**Background**

Fish and crustaceans coexist within marine ecosystems, often exhibiting patterns of co-variation in both their spatial distribution and abundance over time. Given this interdependence, it stands to reason that groups of species may share common trends in their presence and abundance, raising an important question: To what extent can the observed abundance of a subset of species be used to predict the abundance of others? This question is especially interesting when put in the context of abundance modeling. Traditional monitoring of marine species relies on systematic sampling at specific locations and times, with the goal of constructing models of overall abundance. However, such monitoring is both time-consuming and costly. Even in cases where all species are collected in the same trawl, sorting and identifying each species demands additional effort. Therefore, if some species could serve as proxies for the others, it would allow for a reduction monitoring effort while still retaining useful ecological information.

**Problem Statement**

This problem can be broken into two primary components: occurrence and abundance. The occurrence question focuses on identifying groups of species that tend to appear in the same locations, regardless of their densities. The abundance question, on the other hand, examines whether species that co-occur do so in a way that exhibits consistent covariance in their population densities.

Therefore, our first two questions are whether there exist (potentially different) groupings of species that co-occur and species that covary in abundance.

This however is not enough to answer our overall question of whether groups of species can stand in as proxies for others. To answer this, we must explicitly assess how shifting species from the “monitored” pool to the “predicted” pool impacts the total information gained from monitoring.

Here, "information" refers to the extent to which predictions of species abundance match actual observations. In the occurrence case, determining the effectiveness of predictions requires establishing a confidence threshold—how certain must the model be before declaring a species present? This is a practical consideration, as lowering the threshold improves recall (capturing more occurrences) but decreases precision (increasing false positives), while raising the threshold has the opposite effect. This tradeoff can be evaluated using receiver operating characteristic (ROC) curves, where the area under the curve represents the prediction performance. In the abundance case, the proportion of explained variance serves as a direct measure of prediction performance.

From this we arrive at the second set of questions – how the information diminishes as a function of moving species from the “monitored” to “predicted” categories in each of our occurrence and abundance groups.

Finally, to understand how the *monitored* species provide information specifically we must provide a baseline of how much information is contained in a model containing environmental and sampling characteristics only.

In summary, our goal is to assess how well a subset of species can serve as proxies for the occurrence and abundance of a broader set of species. We will do this by answering the following:

1. Do there exist (potentially different) groupings of species that co-occur and species that covary in abundance?
2. For each of these groupings, how does the information captured diminish as the “least informative” species are removed?
3. How much information is captured in models that include only environmental and sampling characteristics, providing a baseline for comparison?

From this we will then be able to assess, as a sensitivity, how much information is gained from monitoring a full superset of species versus a specific subset.

Undertaking this study requires a comprehensive dataset spanning multiple years and a wide range of environmental conditions. The bottom trawl surveys conducted in the Eastern Bering Sea (EBS) and the Gulf of Alaska (GOA) provide an ideal dataset for this purpose, offering extensive species occurrence and abundance data collected over time.

**Data**

The Alaska Fisheries Science Center’s Groundfish Assessment Program conducts annual bottom trawl surveys to estimate biomass and population counts for commercially important fish and crab species. These surveys are carried out with the assistance of several organizations, including the Alaska Department of Fish and Game (ADF&G) and the International Pacific Halibut Commission.

Survey coverage varies by region. The Eastern Bering Sea (EBS) is surveyed every year, whereas the Aleutian Islands (AI) and Gulf of Alaska (GOA) are surveyed biennially, alternating between the two regions. Additionally, there is a biennial survey of the Northern Bering Sea (NBS) and occasional surveys of the Eastern Bering Sea Slope. These efforts provide long-term data on species distributions and population trends across a wide geographic area.

The surveys are conducted between May and September, during which trawl vessels follow predetermined sampling stations. Each trawl involves towing a bottom trawl net at 5–6 km/h for either 15 or 30 minutes, after which the entire catch is brought aboard and sorted by species. For each species, weights and counts are recorded. However, species identification is not always definitive, and so observations include taxonomic confidence indicators and are sometimes grouped into higher taxonomic levels, such as genus.

In addition to species counts and biomass estimates, environmental data is recorded. Most hauls include bottom and surface temperature measurements, and all hauls include fishing depth data. Beyond this each haul is also assigned a stratum, a categorical designation used in stratified sampling estimates to account for environmental and geographic variability across survey locations. These strata help ensure that estimates of species abundance and biomass are representative of broader ecosystem patterns. This environmental information provides context for species distributions and abundance trends.

The bottom trawl surveys have been conducted annually since 1982, with the exception of 2020, when the survey was canceled due to COVID-19 restrictions. The data is publicly available and can be accessed via an online data portal, as well as through dedicated R and Python packages (the latter of which was used for this study).

**Preliminary Exploratory Analysis**

The dataset accessed through the Python API contains a wide range of columns, covering taxonomic, cruise, station, year, survey, weight, haul, count, environmental, and catch-per-unit-effort (CPUE) data. In this section, we describe the features selected for analysis, briefly explain the rationale behind excluding certain columns, and then proceed to explain how we selected species for the analysis.

Each haul is uniquely identified by the columns year, srvy, station, stratum, and haul, which are of integer, string, string, integer, and integer types, respectively. Except for year, which is an ordinal variable, these are all categorical identifiers.

To standardize across varying levels of fishing effort, we selected cpue\_kgha (catch per unit effort in kilograms per hectare) as our primary abundance measure. This variable is continuous but contains a substantial number of zero values, reflecting hauls where a given species was absent. Using CPUE allowed us to remove additional effort-related features that would otherwise be required for standardization.

Each haul also includes several environmental and operational variables:

* Bottom and surface temperature (°C) – continuous variables
* Haul depth (m) – continuous variable
* Net width and height (m) – continuous variables (net height was excluded due to missing values in earlier years)
* Haul duration (hours) – continuous variable
* Distance fished (km) – continuous variable (excluded, as it was redundant with haul duration given the consistent 5–6 km/h trawl speed)

Since net width is effectively captured by the area swept during a haul, it was also excluded from the analysis.

Geospatial and temporal columns, including position and datetime, were omitted from the analysis, as the focus of this study is on species co-variation rather than fine-scale spatio-temporal modeling beyond what species covariance inherently captures.

As a further description of the haul specific environmental and operational covariates we found:

* Haul duration (duration\_hr) follows a bimodal distribution, with peaks at 15 and 30 minutes, corresponding to the two standard trawl durations.
* Surface temperature (surface\_temperature\_c) exhibits a slightly right-skewed unimodal distribution, centered around ~7°C.
* Bottom temperature (bottom\_temperature\_c) has a left-skewed unimodal distribution, centered around ~4°C.
* Depth (depth\_m) is highly right-skewed, with most hauls occurring at shallower depths.

All in all, this data contained 33,958 hauls spanning 42 distinct years, 7,861 stations, and 975 unique taxonomic designations.

**Species Filtering**

The scientific name column contained string values representing the identified taxon, which could include individual species, genera, or broader taxonomic groupings. The taxon confidence column, an enumerated variable, included the following levels: 'High', nan, 'Moderate', 'Low', and 'Unassessed', where nan corresponded to hauls with zero recorded abundance.

To ensure high-confidence species-level data, we applied the following filtering steps:

* Taxonomic Confidence Filtering – Species were retained only if at least 95% of non-zero abundance hauls were recorded with moderate or high taxonomic confidence, reducing the dataset to 339 distinct taxonomic designations.
* Removal of Multi-Species Designations – Entries that represented genus-level or other multi-species classifications were removed, reducing the set to 273 species.
* Minimum Presence Threshold – Species appearing in fewer than 5% of hauls were excluded, leaving a final dataset containing 66 species.

**Summary**

The final dataset covered 33,958 distinct hauls over 66 distinct species and contained the following columns:

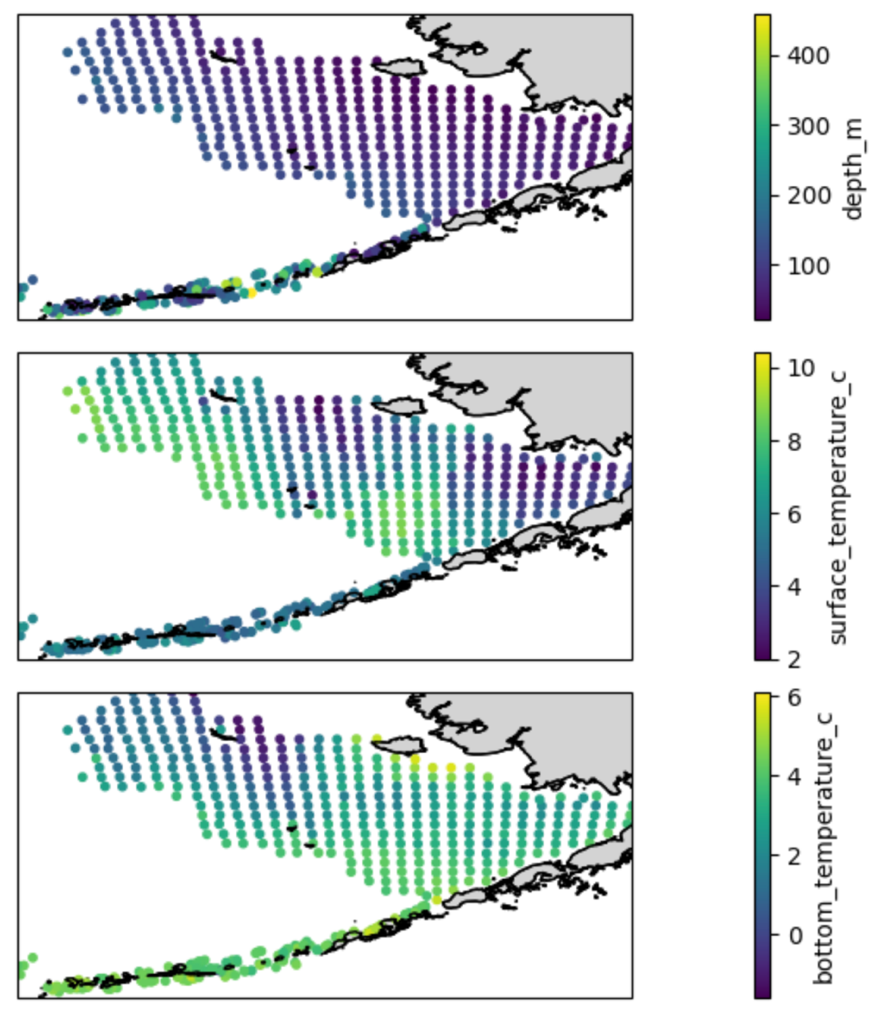
* Haul Identifiers: year, srvy, station, stratum, haul,
* Environmental and Effort Data: duration\_hr, surface\_temperature\_c, bottom\_temperature\_c, depth\_m
* Species Abundance: an additional column per species, indicating its cpue\_kgha value for that haul

**Issues**

A small portion of the dataset contains missing environmental data, with 4.7% of hauls lacking bottom temperature measurements and 2.5% missing surface temperature data. While this introduces some uncertainty, these gaps are relatively minor given the dataset’s overall size. Additionally, since we are treating species presence and abundance as separate problems, the high prevalence of zero values in CPUE data should not pose a significant issue—zeros will be informative in the occurrence analysis while abundance modeling will focus only on nonzero values.

None of the key numerical distributions follow a normal pattern, meaning that normalization or transformation of CPUE data will be necessary to improve model performance and comparability across species. However, the most significant challenge will likely be the sheer number of species retained even after filtering. Developing a robust, automated approach for handling outliers and data inconsistencies will be crucial, as manually reviewing each species is impractical. Finding effective ways to standardize and clean the dataset programmatically will be essential to ensuring reliable and scalable analysis. A map of the united states

AI-generated content may be incorrect.



A graph with a blue line

AI-generated content may be incorrect.

A graph with a blue line

AI-generated content may be incorrect.