**Potential of Using Subsets of Species as Proxies for Broader Occurrence Monitoring**

**Marcel Gietzmann-Sanders**

Monitoring marine communities is costly and time-intensive, raising the question of whether a subset of species can serve as proxies for broader occurrence patterns. Using bottom trawl survey data from the Eastern North Pacific, I tested whether a species subset preserves co-occurrence structure in multidimensional analyses. Species were clustered using Jaccard distances, and their contributions assessed via Kruskal’s stress during stepwise removal. Approximately 60% of species were sufficient to retain key distance structures, and ordinations based on reduced sets closely matched those from the full set. Mantel tests confirmed that depth remained the dominant environmental gradient, consistent across full and reduced species sets. These findings suggest that substantial reductions in species monitored may be possible without sacrificing analytical fidelity—potentially streamlining future survey and modeling efforts.

**Introduction**

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 time1. This raises an important question: To what extent can the observed abundance of a subset of species be used to predict the abundance of others2? It’s a question made especially interesting when put in the context of abundance modeling. Traditional monitoring of marine species relies on sampling at specific locations and times, with the goal of constructing models of overall abundance1. 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 information2.

This problem can be divided into two main components: occurrence and abundance—that is, predicting when species occur and, when they do, their density. In this report, I focus on the first: occurrence. Our initial question is whether there exist groupings of species that consistently co-occur.

However, this alone doesn't answer our broader question: can some species serve as proxies for others? To evaluate this, I test how moving species from the “monitored” pool to the “predicted” pool affects the total information gained from monitoring.

Because multidimensional analyses depend on distance matrices, I define “information” as how well a subset of species preserves the original distance structure. Removing species acts like an ordination, and I use Kruskal’s stress to measure the distortion. By tracking how stress increases as species are removed, I assess their importance—though the key test is whether reduced sets still yield similar analytical results.

In summary, our goal is to assess how well a subset of species can serve as proxies for the occurrence patterns of a broader community in classic multidimensional analyses by (1) identifying groups of co-occurring species (2) measuring how stress increases as species are removed from these groups and (3) comparing analysis based on ordinations from a subset to those based on the full set.

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.

**Methods**

**Data3,4**

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 followed4. 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.

The EBS bottom trawl surveys have been conducted annually since 1982 except in 2020 due to COVID-195. GOA and AI surveys alternate years, BSS was biennial from 2002–2016, and NBS has been surveyed six times since 2010. The data is publicly available and can be accessed via an online data portal, as well as through dedicated R and Python6 packages (the latter of which was used for this study). All in all, this data contains 33,958 hauls spanning 42 distinct years, 7,861 stations, and 975 unique taxonomic designations.

I identified occurrence in a haul as catch per unit effort in kilograms per hectare being greater than 0. As environmental variables I selected (see Figure 2 for an example of spatial distribution):

* 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.
* Haul duration (hours) – continuous variable that follows a bimodal distribution, with peaks at 15 and 30 minutes, corresponding to the two standard trawl durations.

Rows with nulls in any of these columns were excluded from the analysis (32,043 hauls remaining). 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.

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, I 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.

**Clustering**

To identify groupings of species that covary in occurrence, I first generated presence–absence vectors for each species, where each value represented whether the species was present (1) or absent (0) in each haul. These were then aggregated by stratum and year to reflect presence or absence in each stratum–year combination. Pairwise distances between species were computed using the Jaccard distance metric7, which measures the proportion of differences relative to the total number of possible co-occurrences.

For each species, I calculated the maximum and 10th percentile Jaccard distances to all other species. The mean of the 10th percentile values provided a lower bound for expected within-group similarity, while the mean of the maximum distances provided an upper bound. I used the midpoint between these bounds as a threshold for agglomerative clustering (complete linkage) to define co-occurrence groups7. Clusters with fewer than five species were considered potentially spurious and grouped into a single “spurious” category. A permanova8 was conducted to test whether the resulting clusters differed significantly from random groupings.

Finally, I applied non-metric multidimensional scaling (NMDS)7 to the original species presence–absence vectors to produce a two-dimensional ordination for visualization and qualitative assessment of the clustering results.

**Information Gain**

To assess information loss when removing species from co-occurrence groups, I focused on how species omission affects the ability to detect patterns across hauls. Since removing species reduces dataset dimensionality, I drew on the logic of non-metric multidimensional scaling (NMDS) and used Kruskal’s stress to quantify distortion.

I began by constructing presence–absence vectors for each haul using all species, restricting the analysis to the Eastern Bering Sea (EBS) survey region post-2010 (4,826 samples). The following procedure was then applied within each cluster.

First, I computed a Jaccard distance matrix between all hauls using only the species in the cluster of interest, treating this as the reference. I then iteratively evaluated each species by recalculating the distance matrix with that species removed, comparing the result to the full cluster matrix using Kruskal’s stress. The species whose removal resulted in the smallest increase in stress was eliminated, and the process was repeated until no species remained. At each step, stress was recalculated by comparing the current subset to the original full cluster.

This allowed us to track how stress accumulated as the “least informative” species were removed one by one. Following common guidance, I defined the “necessary subset” as the point at which stress exceeded 10%—a threshold beyond which distance patterns are typically considered meaningfully distorted.

**Testing the “Necessary Subset”**

I first combined the necessary subsets from all clusters and assessed whether information loss compounded when subsets were joined. Specifically, I calculated Kruskal’s stress between the full species set and the combined necessary subset across all clusters, to test whether joining subsets introduced additional distortion.

Next, I evaluated whether conclusions about environmental gradients were preserved when using the necessary subset instead of the full species set. I constructed an environmental feature matrix using depth, haul duration, surface temperature, and bottom temperature. Depth was log-transformed, and all features were standardized (mean-centered and scaled by standard deviation).

To assess correspondence between species composition and environmental variation, I performed Mantel tests8 on all combinations of 1, 2, 3, and 4 environmental variables, using 999 permutations per test. For each set size, I retained the combination with the highest Mantel statistic. This procedure was run separately for both the full species set and the necessary subset for comparison.

Finally, I performed a 2D NMDS ordination on a subsample of hauls for each species set. Ordinations were plotted side by side and colored by depth to visually assess any major differences in the spatial structure captured by the full versus reduced species sets.

**Results**

**Clustering and Ordination Performance:**  
The 2D NMDS ordination had a stress of 0.11, indicating a reasonable fit and visual inspection of the clustering over this ordination revealed reasonably clear separation between non-spurious groups (Figure 3). The PERMANOVA test statistic was 6.63 and statistically significant (*p* < 0.05), suggesting that the observed groupings were meaningfully different from random.

**Species Reduction and Stress Accumulation:**  
Across all clusters, the accumulation of stress as species were removed was relatively consistent. The 10% stress threshold was reached with 70–80% of species retained in most groups (Figure 4). One cluster required 90% retention, while another required only 50%, and one fell below 33%. Notably, the group requiring 50% retention was the "spurious" cluster, composed of species that could not be grouped into clusters larger than four. Overall, approximately 40% of species were removed during the reduction process. When all retained species were combined, the final stress relative to the full species set was 6.7%.

**Environmental Covariates Comparison:**  
Mantel tests showed that both the full and reduced species sets yielded the same ordering of environmental variable importance (Table 1). Correlation peaked with depth and then progressively dropped with the addition of bottom temperature, surface temperature, and haul duration (which was not expected to have a strong relationship) (Table 1). The Mantel correlations were similar across sets, differing by less than 0.03, with the largest gap occurring for depth when considered alone. In both cases, *p*-values were at their minimum given 999 permutations used indicating that the test statistics were significant (p<0.05)

**Ordination Comparison:**  
The 2D ordinations of the full and reduced species sets had stress values of 0.134 and 0.123, respectively. Visually, the ordinations were simple rotations and mirror images of one another, with the depth gradient clearly and consistently represented in both cases (Figure 5).

**Discussion**

Our goal was to assess whether a subset of species could serve as proxies for the occurrence patterns of a broader community in traditional multidimensional analyses. Based on the analyses presented here, the answer appears to be yes. Our species-reduction process removed approximately 40% of the original species, and although small differences emerged in the Mantel test correlations, they did not alter the overall interpretation: depth remains the primary environmental gradient driving species composition across hauls in the EBS which aligns with prior findings in the EBS1 and Gulf of Alaska9. Moreover, the ordination of haul subsamples showed that the reduced and full species sets produced nearly identical spatial distributions, with the same clear depth gradient present in both. In sum, the remaining 60% of species acted as an effective proxy for the full community in preserving structure and environmental signal.

One unexpected finding was that our initial guess—that clusters of co-occurring species would be more informative—was not strongly supported. While I successfully identified co-occurrence-based clusters, most required a high retention rate (≥70%) to maintain low stress. One exception was the “spurious” cluster, which required retention of just 50% of its species. An alternative hypothesis is that species diversity within the retained set may matter more than tight co-occurrence. If the spurious cluster was more compositionally diverse, its species may have better captured broader patterns.

Another interesting result was the cluster that required retention of only **33% of its species**—substantially fewer than any other group. While the scope of this study did not allow for a detailed taxonomic analysis of each cluster, this finding warrants further investigation. It is likely that specific kinds of co-variation are required for conditioning on subsets of species to matter2.

While this study demonstrates that a subset of species has potential as proxies for capturing broad patterns in species occurrence, it should be viewed as a **first step**. The retained 60% may not contain sufficient information for all types of analyses. Further work is needed to assess whether the retained species can be used effectively as predictors in species distribution models for the holdouts2.

Finally, this study focused exclusively on **occurrence**, leaving the question of **abundance** unaddressed. However, if future research on abundance patterns reveals a similar result—that a subset of species captures most of the relevant information—it could present a significant opportunity. Such findings would support the potential to **streamline monitoring efforts**, reducing both the cost and logistical burden of ecosystem-wide surveys while still preserving critical ecological insights. Beyond this it would open the door for species distributional modeling1 to be extended to a wider variety of taxa as models could be built for a smaller, well monitored, set of species and extended to a wider set2. Finally, it is worth noting that a key challenge in joint species distribution modeling is fitting many taxa with limited samples10 - reducing the cardinality of species could make such models more tractable.

**Conclusion**

This study provides an example of how a reduced subset of species can act as effective proxies for broader patterns of species occurrence in marine ecosystems. Through ordination-based analysis and stepwise species removal guided by Kruskal’s stress, I found that approximately 60% of species were sufficient to preserve key structural and environmental signals present in the EBS bottom trawl survey dataset in the years since 2010. The reduced set maintained strong alignment with known environmental gradients, particularly depth, and yielded ordinations nearly identical to those based on the complete species set.

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.*

A diagram of different colors of the surface temperature

AI-generated content may be incorrect.

*Figure 2: Spatial Distribution of Features: depth, surface and bottom temperatures by station in the 2024 bottom trawl surveys. Depth in meters, temperatures in Celsius.*

****

*Figure 3: Clustering Results. Each point corresponds to a single species. Axes are from an NMDS ordination in 2D. Colors indicate cluster membership. Pseudo-F value and p-value is from a permanova over the clusters. The cluster indicated by a -1 label is a “spurious” cluster formed by joining all clusters of size 4 or smaller in the original agglomeration.*

A graph of different colored lines

AI-generated content may be incorrect.

*Figure 4: Stress by % Species Remaining. Each line indicates the stress per cluster vs the % of species retained. The dashed line indicates 10% stress - our cutoff for this study.*

*Table 1: Results of Mantel test Search: Each row represents a search through all choices of features for a selection of 1, 2, 3, and 4 features respectively. The features are the choice with the highest mantel test statistic across the full and subset of species. Correlations and p-values are for that set of features. Note that the p-values are all at the minimum possible given the number of permutations used in the mantel tests.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | Correlation Full Set | Correlation Subset | p-value Full Set | p-value Subset |
| Depth | 0.590 | 0.566 | 0.001\* | 0.001\* |
| Depth, bottom temperature | 0.552 | 0.562 | 0.001\* | 0.001\* |
| Depth, bottom temperature, surface temperature | 0.514 | 0.528 | 0.001\* | 0.001\* |
| Depth, bottom temperature, surface temperature, duration | 0.400 | 0.405 | 0.001\* | 0.001\* |

A comparison of different colored dots

AI-generated content may be incorrect.

*Figure 5: Depth vs Ordination: Each point represents a specific haul colored by bottom depth at that haul. Axes are from two ordinations, one for the full set of species into 2D using NMDS and one using the reduced “necessary” subset of species.*

**References**

1. Rooper, C. N., Ortiz, I., Hermann, A. J., Laman, N., Cheng, W., Kearney, K., & Aydin, K. (2021). Predicted shifts of groundfish distribution in the Eastern Bering Sea under climate change, with implications for fish populations and fisheries management. *ICES Journal of Marine Science*, *78*(1), 220–234.
2. Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2020). Improving prediction of rare species’ distribution from community data. *Scientific Reports*, *10*(1), 12230. <https://doi.org/10.1038/s41598-020-69157-x>
3. Alaska Fish Research Surveys - <https://www.fisheries.noaa.gov/alaska/ecosystems/alaska-fish-research-surveys#gulf-of-alaska-bottom-trawl-survey>
4. 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>
5. 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>
6. AFSC GAP Python API - <https://afsc-gap-products.github.io/gap_products/content/foss-api-py.html>
7. Scikit-learn - <https://scikit-learn.org/stable/>
8. Scikit-bio - <https://scikit.bio/index.html>
9. Mueter, F. J. & Norcross, B. L. (2002). Spatial and temporal patterns in the demersal fish community of the shelf and upper slow regions of the Gulf of Alaska. *Fishery Bulletin, 100:559–581*
10. Warton, D. I., Blanchet, F. G., O’Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779. <https://doi.org/10.1016/j.tree.2015.09.007>