Analyzing Bird Species Distribution along the California Current Ecosystem

ANNA F HARVEY

Bellevue University

DSC 680: Applied Data Science

Prof. Catherine Williams

May 9, 2021

Abstract

The California Current Ecosystem plays an important role in the overall health of the eastern Pacific and holds many important fisheries. Every type of plant and animal within the ecosystem plays a role in maintaining its overall balance. Understanding what animals are present within the ecosystem and where they are primarily located can help guide key decisions, such as creating protected areas and tracking endangered populations. By studying specific types or species of animals within an ecosystem, it is possible to determine behavioral trends that in turn offer insight into environmental factors that may play a role in the distribution of species.

The purpose of this study is to examine the distribution of different animal groups observed during ecosystem surveys. 108 different groups were recorded at various stations. The animals recorded are primarily different bird species along with various broad categories of animals such as sharks and seals. Historical information about the locations of these animal is available on a daily basis between 2003 and 2012, providing foundational information to predict the appearance of animals at specific stations in the future. Accurate predictions of animal appearances at specific stations could allow for preemptive efforts to maintain environmental factors and resources in the areas to allow for those animals to flourish. This is particularly important to consider for salmonids and other fisheries in the area.

Introduction

Ecosystem-based management is a strategy employed by the National Oceanic and Atmospheric Administration to provide holistic awareness of ecosystem health.[1] When

addressing an issue such as the number of salmon returning to spawn, multiple factors must be analyzed. While conditions such as commercial fishing rates and predator presence may seem like obvious drivers of salmon population change, the impact of each piece of the ecosystem must be analyzed as well. The relationship between fisheries and the environment can be tenuous but there are ways to manage the impact fisheries have on an ecosystem by analyzing ecosystem variables. A broad-scale approach is necessary to fully understand the past, current, and future condition of an ecosystem.

Ecosystem-based management applies this concept by focusing on how different variables impact each other. The presence of a predator species may impact the population of their preferred source of prey. However, environmental conditions may also impact the population of a prey species which would then in turn affect the number of predators in the area. Climate effects, change in current systems, fisheries, predator-prey relationships, and competition and other factors play a role in the overall health of an ecosystem.

Problem Statement

Sea bird populations are impacted by the presence of prey species in a given area. As prey availability grows, so should the population of predatory sea birds. Additionally, if sea bird populations are particularly high in each area, that may drive down the population of a prey species. The population of sea birds demonstrates an ecosystem's ability to sustain that population. Therefore, if the presence of sea bird species in specific locations can be predicted based on prior data, those predictions can be used as a baseline to determine overall ecosystem health.

Methods

The primary dataset for this project is the "Bird Density by Station in the California Current Ecosystem" dataset from the ongoing project *Juvenile Salmon & Ocean Ecosystem Survey and Salmon Ocean Behavior and Distribution*.[2] The dataset provides information about species (or other objects) observed at oceanographic stations along with details about the segment surveyed, density of the animals observed, latitude and longitude, distance from shore, and date observed. There are 108 different groups represented by species codes. 51 of those groups are identifiable bird groups that follow the American Ornithologists' Union codes for bird species.[3] The other 57 groups are codes created by the project team to identify other animal and object groups such as sharks, seals, and boats. As I was unable to reach anyone on the project team regarding the codebooks for those miscellaneous groups, I chose to focus on the identifiable bird groups. There are 53 different stations along 9 East-West survey transects used as base points for animal population counts that lie off the coast of Washington state and Oregon (Figure 1).

Data was cleaned and visualized in Python and Power BI to understand the data distribution for species. It was discovered that two species of birds were observed at much higher rates than the other bird species. These two species are the Common Murre (code: COMU) and the Sooty Shearwater (code: SOSH). In observation of overall species appearances at different stations, these two species appeared as outliers (Figure 2). An attempt was made to visualize the species density as a map, however the high number of species led to a map that was difficult to read and did not provide significant insight.

PyCaret was used to investigate and implement appropriate machine learning models on the data. “Station name” was chosen as the target variable. The cleaned dataset was used with variables for “Transect Name”, “Station Name”, “Species Code”, “Segment Species Count”, “Segment Distance for Sums”, “Area Surveyed for Sums”, “Density”, “Log Density”, “Decimal Latitude”, “Decimal Longitude”, “Distance from Shore”, and “Date”. 29,973 instances were randomly isolated for modeling and 3,330 instances were kept aside for testing.

Multiclass classification models were compared for overall performance (Figure 3). AdaBoost Classifier (ADA) had the highest accuracy (0.0269), highest Area Under ROC Curve (AUC) at 0.5131, highest recall (0.0273), and highest Kappa (0.0073). Support Vector Machine – Linear Kernel (SVM) had the second-best overall performance and had the highest Matthews Correlation Coefficient (MCC) at 0.0163. Linear Discriminant Analysis (LDA) had the third-best overall performance and the highest F1 score at 0.0198.

Results and Conclusions

ADA, SVM, and LDA models were created, tuned, evaluated, and predicted on the hold-out sample. Almost all metrics improved overall during the prediction stage, with AdaBoost still providing the strongest overall performance (Figure 4). ADA had the highest accuracy (0.0355), recall (0.0349), Kappa (0.0170), and MCC (0.0447). LDA had the highest precision (0.0311) and F1 (0.0194). SVM’s performance decreased making it the third-best performing model.

During evaluation, it was found that the ROC curves for ADA and LDA showed that the SOSH and COMU bird species had the strongest performance when analyzing at target variable level. This is unsurprising as the number of both SOSH and COMU species greatly outnumbered those of the other species. The ROC for the ADA model gave AUC values of 0.87 for COMU and

0.85 for SOSH. The AUC average was 0.55. The ROC for the LDA model gave AUC values of 0.86 for both COMU and SOSH and had an overall average of 0.52. There were two species that were also higher than the average but lower than COMU and SOSH which were PFSH (Pink-footed Shearwater) and RHAU (Rhinoceros Auklet), which were also the next highest overall populations.

On the classification report for the SVM model, COMU had a precision of 0.594, recall of 0.096, and F1 of 0.160. SOSH had a precision of 0.5, recall of 0.206, and F1 of 0.294. SHSP (Stripe-headed Sparrow) resulted in a recall result of 1, which seems to be an error as there is no other data to support the result. PFSH and RHAU did not have significant results from the SVM classification report.

Although model performance remained consistent, the ability to consistently predict the species occurrences at the stations was not significant. A lack of predictability could result from multiple variables including consistent distribution of species along the overall area. If bird species cannot be accurately predicted from occurring at specific stations, that may indicate that species presence may not affect the ecosystem or be affected by the ecosystem as much as other factors, such as the quantity of individuals in a location.

Future Investigations

Due to the high occurrence of Common Murre and Sooty Shearwater, it may be valuable to focus predictive analytics on those two species alone. This is especially true if those species are found to have direct effects on ecosystem health or if the ecosystem's health has direct effects on their populations. Additional ecosystem variables should be added to observations if effective predictive analytics can be expected for analyzing ecosystem health based on species

occurrences. Helpful variables to include could be weather variables (temperature, wind conditions, rain), ocean chemistry conditions (oxygen levels, presence of pollutants), plankton measurements (since both phytoplankton and zooplankton density are significant markers of ecosystem conditions), and levels of human interference (amount of fishing nearby or other vessels). Some of these variables are included in other datasets as part of the *Juvenile Salmon & Ocean Ecosystem Survey and Salmon Ocean Behavior and Distribution* project. Although other data analytics methods can also be helpful in observing and assessing ecosystem health, predictive analytics can help play a role in facilitating caretaking procedures now to help protect ecosystems in the future.

APPENDIX

Figure 1. Map showing survey transects and survey stations.

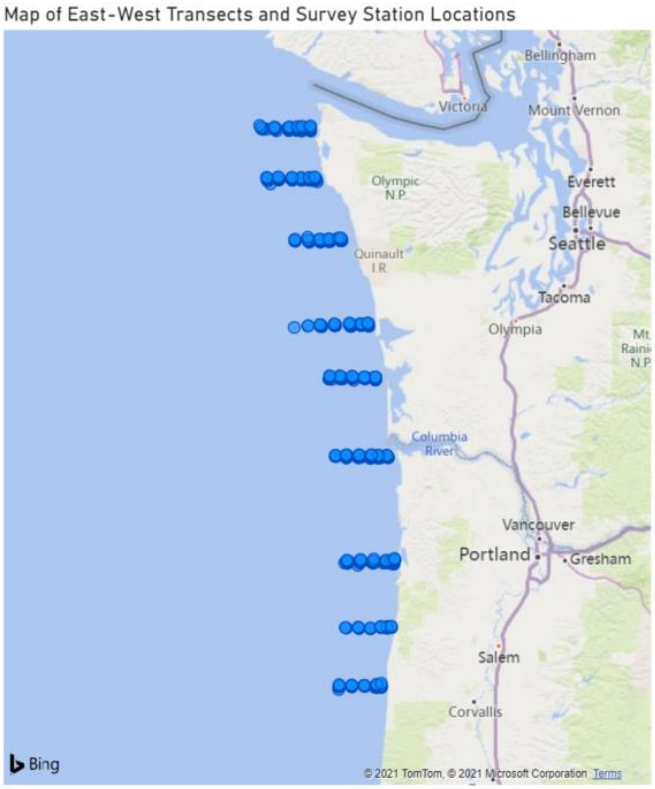


Figure 2: Scatter plot for the top 10 bird species represented along survey routes.

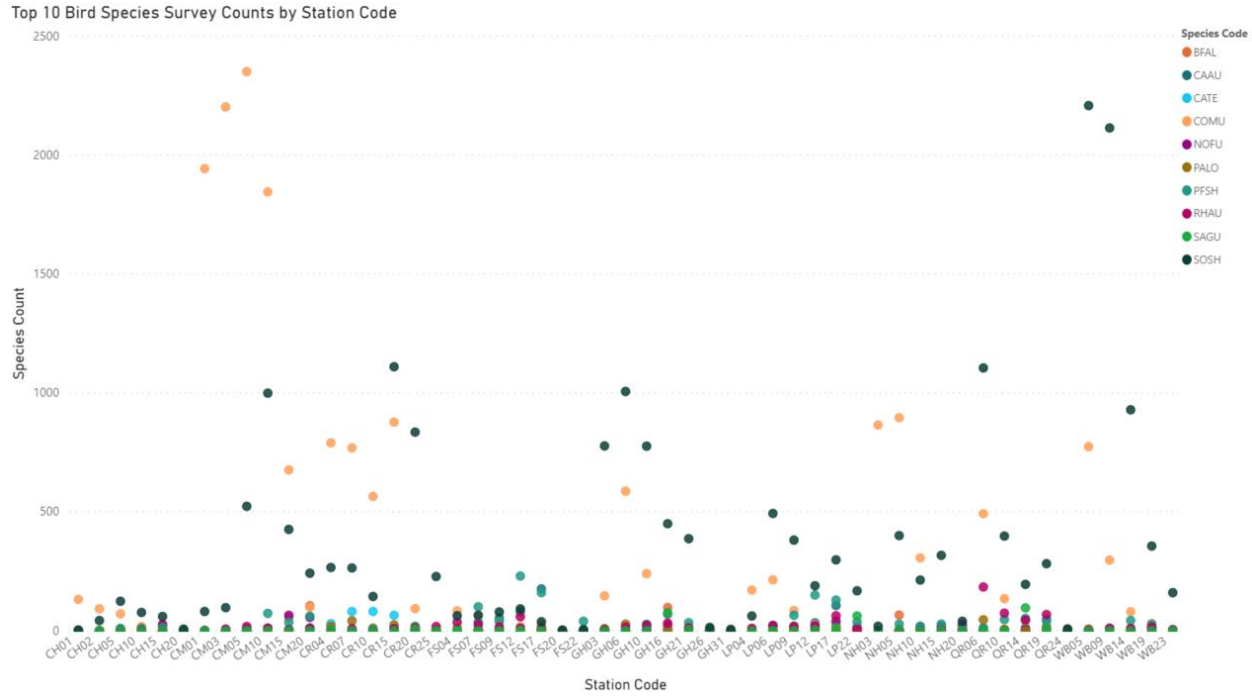


Figure 3. PyCaret evaluation of Multiclass Classification models for the dataset.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.0269	0.5131	0.0273	0.0132	0.0129	0.0073	0.0076	1.1730
svm	SVM - Linear Kernel	0.0264	0.0000	0.0267	0.0116	0.0094	0.0070	0.0163	18.8450
lda	Linear Discriminant Analysis	0.0204	0.4418	0.0208	0.0214	0.0198	0.0007	0.0007	0.2380
ridge	Ridge Classifier	0.0202	0.0000	0.0207	0.0129	0.0143	0.0006	0.0006	0.0390
nb	Naive Bayes	0.0190	0.4676	0.0192	0.0154	0.0107	-0.0008	-0.0008	0.2030
qda	Quadratic Discriminant Analysis	0.0190	0.5002	0.0200	0.0007	0.0014	0.0004	0.0006	0.1690
lr	Logistic Regression	0.0182	0.4621	0.0185	0.0209	0.0179	-0.0016	-0.0016	32.5360
gbc	Gradient Boosting Classifier	0.0125	0.3618	0.0130	0.0208	0.0142	-0.0073	-0.0073	87.6530
knn	K Neighbors Classifier	0.0111	0.4645	0.0116	0.0152	0.0112	-0.0089	-0.0090	0.7790
rf	Random Forest Classifier	0.0109	0.2705	0.0114	0.0101	0.0104	-0.0089	-0.0089	1.2570
lightgbm	Light Gradient Boosting Machine	0.0109	0.2969	0.0114	0.0138	0.0119	-0.0089	-0.0089	6.2960
dt	Decision Tree Classifier	0.0108	0.2596	0.0112	0.0294	0.0135	-0.0093	-0.0113	0.0700
et	Extra Trees Classifier	0.0102	0.2668	0.0107	0.0255	0.0125	-0.0099	-0.0120	1.5740

Figure 4. PyCaret model final results.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Ada Boost Classifier	0.0355	0.5505	0.0349	0.0118	0.0094	0.0170	0.0447

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
SVM - Linear Kernel	0.0257	0	0.0258	0.0247	0.0111	0.0070	0.0309

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Linear Discriminant Analysis	0.0299	0.5210	0.0294	0.0311	0.0194	0.0107	0.0124

RESOURCES

1. NOAA, F. (2020). California Current Regional Ecosystem. https://www.fisheries.noaa.gov/west-coast/ecosystems/california-current-regional-ecosystem.

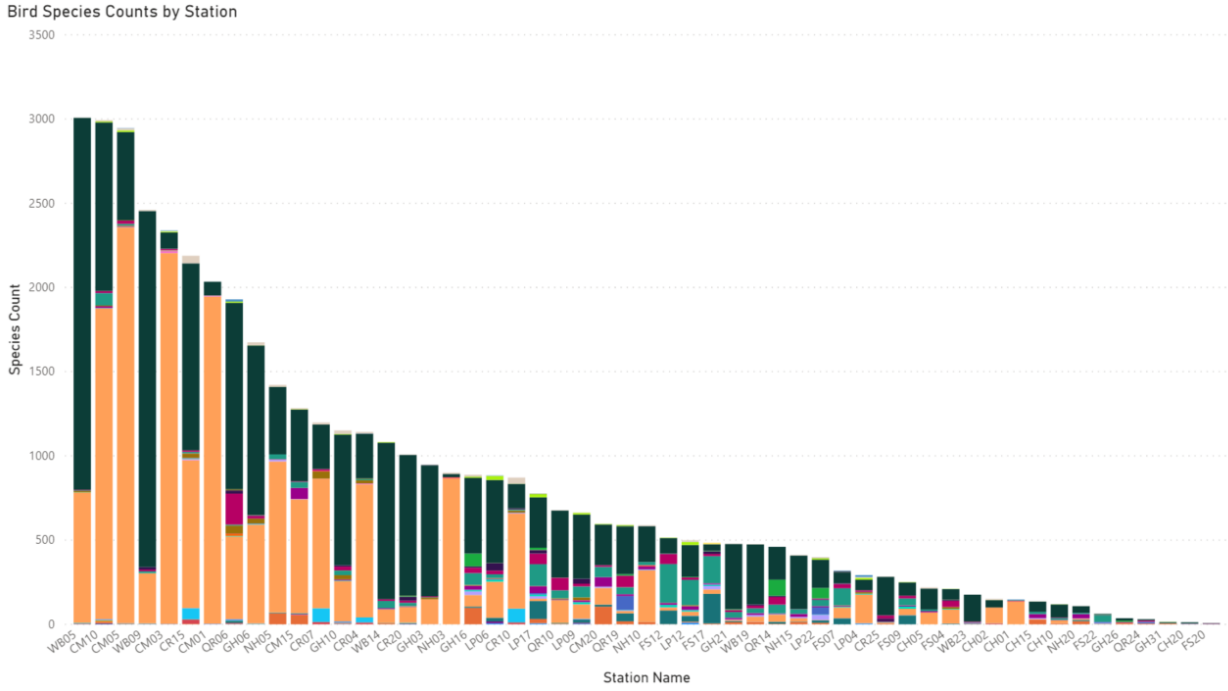
2. NOAA. (n.d.). Bird Density by Station in the California Current Ecosystem. NWFSC PARR Data - Bird Distribution and Abundance - Bird Density By Station in the California Current Ecosystem. https://www.webapps.nwfsc.noaa.gov/apex/parrdata/inventory/tables/table/bird_density_by_station_in_the_california_current_ecosystem.

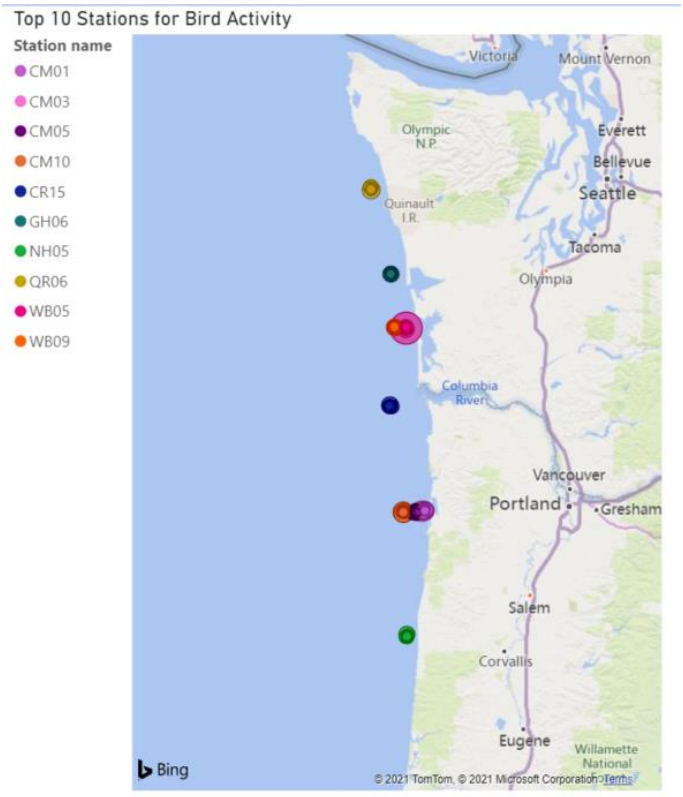
3. IBP. (2020, September 16). Standardized 4- and 6-letter Bird Species ("Alpha") Codes. https://www.birdpop.org/pages/birdSpeciesCodes.php.

https://github.com/anhar421/Portfolio

10 Questions

1. *Which stations showed the most bird activity?*
   WB05, CM10, CM05, WB09, CM03, CR15, CM01, QR06, GH06, AND NH05 were the top ten stations with the most bird activity, in that order.





The top stations were evenly distributed along the entire range, although they only represent 6 of the 9 transects.

2. *What would it mean for a species to have high accuracy for a predicted appearance at a station?*
   If we were able to accurately predict a specific species occurring at a specific station, it would show that the station is a regular habitat for that species, either because it ranges into the area or remains in the area regularly.

3. *What would it mean for a species to have low accuracy for a predicted appearance at a station?*
   Low accuracy could mean a variety of things. It could be that there are too many variables involved in an ecosystem to accurately predict certain species' behavior. In the case of this dataset where there is a specific range of locations to predict species appearances, the locations are close enough together that the same individuals could be traveling between the locations regularly.

4. *What additional variables could increase the ability to predict species appearances at a station?*
   Including information about prey species of the target species could help determine that species' appearance in a location. A thorough understanding of the life histories of those target species would be important too. Certain weather conditions, times of year, or oceanic conditions may need to be present for those species to remain or arrive at a station.

5. *Do Common Murre (COMU) and/or Sooty Shearwater (SOSH) play significant roles in the local ecosystem?*
   Common Murre are only found in North America while Sooty Shearwater are found throughout the world. They both dive into the water to find food, although the murre will swim more. They have similar diets, primarily eating fish but also squid and marine invertebrates. Shearwater populations have declined in North America in recent years, presumably due to warming waters. As fish-eating birds, they would play an important role in indicating the presence of certain prey species in different areas. They likely compete with other sea birds and marine mammals for food.
   https://www.audubon.org/field-guide/bird/common-murre
   https://www.audubon.org/field-guide/bird/sooty-shearwater

6. *What factors might impact the ability for accurate data gathering when doing ecosystem surveys?*

One important factor to consider is the impact of human presence on surveys. It is difficult to observe animals without affecting their behavior. With birds, they could hear vessels approaching and leave the area before they are ever seen. It is important to conduct animal surveys in the least invasive way possible for accurate analysis of animal populations.

7. *What strategies could be employed to collect more accurate data during ecosystem surveys?*
   One way to mitigate human influence on the presence of animals would be to install video installations that could record the area over a given time. Artificial intelligence or manual observation could then be used to identify the animals and number of animals in the video. This would also be beneficial since the video could be relayed anywhere, with a sufficient internet connection and the technology available. It could also simply be retrieved on a regular basis.

8. *What ecosystem variables that impact bird species appearances are controllable by ecosystem managers?*
   The presence of fisheries and ocean vessels can impact bird density in natural environments. Fisheries may attract birds to an area, as fisheries provide an easy meal. However, overfishing would impact the long-term population. If the prey species are overfished, the birds will have to leave to other locations. Ocean vessel traffic can cause birds to leave as well by either interfering with their ability to hunt or limiting availability of prey species as fish are driven away from the noise.

9. *What ecosystem variables that impact bird species appearances are not controllable by ecosystem managers?*
   Large ecosystem variables such as weather, air temperature, water temperature, and water chemistry are generally not controllable by ecosystem managers. Managers can advocate for best practices to limit human impact on climate and ocean ecosystem health, but on a day-to-day or even year-to-year basis, ecosystem managers have to simply observe the affects and adapt to uncontrollable variables such as these.

10. *How would it help ecosystem managers to be able to accurately predict the appearance of specific species in specific locations?*
    We see this scenario play out for many migratory species. If we know that a specific species or specific population will be in a given area at a given time, we can make sure that the ecosystem is healthy enough to sustain them (to the best of our limited ability). For species that exist in the same locations year-round, the absence of a species in a location that they were expected to be in can be an indicator that something has happened to the ecosystem to make their presence unsustainable.