Data Analysis Project

EN.625.603.81 Spring 2022 Module 13

Shelby Golden

5/3/2022

Introduction

The rise of anthropogenic induced environmental changes has precipitated a harrowing trend of habitat loss and shrinking biodiversity. Aquatic life forms are particularly sensitive to changes in their habitat, which can fluctuate for an entire area due to a lack of microclimate formation, like those seen on land. Scientists set out to observe these habitats by measuring factors like temperature, salt content, and the water depth species are found.

Water depth has a strong influence on how heat distributes in an aqueous environment. With the rise of surface ocean temperatures, scientists have observed fish are spending more time deeper than normal. Freitas et al. notes that this behavioral thermoregualtion can deprive species of their regular access to resources, like food. Diving deeper also comes with a reduction in local soluble oxygen that certain species may be less resilient to. Thus, rising surface temperatures can trap sensitive species in smaller and smaller habitable environments that fit their needs.

Changes in aquatic localization and biodiversity challenges the structure of an ecosystem and can lead to a myriad of cascading effects. This paper will examine some of these well established trends. Specifically, changes in biodiversity by area and depth, changes in the probability select species are found at different depths over time, and a soft evaluation of temperature changes by region over 50 years.

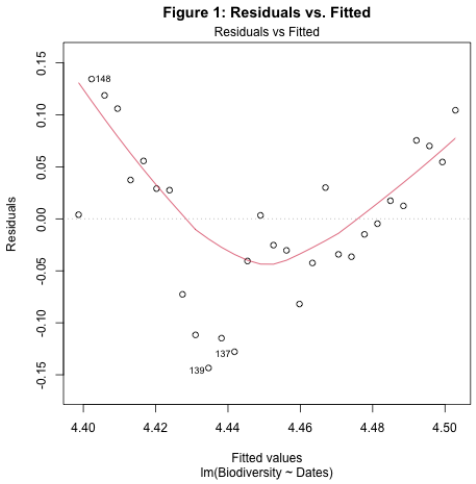
Data Description

In this paper I will be using data from OceanAdapt, which is a collaborative research initiative that includes Pinsky Lab of Rutgers University, the National Marine Fisheries Service (NMFS), and Fisheries and Oceans Canada (DFO). The data has been routinely collected by the NMFS, DFO, NOAA and other agencies over the past 50 years to monitor fish and invertebrate populations in the waters surrounding North America and Canada.

It only includes regions where consistent survey methods were employed and does not include coastlines in an effort to "prevent poleward shifts in distribution". It also only includes species that are caught every year so as to prevent species composition from being a trend effector.

Unfortunately, the data set focusing on species did not clearly indicate the season the observation occurred. Consequentially, seasonal shifts are not granularly evaluated in the species data, and are only generally related to surface temperature changes by year. Latitude and longitude pairs of a given observation were categorized into one of three regions: "Northeast US", "Southeast US", and "Scotian Shelf".

To address the question about how species localize at different depths over time, four species with the most observations over the years (> 275) were identified for further in-depth analysis. These fish also showed a range of observations at different depths. Species: Leucoraja erinacea, Merluccius bilinearis, Squalus acanthias, and Urophycis chuss.



It is entirely possible that they migrate to escape temperature fluctuation, or may naturally migrate around the coast. Indeed, some of the selected species have scientific literature indicating that they migrate up and down the east coast. Despite these possible effectors, because I will be evaluating the changes of depth localization via Bayesian analysis, I am hopeful that it will still indicate general trends of depth localization.

Statistical Methods

Changes in biodiversity (reported as a Shannon Biodiversity Index ($SBI = -\sum_{i=1}^{s} p_i \ln(p_i)$) over 50 years is shown in the key findings section (Figure 2), where each region has their resultant linear fitting plotted. From what I could find about lm() in R, it uses the Least Squares Method to fit linearly related random variables (Madrury, GitHub).

The residual plots for the Northeast and Scotian Shelf linear models look heteroscedastic, one of the requisite conditions for a linear fitting. The Southern region residual plot has a distinct shape that suggests the data may not be linearly related (Figure 1). Because of this, its linear model's slope is not reported and it's values will be excluded from linear modeling with temperature changes.

The probability a fish is found at different depths can be seen as a multivariate binomial distribution, where each $X = 0, 50, 100, ..., 450$ meters. In this paper I will evaluate it as a simple binomial distribution by looking at the marginal PDF. In this case, one depth level will be categorized as a "success" while all the remaining are seen as a "failure". As far as can be discerned regarding the data collection, observations are each random and independent.

	Date	n	sum(pHat)	pHat50	Observed	Estimated
1	1975	323	1.000	0.390	126	125.9
2	1980	532	1.000	0.421	224	226.7
3	1985	503	1.001	0.328	165	168.5
4	1990	507	1.001	0.375	190	196.0
5	1995	483	0.999	0.352	170	167.7

Table 1 shows a snapshot of the probability that Merluccius bilinearis was observed at 50 meters below. The $\hat{p}$ column is calculated using the maximum likelihood estimator for a binomial distribution, $\sum k_i/n$. The estimated column is the expected observed frequency of fish if $\hat{p}$ was the actual probability for a binomial distribution with the respective n observations. The observed and estimated columns show good alignment, which suggests that the data can be assumed to be binomially distributed.

Were we to assume $\hat{p}$ is constant, then a proportions hypothesis test could be employed; $H_o : p_x = p_y$. In order to do this hypothesis test on binomial random variables, X and Y, their proportions, X/n and Y/m, need to be normally distributed to satisfy the Generalized Likelihood Ratio Test (GLRT). As is evident from Figure 3.A., this key requirement not likely to be satisfied.

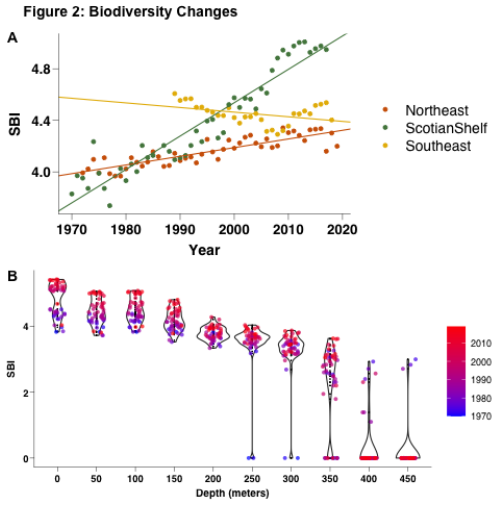
Another method to evaluate changes in $\hat{p}$ over time is through Bayesian analysis. As described above, the probability that a fish is located at a given depth is reasonably binomially distributed. Intuitively, we know that $\hat{p}$ is not constant, and has its own distribution. It will be referred to as θ going forward in the context of this statistical test. In order to leverage a Bayesian analysis I assumed θ to be a Beta distributed random variable. Utilizing the bootstrap technique, I randomly sampled from three similarly trending species (Figure 3.A.) in 1980 to estimate the prior parameters, r and s. For example, for a binomial distribution at 50 meters, $\hat{r} = 393.05$ and $\hat{s} = 505.168$.

As we did in Module 12, the prior distribution was "updated" to the posterior distribution with Beta parameters $r_{new} = r + k$ and $s_{new} = s + n - k$. This process was completed three times, each assuming depth = 50, 150, and 300 meters was a "success". The resultant shift in θ for 1980, 2000, and 2019 were overlayed for each depth analysed (Figure 3.B.). The analysis for observances at 300 meters did not yield a probability distribution. This could be attributed to how infrequently these species visit 300 meters below the water surface. The probability distribution is therefore not displayed.

In order to establish a relationship between surface temperature, region, and year observed two methods were employed: contingency tables for linear independence and linear modeling for trend analysis. Although R has a package for evaluating contingency tables, I calculated this "by hand", so to speak, in the manner that was demonstrated in lecture series 10.I. Earlier in the paper I showed that the the data collected in the Southeast region may not be linearly related, due to a homoscedastic residual plot (Figure 1). Thus, its data is not evaluated.

Key Findings

Changes in biodiversity by area and depth

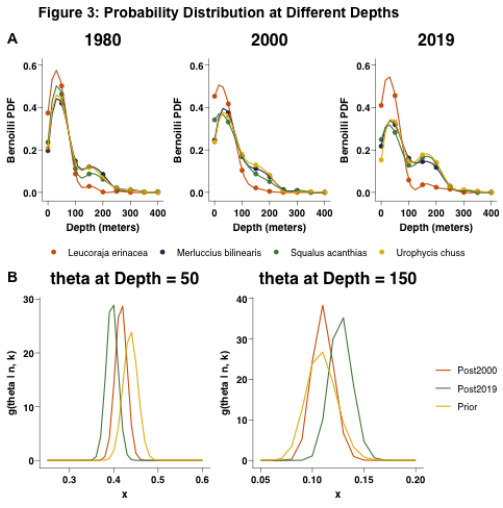


Linear modeling lm(Biodiversity ~ Dates) for each subset of data by the three regions yielded the plot shown in Figure 2.A.. Northeast US biodiversity increased by 0.0067 every year with an $r^2 = 0.745$. Biodiversity in the Scotian Shelf had a more dramatic expected increase at 0.0259 each year and an $r^2 = 0.902$. As was stated earlier, the Southeast region residuals plot indicated that the data may not be linearly related. To further this point, its $r^2 = 0.155$.

Thus, a general increase in biodiversity is associated by year for the Northeast and Southeast. A greater change is observed in the Scotian area compared to the Northeast, at a 3.8 fold difference. Later in this paper these trends will be related to temperature changes over the 50 years recorded.

By visual inspection of the scatter plots, it almost looks like the linear fitting can be further decomposed by time series. This is beyond the scope of my knowledge and was not attempted here. However, a time series decomposition of the linear model could account for additional variation left unexplained with the current fitting as indicated by the model r^2 values.

Very simply, I generated violin plots showing the biodiversity change by depth (Figure 2.B.). On top of each violin plot are scale colored points showing how SBI changes over 50 years. As we intuitively expected, biodiversity decreases the further below the ocean surface we measure. Interestingly, we see that there is some divergence in biodiversity counts by year at 0-50 meters. Here, we see that biodiversity at this depth increased from 1970 to 2010 enough to give the violin plot a bi-modal shape. The two regions spanning 50-150 meters show a slight skew indicating that biodiversity increased at these levels over time as well, but that it was not as pronounced at the 0-50 meter range.



Changes in species depth location over 50 years

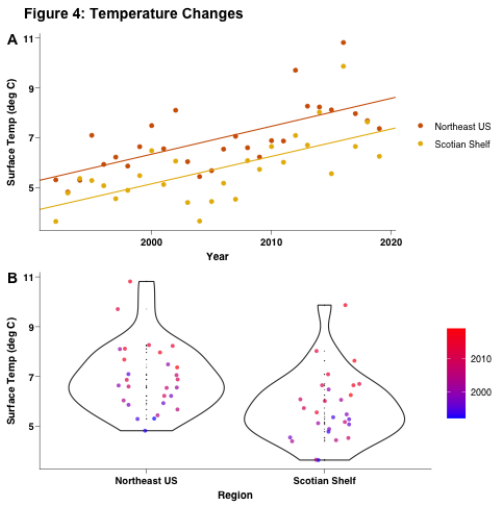
In Figure 3.A. the multivariate probability that the four species of focus are plotted and fitted with a spline curve to emphasize the distribution shape. In general, three fish, Merluccius bilinearis (blue), Squalus acanthias (green), and Urophycis chuss (yellow), trend similarly by depth and even across time. In

order to estimate the random variable θ, I assumed that these fish depth-localization behavior are similar enough to demonstrate how θ varies to the Bootstrap algorithm.

Figure 3.B. shows how θ changes after additional data is considered. The initial, yellow, distribution represents the Beta Prior using the Bootstrap estimated parameters $\hat{r}$ and $\hat{s}$ informed by the three similarly trending species in 1980. For simplicity, I only evaluated changes in θ for Merluccius bilinearis. Each additional posterior variables (n and k) are the total number of observations and number of successful observations seen at each depth in 2000 and 2019.

Based on the observations made by many scientific groups, we would expect that the probability that Merluccius bilinearis is found at higher depths decreases, while lower depths increase. The plots shown in Figure 3.B. suggest the same trend, especially for the probability that we would find Merluccius bilinearis at 50-100 meters. The expected value for θ at 50-100 meters changes from around 0.42 in 1980 to 0.40 in 2019. The expected value for θ at 150-200 meters changes from around 0.11 in 1980 to 0.14 in 2019.

Temperature changes by region over 50 years



In order to avoid the normal temperature swings seen seasonally, the relationship between surface area and location or year were evaluated for data collected in either the fall or spring. No other seasons were captured in the data set. Residual plots for fall data appeared homoscedastic, and so only data collected in the spring is linearly modeled.

The independence of temperature and area was evaluated using contingency tables. Here, each temperature measurement was categorized into first, second, third, or fourth quantile based on the spread of temperature observations in each region. This procedure was repeated for both data collected in the spring and fall.

Fall: critical value = 11.34, test statistic = 1.8491517×10^5

Spring: critical value = 11.34, test statistic = 1.9570095×10^5

Clearly, we would reject the null hypothesis in both cases, meaning that temperature and region are NOT independent. This affirms what we would intuitively expect to see. As can be seen from the test statistic, it is thousands of times higher than the critical cut-off value. I wonder if there is a limitation computing contingency tables over thousands of observations, as was completed here?

In both plots contained in Figure 4 we can see a clear association between surface temperature and the year. The change between the areas is relatively close. Northeast slope: 0.1121 with $r^2 = 0.467$ and Scotian Shelf slope: 0.1099 with $r^2 = 0.445$. A null hypothesis comparing these slopes was not completed in this analysis.

Here the fitting of both models is pretty underwhelming. Based on some oscillating temperature changes, it is possible that certain year's data collection were closer to winter or summer compared to others. Additionally, there appears to be the potential of further time series decomposition of these data, which is not completed here.

Figure 4.B. affirms what is shown in Figure 4.A., that the oceans surface temperature has increased over the 50 years observed.

Discussion

In this paper, I sought to affirm the well established scientific findings of biodiversity change by depth over time. Around this, I sought to show that sea surface temperatures have increased over the 50 years observed,

and that biodiversity changes and sea temperature increase is observed in different regions.

Starting out, it seemed contrary to see biodiversity consistently increase over the span observed. It's plausible that improved data collection on the part of OceanAdapt can explain this. However, if that were entirely the case, I would expect to see an asymptotic curve plateauing at the point where data collection methods did not change much afterward. This is not what is seen. The SBI index instead remains starkly linearly related by year without transformation, and with high r^2 statistics.

In marine biology communities, biodiversity is expected to increase with reasonable increases in temperature (Freitas et al.). This comes down to factors like metabolism and reproduction which are temperature dependent (Sinclair et al.). Even though it is not directly analyzed here, it makes sense that the linear dependence of biodiversity over time is statistically associated with increases in surface temperature. This was not directly analyzed here because the data set for species only stated the year observed and not the time of year, and thus could not be related back to the data set about sea temperatures without ignoring seasonal changes.

The analysis sections evaluating changes in depth affirm the well established behavioral thermoregualtion observation, where fish migrate to deeper waters as the surface temperature increases (Freitas et al.). Here I randomly selected species purely based on how many observations were recorded for each. It would have been interesting to further this by selecting a species that is known to not migrate or consistently localizes within a small depth range. The species analyzed here localized anywhere from 0 to 250 meters below, which might have watered down any general trends to localize at lower depths. Still, a shift, albeit slight, is clear from the Bayesian analysis.

Acknowledgments

I'd like to acknowledge the scientists who have contributed generating the OceanAdapt database and make it available for free. The R community, who shares tips on StackExchange, continues to maintain the platform, and expand its operative tools with indispensible packages.

I'd also like the thank the Dr.'s Bodt and Savkli for putting together the lecture material for the course and the authors of our textbook for such a comprehensive and informative compilation of statistics.

I'd especially like to thank Dr. Wang, who was our dedicated professor for the semester, and offered her expertise and insights throughout the semester.

Reference Links

Data Source: Pinsky Lab at Rutgers Unigersity, Department of Ecology, Evolution, and Natural Resources : Data Download Link

1. How to Calculate Biodiversity, University of Florida
2. Bootstrap Sampling Distribution and Confidence Intervals, Missouri State University
3. Parameter Estimation for the Beta Distribution by Claire Elayne Bangerter Owen
4. Bayes Programming in R by Minin et al.
5. A Deep Dive Into How R Fits a Linear Model
6. Can we predict ectotherm responses to climate change using thermal performance curves and body temperatures? DOI: 10.1111/ele.12686
7. Sea temperature effects on depth use and habitat selection in a marine fish community DOI:10.1111/1365-2656.13497