Natural and social phenomena are characterized by a high level of complexity and multiple factors that tend to be highly correlated. Overfishing in the Gulf of Maine, and its impact on the fisheries that once existed in the Gulf is one such phenomenon. This research project uses statistical analysis to estimate the likelihood that a particular region in the Gulf of Maine was a fishing spot in the past. In addition, the analysis enables us to estimate the many factors, such as, seasonality and the cost of fishing, which affected the productivity of historic Gulf of Maine fishing spots.

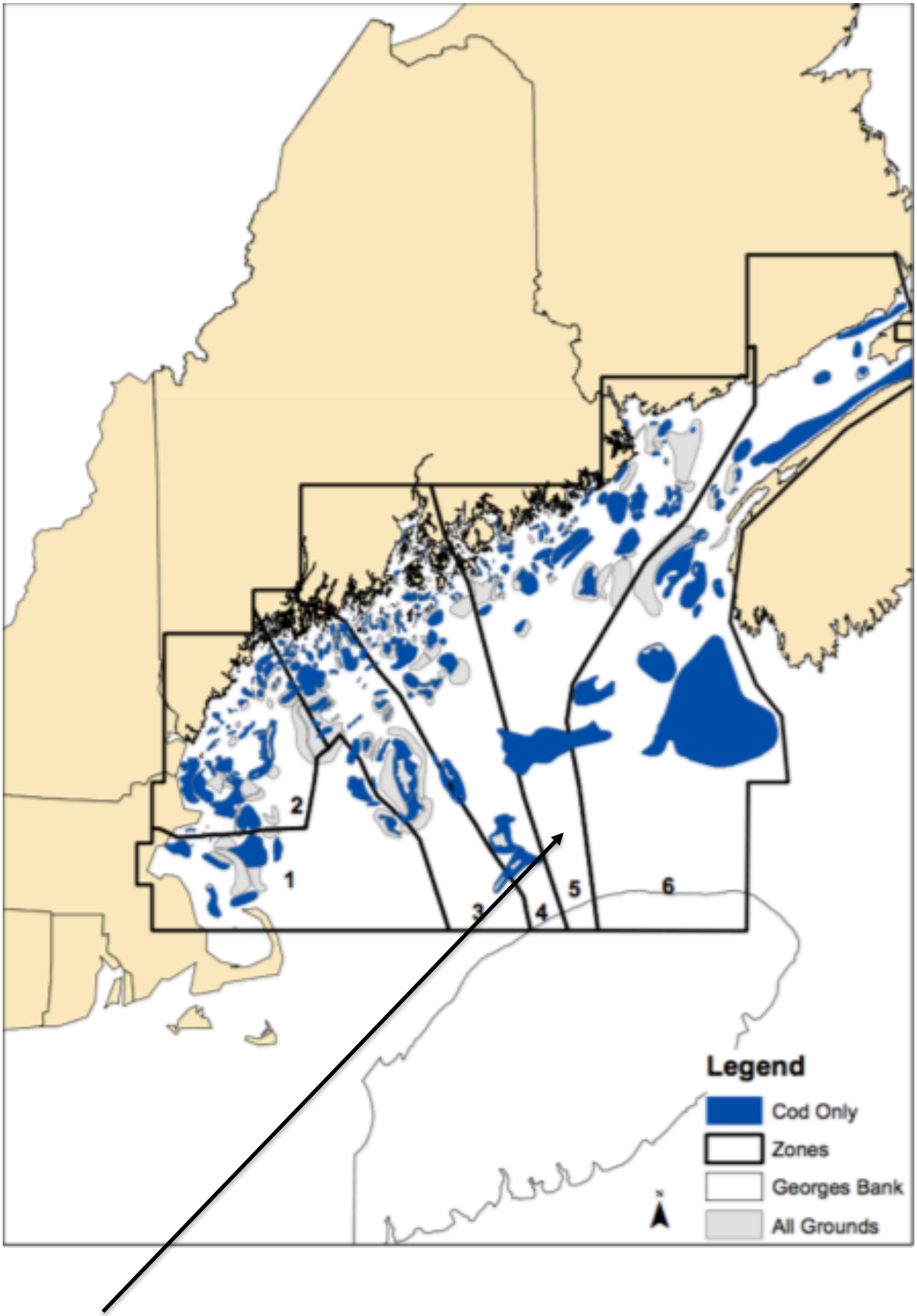
The project, which we have titled "Explaining the presence and productivity of historic fishing grounds in the Gulf of Maine," builds on the insight and information provided by the last few generations of fishermen who made a living catching fish in the Gulf. The data was collected by Ted Ames, a Bowdoin Coastal Scholar. Over the years Ted has interviewed retired fishermen who once caught hake, haddock, and cod. Furthermore, Professor Eileen Johnson and several Bowdoin students have made this data amendable for statistical analysis by digitizing the data that Ted collected. They created a map that indicates the boundaries of the historical fishing sites and their productivity across species and seasons.

The likelihood that a particular region in the gulf of Maine was a fishing spot in the past can be expressed mathematically using a relation that maps a set of explanatory factors to their likelihood. This set of explanatory factors or variable includes mean water depth (in meters), mean ocean floor slope, distance to the shore (in kilometres) and distance to a river (in kilometres). As part of my work on this project, I explored ways to make the fishing model as accurate as possible. For example, using the square of the explanatory variables, say distance to shore squared, leads to a result that implies that at first additional depth increases the likelihood of a spot being a good fishing spot, but, at some point more depth becomes disadvantageous. This is more realistic because some depth and distance is needed to find fish in a boat.

However, the high level of correlation within the set of explanatory variables makes it essential that the estimation is rigorous and that the model is robust enough to capture the complexity of the biophysical phenomena observed in the case of fishing spots. For instance, if the model omits one or more spatially correlated processes that affect fishing site productivity then the standard errors will be biased. Thus, one of my main aims is to explore statistical means to check and reduce the amount of bias and standard error that might be observed in the analysis.

This research was materialized and effectively executed through the generous support of Grua/O'Connell Fund. The research project was a great opportunity to gain insight and experience in the field of economics and applied mathematics. The project enabled me to build on coursework and academic perspective from my Mathematics and Economics majors in a new academic setting, beyond the classroom. Furthermore, the focus on fishing spots in the Gulf of Maine, helped me link my academic scholarship to cultural and natural facets of the state of Maine, thereby enhancing my understanding of the place I am grateful for the kind consideration and support provided by Professor Nelson and the Grua/O'Connell Fund.

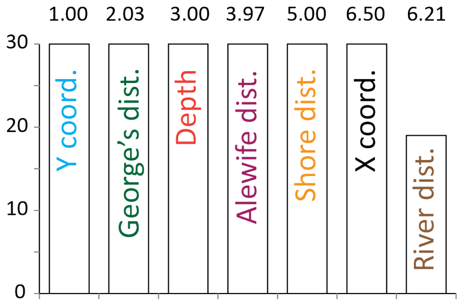
During the course of research I could also take a mile deep dive into machine learning techniques to answer some of questions stated above as well as to introduce a new approach to data driven investigations. These techniques move beyond traditional regression estimations where all else is held constant and address the determinants of a dependent variable all factors considered. The following figure visualizes the region of focus. The data at hand contains around 2.6 million coordinates and 21 bio-physical and economically relevant variables such as distance to port which can be used as a proxy for valuation of fisheries by fishermen. The results reported in this document are just for region 5 but a similar methodology was applied for all regions.



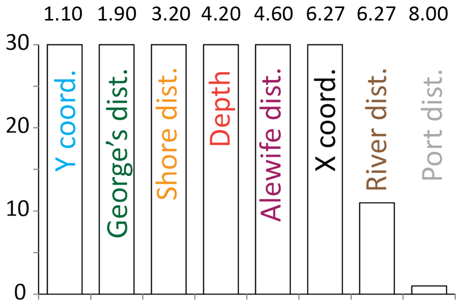
Region 5

Variable Selection and Importance

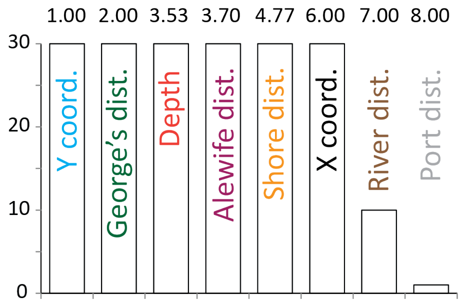
The first question to ask: what covariates seem to explain whether or no a spot is a fishing spot? To answer this we used random forest algorithms on data for region 5. Random forests can be used to rank the importance of variables in a regression or classification problem. The R function VSURF uses random forests to select the subset of independent variables that best interpret an outcome of interest, in this case whether or not a spot was a cod fishing spot in any season, in the spring, in the summer, in the fall, or in the winter. Because we ran 30 random forest algorithms for each unique dependent variable the most a variable can be selected as important to dependent variable interpretation is 30 times. The number above each bar gives the variable's average importance rank in the iterations it was selected.



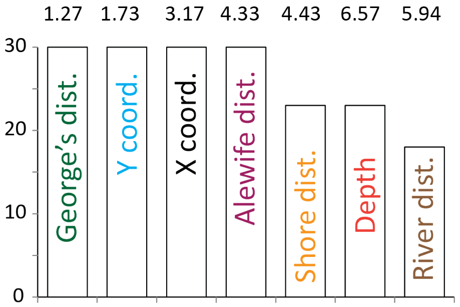
Any



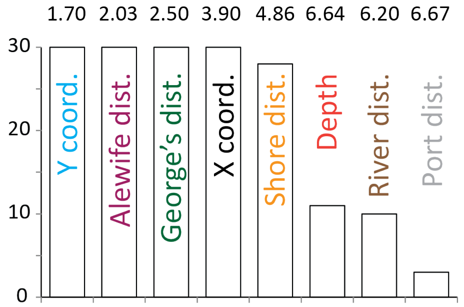
Spring



Summer



Fall



Winter

Estimating a model of fishing behavior
Next we estimated the following model,

$$\begin{aligned} Y = \alpha &+ \beta_1 Slope + \beta_2 Slope^2 + \beta_3 Depth + \beta_4 Depth^2 + \beta_5 George's\ Bank\ Dist. \\ &+ \beta_6 George's\ Bank\ Dist.^2 + \beta_7 Port\ Dist. + \beta_8 Port\ Dist.^2 + \beta_9 Shore\ Dist. \\ &+ \beta_{10} Shore\ Dist.^2 + \beta_{11} River\ Dist. + \beta_{12} River\ Dist.^2 \\ &+ \beta_{13} Hist.\ Alewife\ River\ Dist. + \beta_{14} Hist.\ Alewife\ River\ Dist.^2 \\ &+ \beta_{15} Season\ Lag + \varepsilon \end{aligned} \tag{1}$$

with logistic regression. In this case Y_{is} is equal to 1 if spot *i* is a fishing spot in season s and equals 0 otherwise. Season *s* includes any (it is a fishing spot in at least one season), spring, summer, fall, or winter. We estimate (1) once for each season where $\beta_{15} Season\ Lag$ is included and again when it is not included. *Season Lag* is given by Y_{is} for the previous season.

Ultimately we choose to ignore the estimates of model (1) that include $Season\ Lag$ because in those models *Y* is predominately explained by $Season\ Lag$. The impact of the other covariates on fishing behavior are masked in this version of the model.

While the estimated coefficients themselves are hard to decipher (see Table 1) in isolation we give two additional analyses to help interpret them. First because we include the quadratic form of each variable we are able to indicate the value of a variable that maximizes, or in a few cases, minimizes the likelihood of a spot being a fishing spot, all else equal. See Table 2. In Figure 2 we maps the expected values of Y_{is} , given by $\widehat{Y}_{is}$ for all *s*.

Table 1: Estimated Logistic Models (Y = {0,1}) for Region 5

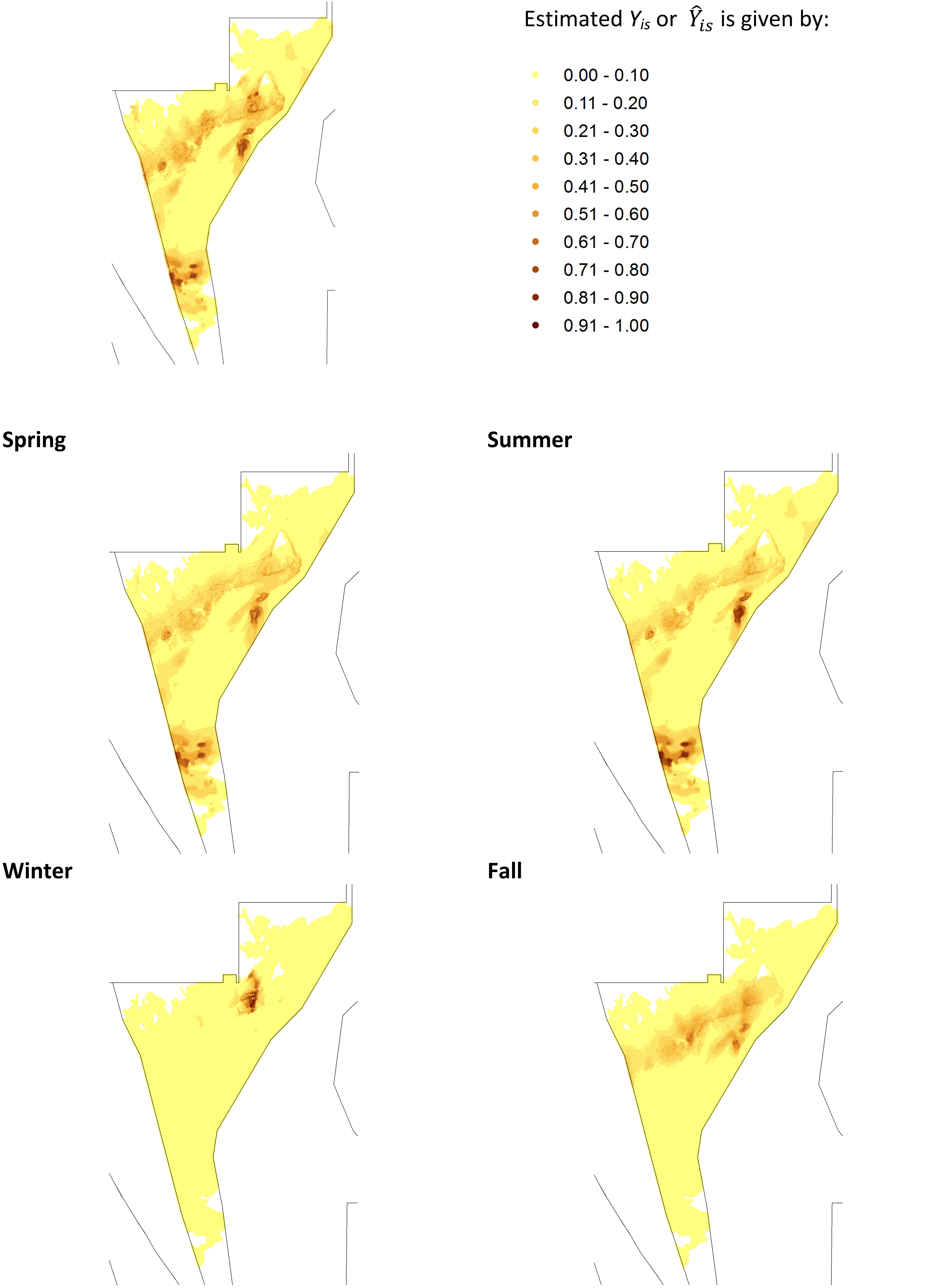
	Any	**Spring**		**Summer**		**Fall**		**Winter**	
Constant	-10.4672	-7.5433	-1.7504	-9.5735	-16.2249	-44.9991	-26.6343	-166.1628	-219.1617
	(0.334)	(0.3621)	(0.4157)	(0.3646)	(0.6549)	(0.697)	(0.946)	(1.5673)	(2.9875)
Slope	0.5834	0.6364	0.5647	0.4619	0.0885	0.3803	0.4173	1.3178	0.6511
	(0.0137)	(0.0142)	(0.0147)	(0.015)	(0.0245)	(0.0171)	(0.025)	(0.0529)	(0.0729)
Slope Sq.	-0.0738	-0.0732	-0.052	-0.0393	0.0165	-0.0681	-0.106	-0.3662	-0.1208
	(0.0036)	(0.0036)	(0.0036)	(0.0038)	(0.006)	(0.0047)	(0.0069)	(0.019)	(0.0229)
Depth	-0.0403	-0.037	-0.0453	-0.0484	-0.0515	-0.0072	0.0276	-0.1231	-0.2465
	(0.0004)	(0.0004)	(0.0005)	(0.0005)	(0.0008)	(0.0007)	(0.0009)	(0.0034)	(0.0053)
Depth Sq.	-0.0003	-0.0002	-0.0003	-0.0003	-0.0003	-0.0001	0.0001	-0.0011	-0.0015
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
George's Banks Dist.	0.0524	0.0313	-0.0159	0.0283	0.0401	0.3264	0.2036	1.1232	1.3239
	(0.0025)	(0.0028)	(0.0032)	(0.0027)	(0.0048)	(0.0054)	(0.0075)	(0.0114)	(0.0198)
George's Banks Dist. Sq.	-0.0001	-0.0001	0.0000	-0.0001	-0.0001	-0.0007	-0.0005	-0.0021	-0.0023
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Port Dist.	0.2244	0.1331	0.1782	0.1689	0.1987	-0.0897	-0.0301	0.246	0.8821
	(0.003)	(0.0033)	(0.0034)	(0.0035)	(0.0054)	(0.0065)	(0.0072)	(0.0164)	(0.0246)
Port Dist. Sq.	-0.002	-0.001	-0.0012	-0.0014	-0.001	0.0063	0.0057	-0.0128	-0.0266
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0005)	(0.0007)
Shore Dist.	-0.3457	-0.1945	-0.2818	-0.279	-0.3567	-0.012	-0.1307	0.1733	0.2163
	(0.0035)	(0.0038)	(0.0043)	(0.0041)	(0.0062)	(0.0061)	(0.007)	(0.0158)	(0.0197)
Shore Dist. Sq.	0.0049	0.0031	0.0031	0.0049	0.0063	-0.0043	-0.0023	0.0114	0.0127
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0005)	(0.0007)
River Dist.	0.0831	0.0907	0.2122	0.1138	0.116	0.0109	0.1797	-0.55	-1.6055
	(0.0024)	(0.0028)	(0.0034)	(0.0028)	(0.0049)	(0.0047)	(0.0059)	(0.0123)	(0.0233)
River Dist. Sq.	-0.0022	-0.0017	-0.0024	-0.0028	-0.0036	-0.0019	-0.0051	0.0075	0.029
	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0003)	(0.0005)
Hist. Alewife River Dist.	0.1503	0.0695	-0.0129	0.1556	0.2223	0.3005	0.1833	0.8148	1.1756
	(0.0021)	(0.0024)	(0.003)	(0.0024)	(0.0037)	(0.0052)	(0.0061)	(0.0125)	(0.0206)
Hist. Alewife River Dist. Sq.	-0.0009	-0.0003	0.0001	-0.0012	-0.002	-0.0029	-0.0017	-0.0106	-0.017
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0001)	(0.0001)	(0.0002)	(0.0003)
Season Lag			4.764		5.7452		4.9105		10.4623
			(0.0356)		(0.0206)		(0.0229)		(0.1378)
McFadden's R^2	0.23	0.21		0.27		0.32		0.54	

Table 2: Inflection Points in Region 5 Estimated Models. Bolded values indicate a value that maximizes the likelihood of a spot being a fishing spot, all else equal. Un-bolded values indicate a value that minimizes the likelihood of a spot being a fishing spot, all else equal. A black cell means there is no minimum or maximum value.

		Spring		**Summer**		**Fall**		**Winter**	
	Any	**No Lag**	**Lag**	**No Lag**	**Lag**	**No Lag**	**Lag**	**No Lag**	**Lag**
Slope (degrees)	**3.95**	**4.34**	**5.43**	**5.88**		**2.79**	**1.97**	**1.80**	**2.70**
Depth (meters)	**-74.07**	**-74.63**	**-84.50**	**-72.22**	**-78.82**	**-36.27**	-226.26	**-57.65**	**-84.98**
River Dist. (km)	**18.88**	**26.26**	**43.52**	**20.46**	**16.11**	**2.80**	**17.68**	36.85	27.72
George's Banks Dist. (km)	**176.35**	**160.10**		**166.24**	**267.37**	**236.91**	**211.15**	**266.50**	**287.21**
Port Dist. (km)	**57.08**	**49.54**	**71.49**	**60.22**	**96.50**	7.07	2.62	**9.63**	**16.61**
Shore Dist. (km)	35.6	31.45	45.93	28.53	28.31				
Hist. Alewife River Dist. (km)	**79.86**	**100.25**	54.72	**66.56**	**54.44**	**50.98**	**54.74**	**38.57**	**34.59**

The coefficients above enable us to predict the likelihood of being a fishing spot for all 2.6 million coordinates given the 14 variables we use as independent variables.

Figure 2



Any season

Spring

Summer

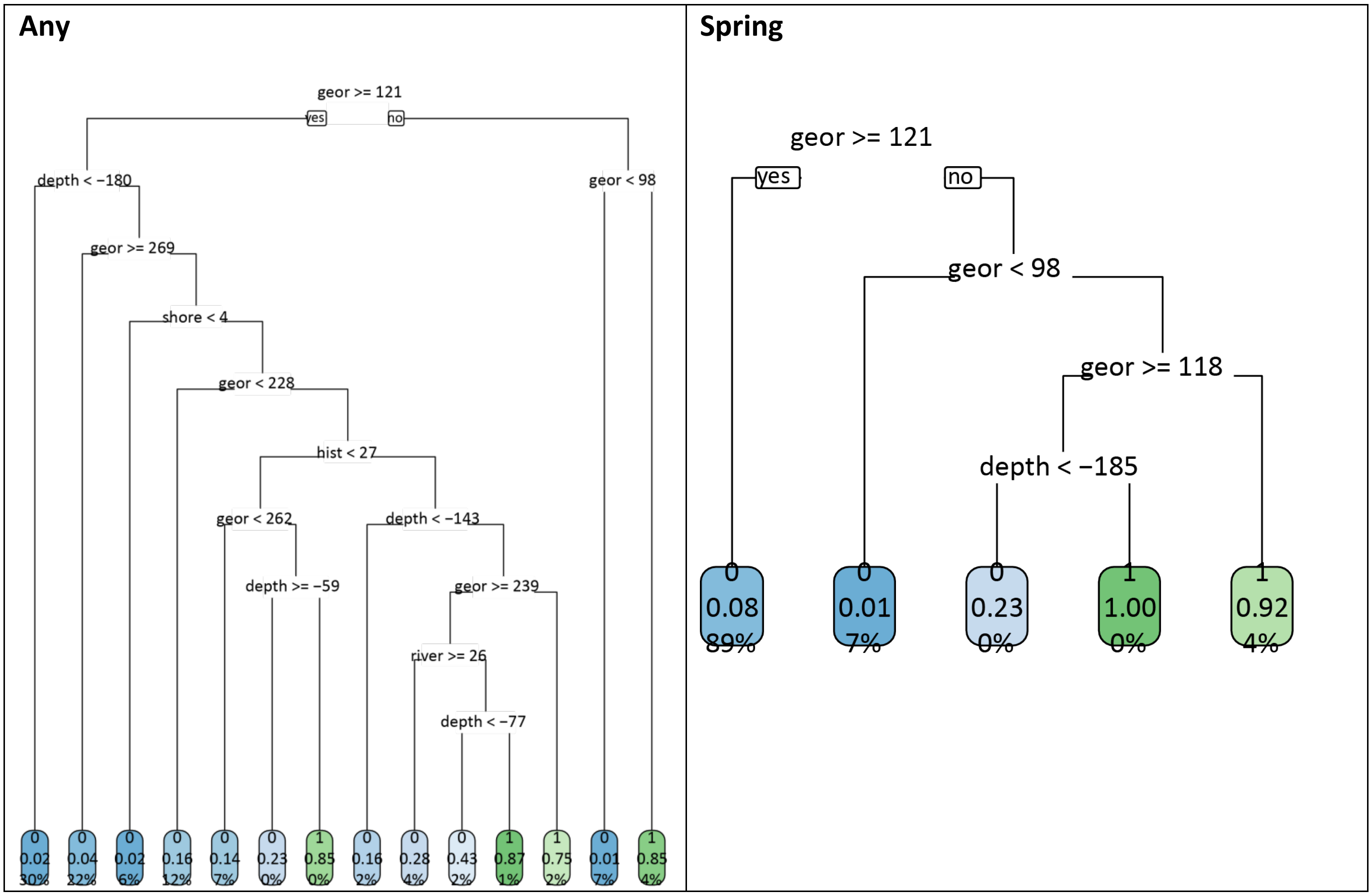
Winter

Fall

Classification Trees

With linear regression we are trying to explain or predict what makes a fishing spot. What covariate combinations make it more or less likely a spot was using for cod fishing? Classification trees can also be used to predict the combination of covariates that are associated with fishing spots. Using the same variables we used in model (1) (expect we do not include the quadratic terms) we build “pruned” decisions trees for each season in region 5.

Before building the decision we split the region 5 dataset is split into two. The “training” subset is used to find the different combinations of covariate values that are associated and not associated with fishing spots. The trees for any season and spring in region 5 are given below.



Any

Spring

The numbers in each leaf are 1) the value of Y, 2) the percentage of spots with these attributes that take on the given Y value, and 3) the percentage of all spots with these attributes.

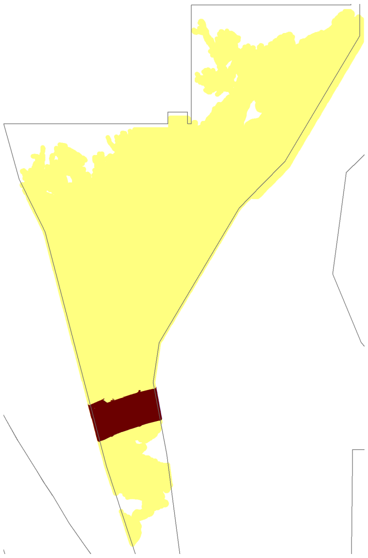
Then we test the decision tree with the other half of the data and calculate each tree’s error rate: how many times does the tree predict the opposite of the actual value. See Table 3.

Table 3. Classification Tree Error for Region 5

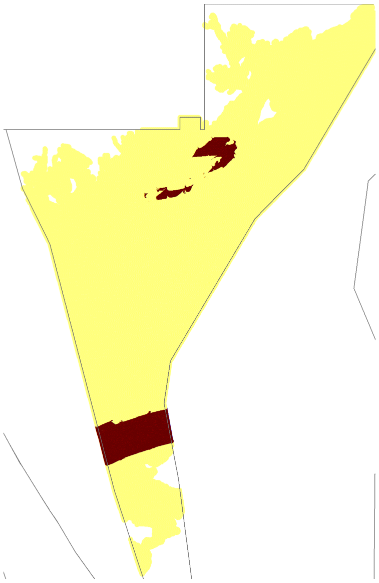
Season	Error
Any	0.082
Fall	0.108
Summer	0.082
Spring	0.100
Winter	0.122

We can also generate maps of the tree's predicted fishing spots. Where the red regions are predicted to be fishing spots by the classification trees.

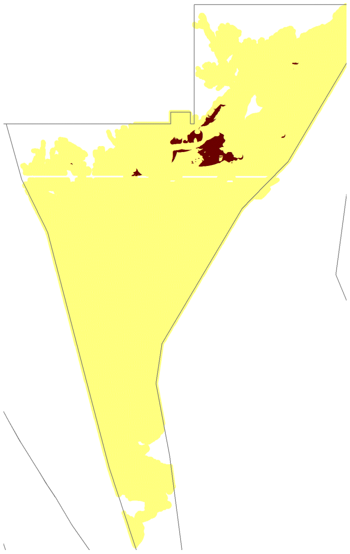
Spring



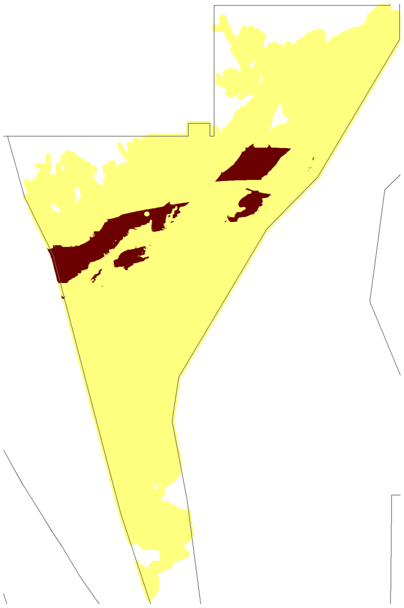
Summer



Winter



Fall



Variable Selection and Model Estimation (LASSO)

The underlying model is a linear probability model. For example, for every additional 1 km closer to an historical Alewife river, the likelihood a region 5 point is a fishing point in summer decreases by 0.18%, all else equal.

Variables	Any	Spring	Summer	Fall	Winter
Constant	-11.1581	0.1047	-15.4441	-18.9491	-0.1334
Slope (degrees)	0.0496	0.0487	0.0358	0.0328	0.0047
Depth (meters)	0.0012	0.0009	0.0011	0.0004	0.0003
River Dist. (km)	-0.0057	-0.0017	-0.0056	-0.0043	-0.0040
George’s Banks Dist. (km)	-0.0089	-0.0041	-0.0082	-0.0093	-0.0023
Port Dist. (km)	NA	-0.0010	-0.0002	0.0022	NA
Shore Dist. (km)	-0.0029	-0.0007	NA	-0.0046	0.0009
Historic Alewife River Dist. (km)	0.0023	NA	0.0018	NA	0.0002
X Coordinate (km)	0.0026	0.0019	0.0020	0.0020	0.0013
Y Coordinate (km)	0.0025	NA	0.0034	0.0041	NA
Minimum CV Error	0.106	0.094	0.088	0.059	0.016