out into the Gulf of Alaska and conducted multiple bottom trawls, with each bottom trawl being referred to as a particular haul of fish. After running the bottom trawl for long enough, the nets would be brought up to classify the types of species captured in the net. The survey ran for multiple months during the fishing season in order to sample all of the Gulf of Alaska; in particular,

researchers broke up the Gulf of Alaska into many different stratum each year, and made sure to get sufficiently many samples from each stratum during the fishing season. The survey paid particular attention to important fish in Alaska, such as the walleye pollock, one of the most commercially important fish in terms of global food supply and the Alaskan economy.

Each row of the dataset represents a species caught during one of these hauls, and includes information about where the haul was conducted, the water temperature, depth, net size, species caught, and the weight/count of the species caught, among other variables. Table 1 shows all of the variables of interest that will be used in this report. To collect the data, we pulled 218,838 observations of species' hauls from NOAA's GOA API. In particular, walleye pollock populations vary significantly by depth and season, so we will pay special attention to these variables in our analysis (Charles F. Adams 2009).

Table 1: Variables of Interest

Name	Units	Variable
Depth	Meters	depth
Bottom Temp	Celsius	bot
Surface Temp	Celsius	surf
Net Width	Meters	net-width
Net Height	Meters	net-height

CPUE: Catch per unit effort

The primary goal of NOAA's fishing surveys in Alaska is to assess if fishery resources are being used sustainably across the Alaskan waterfront. Specifically, NOAA is interested in how populations of economically important fish species are changing over time, so that fisheries can preemptively take actions to avoid population crashes, which can be economically and biologically devastating to the region.

In order to assess the population of a particular species of fish, fisheries commonly use an indirect measure of population effort called Catch per Unit Effort (CPUE), defined informally as how many fish you would expect to catch per 1 unit of effort expended on the part of a fishery. In the case of hook fishing, the CPUE is often defined as the number of fish caught per 1000 hooks used. CPUE is proportional to the abundance in the population, since a high CPUE indicates it is easy to catch a fish, while a low CPUE indicates the opposite. More details on the relationship between CPUE and abundance are given in (Mark N. Maunder 2006).

Since the Groundtrawl survey uses nets across the ocean floor, the CPUE is typically defined as the number/weight of fish caught per unit area fished. In this analysis, our definition of CPUE wil be the kilograms of fish caught per hectacre fished, which will be denoted `cpue_kgha`.

Aims

Aim 1: How can we best predict CPUE for the walleye pollock across both spatial and temporal dimensions in the Gulf of Alaska, and what will happen to the population in the near future?

Since CPUE is supposed to be proportional to abundance, it is of value for an organization like NOAA to be able to predict CPUE in order to predict fish abundance. Surveys like the GOA Bottom Trawl Survey are expensive and time consuming to conduct; additionally, they only occur biannually. It would be useful for NOAA to be able to predict CPUE into the future for different locations, so that between surveys, fisheries will have a good estimate of what their population looks like. This is the major goal of this report, and it will be assessed by comparing a zero-inflated generalized linear model (GLM) to a specific fishery model VAST (Vector Autoregressive Spatiotemporal model), and then predicting abundance from the model that performs the best.

Aim 2: How do different covariates present in the dataset predict CPUE?

It is also of use to understand how different variables affect CPUE, as without controlling for other effects on the CPUE, CPUE is not directly proportional to population abundance. For example, suppose we were interested in the population of a particular kind of surface fish; if all of our samples consisted of depths in the deep sea, we would expect to catch no fish on the surface. This would result in a near 0 CPUE, which would be not at all proportional to the abundance of our target species. (Mark N. Maunder 2006) gives many more situations in which the CPUE may give an inaccurate estimate of species abundance, such as using the wrong kind of hook/net, seasonality, and shifting water temperatures, to name a few.

Thus, it is of value to try to understand how different covariates impact the CPUE in our model in order to try to get a more accurate estimation of species abundance over time and space. This will be assessed by comparing our best model to a model with no covariates to see how much more of our variance is explained by adding our covariates of interest.

EDA and Data Processing

Missing Data.

There are some missing data concerns when modeling the walleye pollock. For one, hauls where no walleye pollock were caught were not initially present in the data. That is, they were implicit zeroes, as any haul that did not catch walleye pollock did not contain a row for this species.

To address this, if a haul was present in the data but contained no row for walleye pollock, we added a row for that haul representing walleye pollock with a count of zero. Luckily, all of our covariates of interest are the same within the same hauls, so we was able to copy over covariate data from other fish in the haul (i.e. net width, height, depth, etc.). Thus, all of the implicit zeroes where at least one other fish was caught in a haul were able to be made successfully into explicit zeroes for walleye pollock.

There were some hauls that were completely missing from the dataset; it is likely that these hauls had entirely empty nets, so no information was recorded about these hauls. Unfortunately, these data are irretrievably missing, as no data is known about any of the covariates of these hauls. This survivorship bias could potentially lead to us overestimating the CPUE for certain regions, as there may be zeroes present in a region that we cannot account for due to missingness. For the walleye pollock, there exist 1074 missing observations of this type that we have no way to effectively analyze, and as such these will not be included in our analysis.

There also existed some missing values of the covariates; specifically, depth and surface/bottom temperature all had naturally missing values in our dataset. These covariates were imputed by taking the mean of all the points in their same sampling stratum in the same month of the year; this was done so that mean imputation captures both points close in space (within the same stratum) and within the same season ofthe year (month).

EDA Plots

The EDA plots presented here help to understand some of the underlying relationships in our data before tackling our main modeling question of interest.

As we can see from Figure 1, there do exist some relationships between the `cpue_kgha` and other covariates in our data. This suggests that we should be considering a GLM component to our spatial analysis, as cpue does not remain constant in different parts of the Gulf of Alaska, where things like temperature vary.

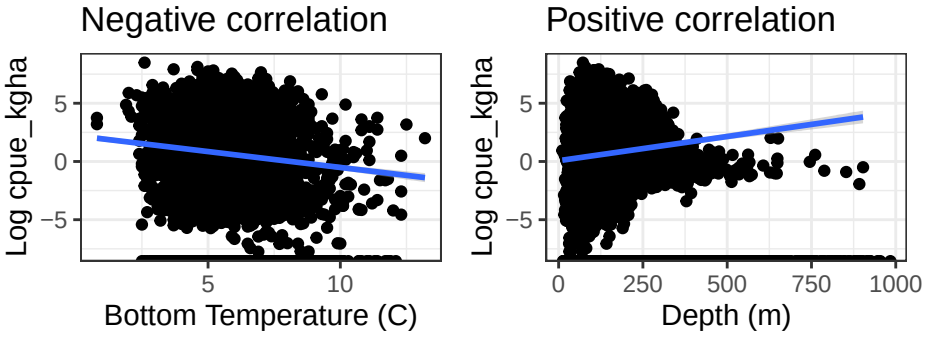

Figure 1: Relationship between cpue_kgha and covariates

As we can see from Figure 2, our data was collected approximately bianually, and most of the counts of our fish are relatively small, with a few very large outliers. This may make it difficult to pick up on some of the larger hauls in our data.

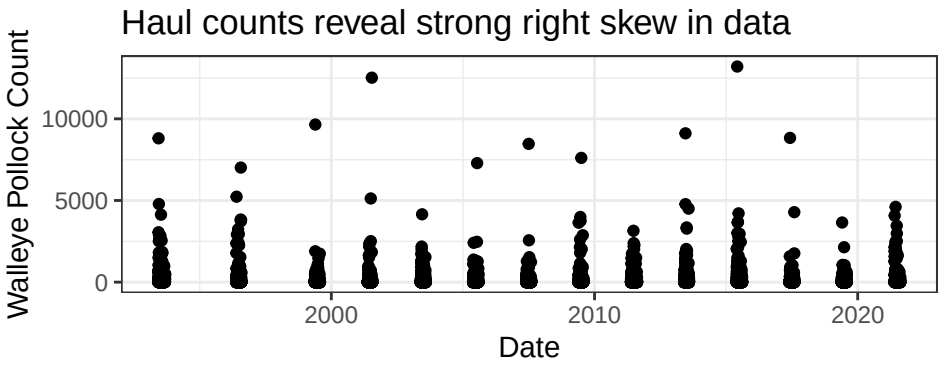

Figure 2: Amount of walleye pollock in each haul, 1993-2021

Finally, the spatial distribution of our data in Figure 3 reveals that we will likely need to aggregate our data to the stratum level, rather than work with the raw latitude/longitude of our data. This is because almost all of our cpue_kgha values are very small, since on most hauls, almost no walleye pollock are caught. This may prove a challenge when modeling, as it may be difficult for our models to pick up on the spatial/temporal correlation, given the amount of low cpue_kgha values in our data.

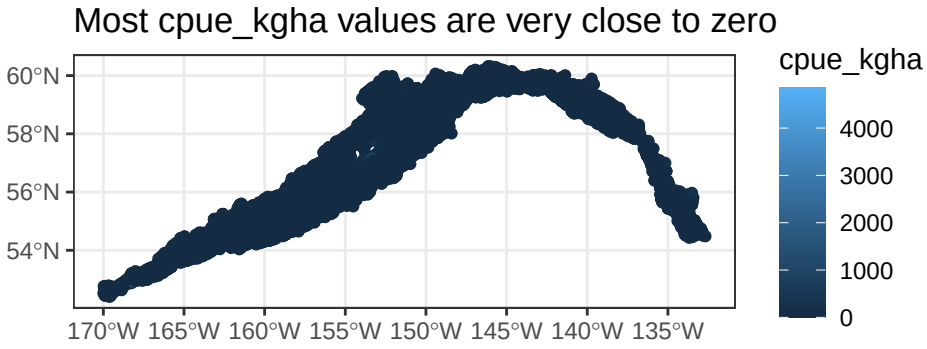

Figure 3: Spatial distribution of cpue_kgha in Gulf of Alaska

Methods

Zero inflated GLM.

To try to see if adding a spatiotemporal component is important when predicting CPUE, we fit a Poisson zero inflated GLM on the count of fish in a given haul. A zero inflated regression model is one where we first predict (using a classifier, such as logistic regression) whether a given value is zero, and then predict the non-zero points with a GLM. In this case, we fit logistic model to predict if a given haul was zero, and then we fit a Poisson model on the count in that haul if it is deemed non-zero. In mathematical notation, the model is as follows:

$$\mathrm{P}(\mathrm{Count}_i = 0) = \pi_i + (1 - \pi_i)e^{-\lambda_i}$$
$$\mathrm{P}(\mathrm{Count}_i = y_i) = (1 - \pi_i)\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

where:

$$\mathrm{logit}(\pi_i) = \beta_0 + \beta_1 \mathrm{bot}_i + \beta_2 \mathrm{surf}_i + \beta_3 \mathrm{depth}_i + \beta_4 \mathrm{netheight}_i + \beta_5 \mathrm{netwidth}_i + \beta_6 \mathrm{vessel}_i + \beta_7 \mathrm{year}_i$$

$$\log(\lambda_i) = \beta_0 + \beta_1 \mathrm{bot}_i + \beta_2 \mathrm{surf}_i + \beta_3 \mathrm{depth}_i + \beta_4 \text{net-height}_i + \beta_5 \text{net-width}_i + \beta_6 \mathrm{vessel}_i + \beta_7 \mathrm{year}_i$$

This model made sense to fit, as nearly 20% of our walleye pollock data is zeroes. Fitting a poisson model also made sense since we are modeling count data. To fit this model, we used the `zeroinfl` package in R. Sensitivity analysis was conducted by performing backward selection from our full model without interaction effects, which were not included due to limited support for their conclusion and lack of support for their inclusion in EDA.

Zero-Inflated Poisson Diagnostics and Results

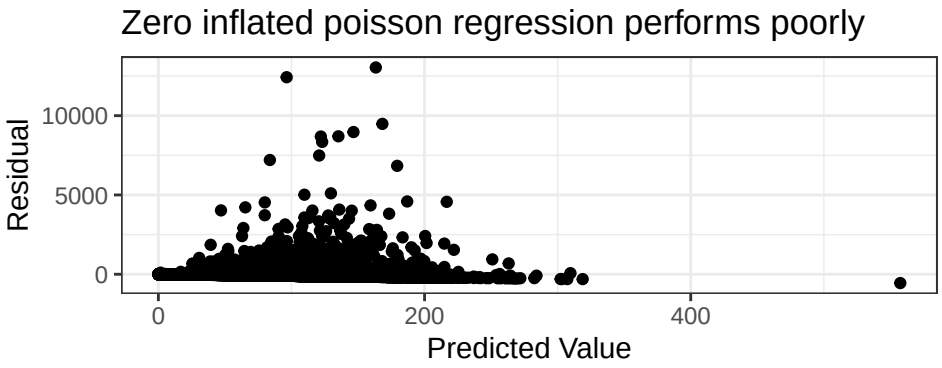

Figure 4: Residuals for Zero Inflated Model

As we can see by Figure 4, this model does exceptionally poorly, with the errors being in the hundreds of fish. The RMSE is 433, which is fairly bad, considering that most of the counts of fish in our dataset are quite small. The AIC is also quite high; this should be expected, given that our model does not account for spatiotemporal correlation between observations. As a result, we will not spend much time analyzing this model's output, since VAST ends up performing significantly better; however, this is a helpful baseline to understand how much better VAST performs when compared to a more traditional model for this kind of data.

VAST Model Description

VAST (Vector Autoregressive Spatiotemporal model) is a model created by James Thorson at NOAA that efficiently models fishery data; in particular, it can be used to measure species abundance directly, unlike our previous model, which only had the ability to predict the counts of fish/CPUE.

The model is rather complicated, and a full description of the model is given in (Thorson 2018), but a high level version of the model will be presented here. VAST is a spatiotemporal model where we fit two linear predictors p_1 and p_2 for each of the ith samples in our survey. The full form for the first linear predictor is given below, as seen in (Thorson 2019) in slightly more generality:

where $\beta_1(t_i)$ represents the intercept term for each time step, the second and third sum represent the spatial and spatiotemporal effect on the model (repsectively) and the last two sums represent density covariates (those affecting the underlying distributWion of where fish are) and catchability covariates (those affecting solely the ability to catch fish, but not their underlying distribution). For example, a catchability covariate in our model would be the width of the net, since it does not affect the underlying distribution of fish, but only how often we

catch them, whereas surface temperature would affect where walleye pollock is distributed in the Gulf of Alaska, so it would be a density covariate.

$p_1(i)$ and $p_2(i)$ are then used to create a probability that a given point is zero and to predict the actual counts of the non-zero points, respectively. Specifically, we calculate two quantities:

$$p_1(i) = \beta_1(t_i) + \sum_{f=1}^{n_{w1}} L_{w1}(f) w_1(s_i, f) + \sum_{f=1}^{n_{e1}} L_{e1}(f) e_1(s_i, f, t_i) + \sum_{f=1}^{n_{n1}} L_1(f) n_1(f) + \sum_{p=1}^{n_p} \gamma_1(t_i, p) X(s_i, t_i, p) + \sum_{k=1}^{n_k} \lambda_1(k) Q(i, k)$$

$$r_1(i) = \text{logit}^{-1}(p_1(i))$$
$$r_2(i) = a_i \exp(p_2(i))$$

These are then used in the prediction of zero count hauls ($r_1(i)$) and the rate of fish in the area $r_2(i)$ respectively. Finally, the model uses these various calculated quantities to calculate the density at each spatial location and time, as well as an abundance index for each spatial stratum defined in our model.

In order to fit the model, we specify a certain number of knots (in our case, due to computational constraints, 100 will be used) which are points selected in the Gulf of Alaska that represent fixed cells that can be used for our model prediction. This is because the locations need to remain constant over time for our model to predict, so we summarize our covariates and data to these 100 knots in order to account for spatiotemporal correlation.

There are many more details to the model that are too long to get into in this report, but for readers interested in further discussion of the model, (Thorson 2018) delves into more detail.

VAST Modeling Choices

When fitting VAST, there exist many different modeling choices that must be made by the user. These are detailed in (Thorson 2019): We won't go entirely into the choices made here, but we will detail the most important ones.

Firstly, we decided to include a vessel effect in the model, which controls for the effect of vessel on the amount of fish caught. This controls for unknown variables that could have affected how many fish were caught (i.e. skill of crew, noise of engine, etc) that we would expect to vary between different vessels.

Additionally, we decided upon 5 total covariates to use in the model; three density covariates and two catchability covariates; these covariates are below. This is both because there are

conceptually good reasons for fitting a model with these covariates, and becauseour EDA revealed slight associations between these variables and CPUE. To assess if these variables are useful in prediction, we also fit a VAST model with no covariates to assess how much predictive performance changes.

We also fit a first order autoregressive process for the intercepts for each linear predictor, as suggested by (Thorson 2019) when survey data is not conducted annually, as in the case of our ground trawl survey here. The model was fit with 100 knots, largely due to computational constraints; with more compute power, we could scale up the model to have more knots, and we would expect that predictive performance should improve with a finer spatial density.

Table 2: Variables for VAST

Name	Type of Covariate
Depth	Density
Bottom Temp	Density
Surface Temp	Density
Net Width	Catchability
Net Height	Catchability
Vessel	Vessel Effect

VAST Model Limitations

Due to the model being so new, many methods that we would like to use to assess model output are incapable of being used as we expected. For example, prediction onto new data not seen in the model is still currently on the dev branch, and it threw errors that were not solvable within the time constraints of this case study. This means that all of our performance metrics will be evaluated on the training data for the model. As a result, all future evaluation metrics will be in-sample, both for the zero inflated approach and VAST. This is not the best way to evaluate our model, but given the complexity of the model and novelty, this is the best approach to be fair to both approaches.

VAST Diagnostics

After both of the VAST models were fit (covariate and no covariate) model convergence was checked by the built in helper function `check_fit`. Both models appropriately converged, and there were no issues with the coefficient values of the covariates.

Residuals and Convergence

The residuals in Figure 6 indicate that our model is doing rather well; our QQ plot looks roughly normally distributed, and our rank transformed standardized residuals are mostly below 1. This appears to be a significant improvement over our other model, although it appears to have a problem with overpredicting the CPUE, given that the residuals tend to be above zero. It's unclear why the model is having this difficulty, but it should make us cautious when interpreting results into the future with our model, as it seems likely that it will overpredict the fish abundance given that it overpredicts on the training data.

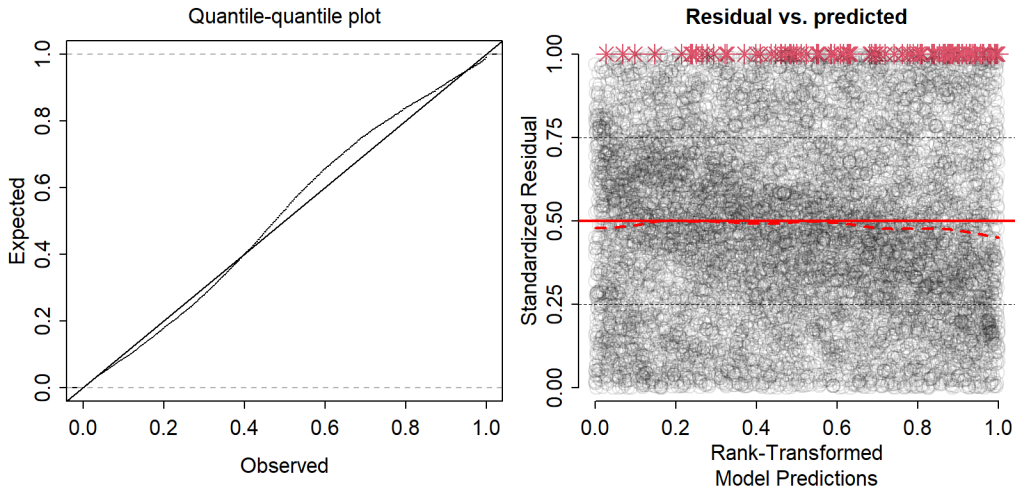

Figure 5: Standardized Residuals for VAST model

Results

Model Comparison

From our modeling, its pretty clear that the VAST model outperforms our zero inflated model. In Table 3, we can see that the VAST model with covariates performs the best on our training set, and that both VAST models perform significantly better than our zero inflated approach. This suggests that our specific spatiotemporal model performs the best when estimating abundance. However, the small difference in AIC between the covariate model and no covariate model suggests most of the work is being done by properly modeling spatiotemporal dependencies.

Table 3: AIC for each model

AIC_Zero_Infl	AIC_VAST_NoCov	AIC_VAST
3097753	81477.17	81375.96

Future projections

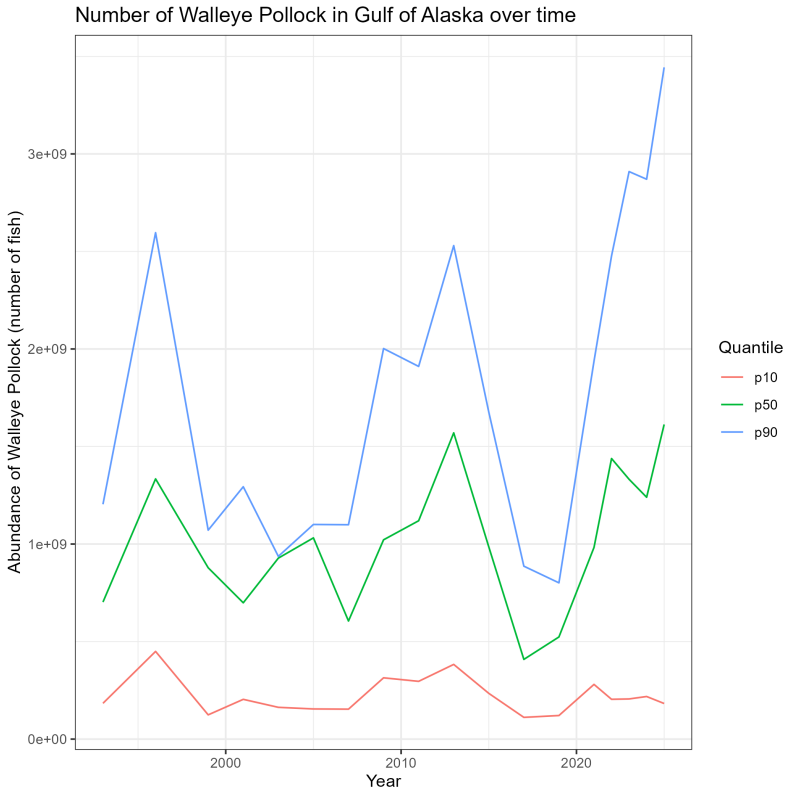

Figure 6: Projections of Walleye Pollock into 2025

As we can see in Figure 6, the walleye pollock population is predicted to remain the same or grow in the following years. The 10th percentile prediction for our fish abundance indicates that the population has remained mostly constant over time, and that walleye pollock will remain roughly the same through 2025. The 50th percentile of our prediction interval shows that the population will grow back to its peak in 2012, and continue to grow after that, while the 90th percentile indicates intense exponential growth in the walleye pollock population in the Gulf of Alaska. As mentioned in (Thorson 2019), we should likely trust the more conservative assessment of moderate growth/stagnation suggested by the 10th percentile. This is because we fit a first-order autoregressive process on the intercept for our covariates. This can cause inaccurate exponential growth in our model into the future, and given that our residuals

suggest that our model is overpredicting on our training data, a conservative estimate of mild growth seems most accurate.

Covariate Analysis

Understanding covariates in our VAST model is difficult, as the output does not support easy interpretation of these covariates. In addition, work is still in development on performing inference on these covariates. However, the extremely small difference in predictive performance between the intercept only VAST model and the covariate VAST model indicates that our covariates are not very useful in predicting fish populations. The deviance difference between the two models is quite small, and they perform roughly the same in all available metrics between these models. Thus, we tentatively conclude that our analysis did not show that the covariates we included dramatically impact abundance of walleye pollock in the Gulf of Alaska. However, this is an extremely tentative conclusion, and we will discuss more about its limitations in our limitations section.

Limitations and Future Work

The limitations of our approach are apparent. Given the novelty of the VAST model, many important kinds of analysis weren't available as a part of this analysis (namely, out of sample validation). Due to computational constraints, we were limited in the number of knots that could be fit within a reasonable period of time, which additionally limited sensitivity analysis on our final VAST model. In addition, VAST model outputs are difficult to work with and interpret, which makes our results mostly useful for prediction rather than inference. Finally, there may have been more models that we could have compared VAST to in order to give a fair comparison (such as a GLM with a Gaussian Process spatial effect) but due to time constraints and space constraints, these alternative approaches were not considered.

For future work in improving the models presented here, we would encourage the VAST model to be fit with additional knots, and for more sensitivity analysis to be done in tuning some of the more advanced model parameters. It would also be interesting to perform out of sample validation when functionality is added to predict at specific locations in the spatial domain, rather than only being able to predict at generated knots or in aggregate.

References

Charles F. Adams, John J. Kelley & Kenneth O. Coyle, Robert J. Foy. 2009. "Seasonal Changes in the Diel Vertical Migration of Walleye Pollock (Theragra Chalcogramma) in the Northern Gulf of Alaska." *Environmental Biology of Fishes* 86 (August): 297–305. https://doi.org/10.1007/s10641-009-9519-y.

Fisheries, NOAA. 2023. "Alaska Pollock," January. https://www.fisheries.noaa.gov/species/alaska-pollock.

Mark N. Maunder, Alain Fonteneau, John R. Sibert. 2006. "Interpreting Catch Per Unit Effort Data to Assess the Status of Individual Stocks and Communities." *ICES Journal of Marine Science* 63 (January): 1373–85. https://doi.org/10.1098/rspb.2002.2218.

Thorson, James T. 2018. "Forecast Skill for Predicting Distribution Shifts: A Retrospective Experiment for Marine Fishes in the Eastern Bering Sea." *Fish and Fisheries* 20 (November): 159–73. https://doi.org/10.1111/faf.12330.

———. 2019. "Guidance for Decisions Using the Vector Autoregressive Spatio-Temporal (VAST) Package in Stock, Ecosystem, Habitat and Climate Assessments." *Fisheries Research* 210: 143–61. https://doi.org/https://doi.org/10.1016/j.fishres.2018.10.013.