**Investigation into Using Subsets of Species as Proxies for Broader Abundance Monitoring**

**Marcel Gietzmann-Sanders**

**Background**

Fish and crustaceans coexist within marine ecosystems, and it stands to reason that groups of species exhibit patterns of co-variation in both their spatial distribution and abundance over time. This raises an important question: To what extent can the observed abundance of a subset of species be used to predict the abundance of others? It’s a question made 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 in monitoring effort while still retaining most if not all the useful abundance and distribution 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, we will take a specific interpretation of "information" – the extent to which predictions of species abundance match actual observations. In the occurrence case, given this is a classification problem, we can use 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 *monitoring of* species provides 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.

**Data1,2**

The Alaska Fisheries Science Center’s Groundfish Assessment Program conducts annual bottom trawl surveys to estimate the 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 (Figure 1). 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. If the catch is small enough it is entirely sorted otherwise a standardized subsampling procedure is followed2. 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.

A map of the united states

AI-generated content may be incorrect.

*Figure 1: Stations by Survey. The positions of each of the stations in the 5 different survey areas used in the bottom trawl survey.*

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 restrictions3. The data is publicly available and can be accessed via an online data portal, as well as through dedicated R and Python4 packages (the latter of which was used for this study).

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

**Preliminary Exploratory 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 zeroes, 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 with a left-skewed unimodal distribution, centered around ~4°C for bottom temperatures and a slightly right-skewed unimodal distribution, centered around ~7°C for surface temperatures
* Haul depth (m) – continuous variable that is highly right-skewed, with most hauls occurring at shallower depths.
* Net width and height (m) – continuous variables that were not selected for this analysis (net height is missing in earlier years and net width is effectively captured by the area swept during a haul)
* Haul duration (hours) – continuous variable that follows a bimodal distribution, with peaks at 15 and 30 minutes, corresponding to the two standard trawl durations.
* Distance fished (km) – continuous variable (excluded, as it was redundant with haul duration given the consistent 5–6 km/h trawl speed)

A comparison of a graph

AI-generated content may be incorrect.

*Figure 2: Distributions of Selected Features. Histograms of depth, surface and bottom temperatures, and haul duration.*

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.

**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.

A diagram of different colors of the surface temperature

AI-generated content may be incorrect.

*Figure 3: Spatial Distribution of Features: depth, surface and bottom temperatures by station in the 2024 bottom trawl surveys.*

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.

In general, the abundance per species per haul (cpue\_kgha) is a highly right skewed distribution with a large number of zeros. See Figure 4 for some examples.

**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

A graph with a blue line

AI-generated content may be incorrect.

*Figure 4: Example Species Distributions. Distributions over non-zero hauls of the fourth root of cpue in kg/ha for the least common (by occurrence in hauls), 50th percentile common, and most common species (from left to right).*

**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.

**References**

1. Alaska Fish Research Surveys - <https://www.fisheries.noaa.gov/alaska/ecosystems/alaska-fish-research-surveys#gulf-of-alaska-bottom-trawl-survey>
2. Results of the 2023 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna - <https://repository.library.noaa.gov/view/noaa/61492>
3. The 2022 Eastern Bering Sea Continental Shelf Trawl Survey - <https://apps-afsc.fisheries.noaa.gov/plan_team/resources/draft_ebs_crab_tech_memo_2022.pdf>
4. AFSC GAP Python API - <https://afsc-gap-products.github.io/gap_products/content/foss-api-py.html>

**Postscript**

I have outlined five different questions that are needed to answer my overall question around the ability to use subsets of species to proxy supersets. This might seem like a rather large project to tackle and therefore I thought it would be pertinent to explain why I think this is rather straightforward.

The first two questions are where I bring in the tools discussed in this class – either by using clustering or dimensionality reduction techniques I will look for groupings of species that covary.

The latter three questions are where I bring in my own expertise to “put a bow” on the problem. Determining the information lost as we reduce the number of species in the subset should be as simple as the following:

1. Take each of the species not yet in the “predicted” category and build a random forest for that species given the species still in the monitored category + environmental and haul covariates.
2. Evaluate the “information” metric in question and select the species that causes the minimal reduction in information
3. Add that species to the “predicted” category and repeat from step 1 until there are no species in the monitored category (giving us our environmental model)

Building random forests is rather straightforward and so all this comes down to is building a loop that can work its way through each of the groups. With this done I’ll have all the information I need to report on each of the questions outlined above. Most of the real work for me will be in doing the clustering/dimensionality reduction to identify groups. I’ll just be applying <https://scikit-learn.org/stable/> to do the rest.