Analysing the Fish Hunting Numbers in North Pacific Ocean*

Using Bayesian Modeling, to find that number of fish caught has increased

Shreya Sakura Noskor

November 30, 2024

Abstract

We analyzed data on fish catches in the North Pacific Ocean using Bayesian modeling. Our analysis shows that the number of fish caught has increased over time. This suggests that fishing activities in the region have intensified. Understanding this trend is important for managing fish populations and ensuring the sustainability of the ocean's resources. This is important as the Pacific is full of different kind of fish like, trout and salmon that are essential in the food chain, but are being overhunted for food and recreational purposes.

Table of contents

*Code and data are available at: https://github.com/NotSakura/FisheriesData.git.

1 Introduction

Estimand is the number of fishes that were caught by each country, for every year. Or more specifically what the rate was and if they are more likely to be fished for commercial purposes.

2 Data

2.1 Overview

The data was downloaded from (NPAFC) (2024) and was cleaned using R (R Core Team 2023). The data was read using Schauberger and Walker (2024) and Wickham and Bryan (2023), while the data was cleaned using Wickham et al. (2019), Wickham et al. (2019), Firke (2023), Wickham et al. (2023), Xie (2023). The data was modeled using Arel-Bundock (2022), Robinson, Hayes, and Couch (2023), Goodrich et al. (2022), and Bürkner (2017).

To download the data go to NPFAC's official data portal and look for "NPAFC Catch Statistics (updated 28 June 2024)". Click on that link to get the csv file containing all the data.

2.2 Methodology and Measurement

NPAFC has this data to download from their website. The way they gathered this data was that they are an inter-government organisation so they have access to government data based on how much fish were hunted in the respective countries. The countries in the data include Canada, Russia, Korea, Japan and Unites States of America. The way each of these countries measured this data was that when fish are being caught on international waters and report it to each other. This is strictly enforced, especially after the fall of salmon and trout population in the Pacific, majorly due to environment purposes.

This paper look at multiple variables. We will go through them one by one: - First, variable we look at is Country which are either "Canada", "Russia", "Korea", "Japan", and "United States", all representing the countries that are members of this organization. This data was left unchanged. - Next variables we look at is "Whole Country/Province/State", where the instances of these variables may either be Whole country, or the different states or provinces that was fishing and gathered that data. So for example, if the value was British Columbia then the corresponding number of fishes caught reported is the number of fish caught by the province. Throughout our data we filtered for the "Whole country" value, assuming that the numbers in each province and/or state would add up to the number in "Whole country" (which it did). This was because we were more interested in comparing the fishing trends between countries rather then within a country. - The next variable is "Reporting Area" which accounts for where the fishes were caught. This was also filtered by "Whole country" due to the previous reasoning. The next variable that we filtered was "Species" which contained "Cherry", "Chinook", "Chum", "Coho", "Pink", "Sockeye", "Steelhead" and "Total". These are all types of salmon except for Steelhead which is a trout and "Total" which represents all the fishes that were hunted. Although there is a lot of interesting information to uncover if we did a deeper analysis on each fish, but, we decided that the best way to compare the fishing trends between countries would be to just look at the total fishes caught. - The next variable we used and actually analyse is the "catch type". This tells us whether is the fishes were caught for commercial purposes (caught for profit purposes like selling), sporting purposes (which means they were caught recreationaly) or subsistence purposes (which means they were caught to provide food, not as a profit). - The last variable we filtered was "Data Type" which was the unit that these numbers were reported in. There was Numbers in 1000s or Round weight in metric tonne. We chose to filter with numbers in 1000s as the other option was done only by the US, who provided both units.

These were the variables that we filtered but the variable we actually analyse is the number of fishes caught. That number in the raw data was provided as the year as a column and the corresponding value as the number of fishes caught. This format meant that when we are creating models or graphs it is very difficult to work with it. And so we actually shifted, rather pivotted the table so that analysis is easier. To pivot the table we created 2 new rows to the dataset, "Year" and "Catch". The year corresponds to the column's

title which is the year this data is for and the "catch" refers to the number of fishes that were caught in that year, in that country. This helped significantly with making the models and such.

2.3 Data Visualization

we make the assumption that the columns that say whole country it also include the provinces and different areas number as well.

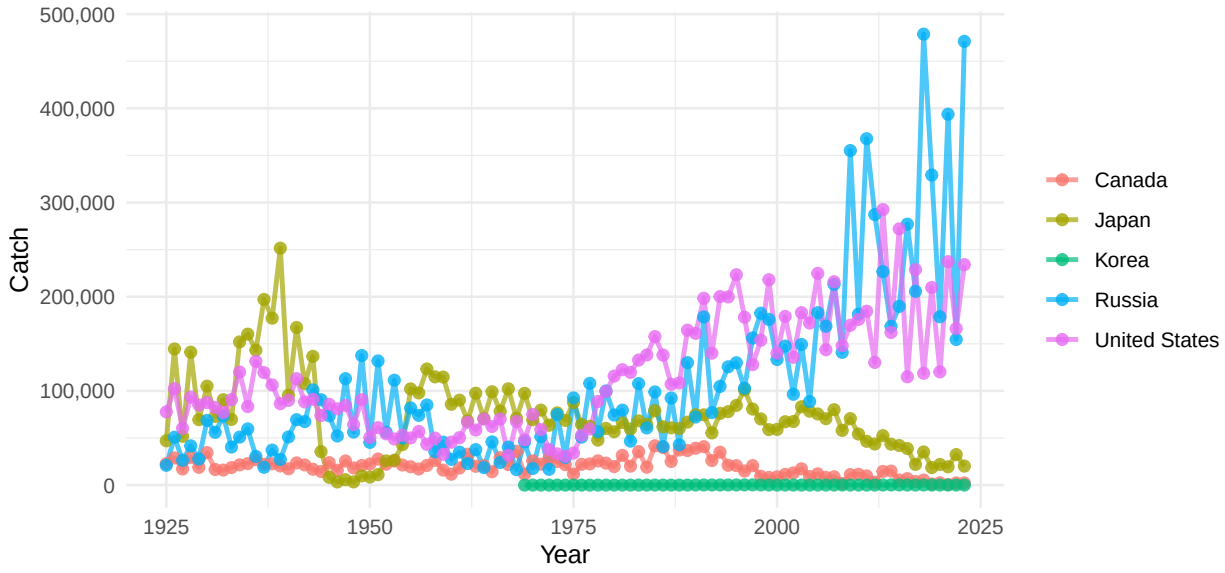


Figure 1: Catch over Time by Country

From Figure 1 we are able to see the number of fishes caught in each country as time goes by. They fluctuate a lot because over fishing in one year means that the next year there is not a lot of fish left as not as many survived to repopulate. However we see that there is a increase in the number of fishes caught for the US and Russia. Meanwhile, Japan and Canada are declining. This is interesting as Japan is surrounded by the Pacific Ocean more than all of these other countries and Canada coming in a close second. An educated guess as to why this may be the case is that, the US and Russia may be fishing more as they have a stronger fishing industry with the technology and money to fund longer trips in the ocean.

In Figure 1 we can barely see the fluctuation for Korea; it seems almost constant. Hense we graph it sepreatly here in Figure 2. We see that like the other countries in our data, they also fluctuate in numbers. From this an guess may be that they are decreasing the number of fishes they catch but, our guess is quite the opposite. We think that in 2023 it was just low due to one of the fluctuations where they fished too much the previous year. We expect a rise in the next 2 or 3 years.

Figure 3 is a great way to analyse the different reasons why these fishes are caught. We narrowed the year to 2022 as we predict the fishing industry has recovered after covid and will show accurate results. Here we see that most fishes are caught due to commercial purposes where they sell the fish rather then for sporting or subsistence purposes. It also seems that Russia fishes the most out of the Norther Pacific region with its number being almost doubled from the second most country to fish, the US. Russia has fished around 309,142,000 fishes in the year 2022 next to US who has fished around 166,114,000. With these numbers, no wonder we see fluctuation in the number of fishes caught in Figure 1.

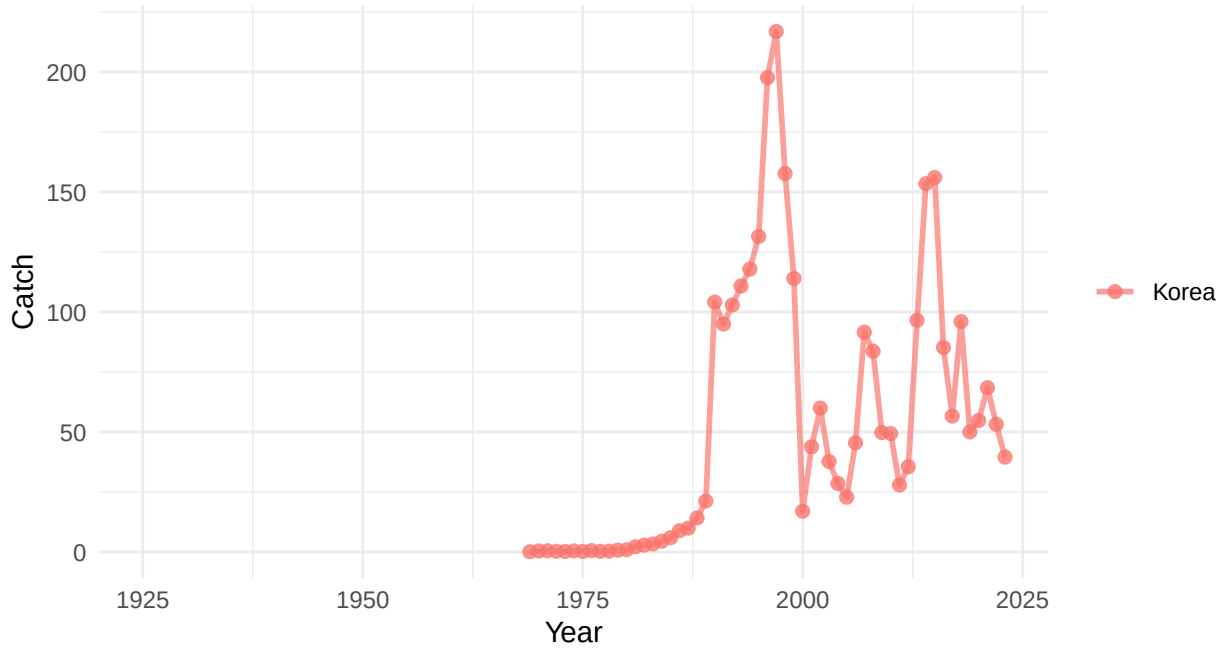


Figure 2: Catch over Time by Country

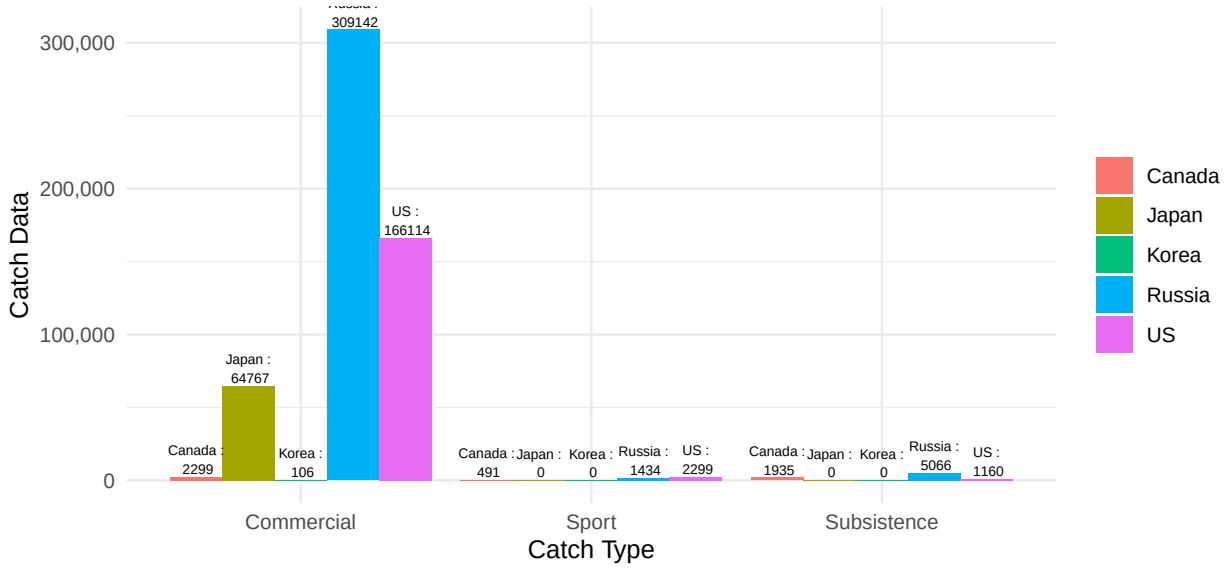


Figure 3: Catch Data by Country and Catch Type (2022)

3 Model

Here we model the data in 2 ways. The first way is to model the rate at which the fishes are being caught for any country. And the second model looks at what is the probability that a country is fishing for commercial purposes.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. We also use `brm` package from Bürkner (2017)

3.1 Number of Fishes Caught (dependent on Year and Country)

Define $Catch_{ij}$ to be the number of fishes caught. We are trying to see how will $Catch_{ij}$ change as we increase $Year$ which is our other variable. We are finding the correlation with the countries in mind which is why we account for the random effect of the country.

$$\begin{aligned}
\text{Catch}_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma^2) && (1)\\
\mu_{ij} &= \beta_0 + \beta_1 \times \text{Year}_{ij} + u_j && (2)\\
u_j &\sim \text{Normal}(0, \sigma^2_{\text{Country}}) && (3)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (5)\\
& && (6)
\end{aligned}$$

3.1.1 Model Justification

- $Catch_{ij}$ is the number of fishes caught, the variable we are modeling
- $Year$ is the independent variable that shows the year
- u_j is the random effect of the countries. This is important as the number of fishes caught differentiate between country so to get a weighted result we add this variable.
- β_0 and β_1 tells us the intercept and the slope respectively, of the Bayesian Generalised Linear Mixed Model.

We predict that even thought 2 (may be 3) out of the 5 countries in the data set have a positive trend in terms of the number of fish being caught (Figure 1), because the rate at which Russia and USA is increasing is way higher than Japan, Korea and Canada. This can be seen visually. So I think that the rate of increase is going to be positive.

3.2 To be Commercial or Not to be Commercial

Define $IsCommercial_{ij}$ to be the probability that the fishes caught were commercial (for buisness purposes). We are trying to see how will $IsCommercial_{ij}$ change as we change $Country$. We esstially use, again, a Bayesian Generalised Linear Mixture Model to see how does the likelihood of fishes being caught for commercial purposes varies from country to country.

$$\begin{aligned}
\text{IsCommercial}_{ij} &\sim \text{Bernoulli}(p_{ij}) && (7)\\
\text{logit}(p_{ij}) &= \beta_0 + u_j && (8)\\
\text{u}_j &\sim \text{Normal}(0, \sigma^2_{\text{Country}}) && (9)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (10)\\
\sigma_{\text{Country}} &\sim \text{Normal}(0, 2.5) && (11)
\end{aligned}$$

3.2.1 Model Justification

- $IsCommercial_{ij}$ is the probability that the fishes caught were commercial. This is what we are modelling using Bernoulli Distribution.
- *Country* is the independent variable that shows the countries.
- u_j is the random effect of the countries. This is important as the number of fishes caught commercially is different between countries.
- β_0 tells us the intercept so, what is the likelihood of it being commercial for a "baseline" country. This is essentially the average of all the intercepts for all the countries.
  Note that this model doesn't look at the number of fishes caught rather the number of occurrences in the year 2022 where fishes were caught commercially. We focus on the year 2022 because it is recent data and modeling with all years between 1925 to 2023 would have not given us a result that was relevant to us. We would also loose information as the liklihood may have not been as high in the 1900s due to the lack of equipment.

We predict that the outcome or rather the probability that the fishes caught were commercially would be positive. This is because as we were looking at the graphs in the data section, we notice that Figure 3 shows most of the fishes caught were commercially and then as a sport and then as a subsistence.

4 Results

4.1 Model Results

Table 1: Model Summaries

(a) bayesian model summary for predicting the probability of country fishing for commercial purposes.

	Second model
b_Intercept	0.84
	(1.03)
sd_Country___Intercept	2.32
	(1.07)
Num.Obs.	45
R2	0.267
R2 Marg.	0.000
ICC	0.7
ELPD	−26.1
ELPD s.e.	2.6
LOOIC	52.2
LOOIC s.e.	5.2
WAIC	51.6
RMSE	0.41

(b) bayesian model summary for predicting the rate of fishes being caught

	First model
(Intercept)	−1 367 335.76
	$(1.902\,962 \times 10^5)$
Year	721.74
	$(9.452\,000 \times 10^1)$
Sigma[Country × (Intercept),(Intercept)]	3 487 096 712.07
	$(2.258\,065 \times 10^9)$
Num.Obs.	451
R2	0.448
R2 Adj.	0.439
R2 Marg.	0.077
Log.Lik.	−5547.420
ELPD	−5555.7
ELPD s.e.	36.7
LOOIC	11 111.3
LOOIC s.e.	73.4
WAIC	11 111.1
RMSE	53 056.21
r2.adjusted.marginal	0.438511212758486

In Table 1a we see the model summary for Model 1 which was finding the probability of a country fishing for commercial purposes. And in Table 1b we see the results for model 2 where we look at the rate at which fish are being hunted for the "baseline" country. There are a lot of numbers here that we don't need to look at to see if our predictions were right. We only need to know the intercepts and the slopes or as shown in the model β_0 and β_1. In Table 1a we see that the log likelihood of a "baseline" country fishing for commercial purpose is 0.84 which is 69.8%. This means that there is 69.8% probability that the country is fishing for commercial purposes, without even knowing the country. The "sd country Intercept" just tells you how much variation there is in that result. Next in Table 1b we see that the intercept is -1,367,335 which tells you that at Year 0 somehow a baseline country fished negative -1,367,335 fish. We will discuss more about this number and interpret it in the next section (Section 5). The "Year = 721.74" tells you that with every increase in year there is about 721,000 more fish that are being caught. And finally similar to the other model summary "sd country × Intercept Intercept" tells you how much variation there is in the results. We don't need these variation values unless we are analysing the model (and not the data).

4.2 Other statistics

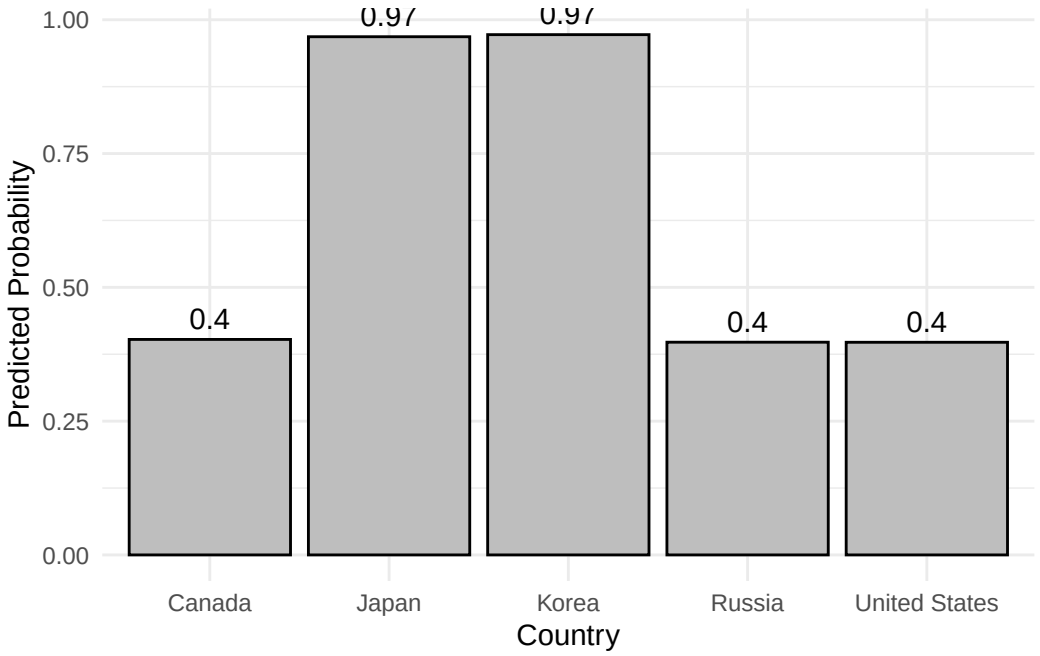


Figure 4: Predicted Probability of Commercial Catch by Country

We also calculate the predicted probability of a country fishing commercially for all 5 countries in the NPFAC. In Figure 4 we see that Canada, US and Russia has a 40% probability that they fish commercially, Japan and Korea both have 97% probability. We will see why that is the case in the next section (Section 5).

5 Discussion

5.1 Story of the Data

5.2 Weaknesses and next steps

Korea didn't report data until 1969 so they are left out of the first model in general.

A Fisheries and Surveys/Observational Data

The data that we gathered for this paper is definitely observational data. This is because the in the NPFAC's official website they state that they gather the data by contacting each member of government and asking them to submit this data. In our own Canadian government, we see that we have Fish Monitoring Policy where people are expected to report their catch data if they caught a lot (Fisheries and Oceans Canada 2021). For example, if some one has fished a rare specie of fish or a large number of fishes were caught they must report it to the government. This is how the government gains the data and passes it on to NPFAC.

The common concern with datasets of our type is often 2 paradoxes know as Simpson's Paradox and Berkson's Paradox. We assume that Simpson's paradox does come into play for our model in Section 3.1. Simpson's Paradox is when we have relationship between subsets of data show a result but when we model the entire result together, the result differs. This can be seen in our first model because we see that in Figure 1 that different countries have a different rate of change for the number of catches with respect to year. We see that Japan and Canada has a decreasing rate. However the over all model outputs a positive rate of change. This shows that the subset of the data has a different result then when we combine the entire data. So this paradox does apply to us. Next we also say that the Berkson's Paradox also affects us as in our data set we only take a look at the data for Salmon and Trout Population. However, what the dataset doesn't account is the other fish species that are also fished. Yes it is true that over fishing is an issue but the result that our model shows may not be the most accurate number due to the fact that this key information is left out. The Mountain House (2024) shows that salmon, tuna and trout are one of the most fished anadromous. And North Pacific Anadromous Fish Commission (2024) shows the other anadromous fishes that are in the North Pacific Ocean.

We observed that one of the reasons why these paradox applies to us is because our data doesn't sample the entire data. The way they gather their data is similar to an Integrated method (mentioned in Stantcheva's supplemental appendix in Stantcheva (2023)). Although they are not an open community where anyone who is a member of the community may be rewarded for their participation in giving their data, all of the countries is under an organisation and they give their data as they all actively participate in the organisations goal for sustainable fishing. The issue we were talking about earlier- the one that causes these paradoxes to affect our data - is the fact that the way this organisation collects data also presents a sampling bias. A sampling bias is when the data collected is not a representitive of the entire population. This can be caused when data is collected only from one place and we ignore all other factors. This is what happens here. Because NPFAC only consists of the 5 countries in the data (Japan, Korea, Russia, US, Canada), our data has gaps for the number of fishes caught by other countries that fish in the North Pacific Ocean such as other East Asian countries like China, Phillipines and Taiwan. Realistically China has a large fishing industry as well so our analysis would look very different when we include their data. Another thing we are missing is the fact that there are other Anadromous fishes in the Pacific that could also skew the data and causes the Berkson's Paradox.

We will now move on from critiquing the data to our analysis. Our modeling is a prime example of convenience sampling. This is when we filter data to our own rules and analyse those results because it is the most convenient for us. In this paper it was done for Figure 1. This was due to the fact that not a lot of the data dated back to the 1925 so, initially Korea was left out from our data. However, because we wanted a chance to compare all 5 countries of the organisation, we combined data from Korea which started 1969. We purposefully chose filters that will shorten our data while getting getting the attributes we want to study. The effect of this is that our model where we check the rate of change of fishes caught Section 3.1, is inaccurate as we don't have the data from 1925 to 1969 for Korea and we filtered the data so much to fit Korea that we loose information like the type of fish and why they were fished (commercial, sporty, subsistence).

Now we will move to the survey analysis. Because this data was not gathered through a survey we will simply talk about how we would gather this data. It would not be very complicated. Our hope is that 6-months before we send the survey we ask the fisher to keep track of their weekly fish caught. That they would record the number of fish caught of each type, the area they were in and when they caught them (time of day would be appreciated but the date is minimum). Then 6 months after the notice I send out a survey that asks for their information. It would be nothing private like the area they live in and how long

they spent fishing. I would send out this email to corporations for their data as well as regular fisherman that fish for subsistence. This would be to make sure there is not sampling bias when collecting the data. This was an issue in our original data as we noticed there were a lot more commercial fishes caught and we did not know if this was due to not a lot of fish caught for subsistence or they were not reached out to for their data. This would make sure that we have as minimal gaps in our data as possible.

B Data Sheet

1. For what purpose was the dataset created? The data set was originally created by North Pacific Anadromous Fish Commission ((NPAFC) 2024) created this dataset to see how much fishing of trout and salmon was done in international waters. They strictly prohibits mass fishing of these highly demanded fishes and hence the dataset. We use the dataset to analyse the number of these fishes caught and see if there is a reason for NPAFC to be worried, and if there is a solution to this.

2. What do the instances that comprise the data set represent (for example, documents, photos, people, countries)? The instances of these data is either numeric or categorical. Thee first couple of columns tells you the country the data is from, where the fishes were hunted as well as getting to specific regions. Why they were hunted and what the unit were, was also in the data set. The majority of the dataset is numbers expressing how much fish was hunted, either in thousands or tonnes.

3. Is any information missing from individual instances? There are several instances of data missing. Most are from the number of fishes column as some of the data goes back to 1925 but, not all of them so, there are some rows of data where the total number of fishes collected in 1928, for example, is not present.

4. How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/ derived from other data (for example, part-of-speech tags, model-based guesses for age or language)?

This dataset was reported by the subjects of each country. Meaning this organisation asked the government of Canada, US, Korea, Russia and Japan, send in the numbers and they compiled this data.

C Model Card - North Pacific Anadromous Fish Data

Model Details - Model was created by Student at University of Toronto on November 27th 2024. - Bayesian Generalized Linear Model (GLMM) - Ver 4 as the other versions tried to take all variables like region and country into consideration but the model was too complex to implement and even more to analyse.

Intended Use - Analyse the number of Anadromous fish caught in the North Pacific Ocean by looking at if they are commercially caught and the rate at which they were caught. - Intended to see if sustainability measures put on fish hunting has worked - Not intended to see what measures can be put into change the outcome of the results seen here (such as laws to prevent the rate being so high)

Factors - Relevant factors are year, number of fishes caught and if the fishes were caught for commercial purpose or not (true or false value). Year is a predetermined value and the countries that reported the data to NPFAC was in charge of identifying the number of fishes they caught and why they caught it. Details available in data section of the paper. - Evaluation factors or the values that are being reported are rate of number of fish caught Figure 5b and probability that the fish caught was commercial Figure 5c. These are being evaluated to understand if certain sustainability measures are helping the fish being over hunted.

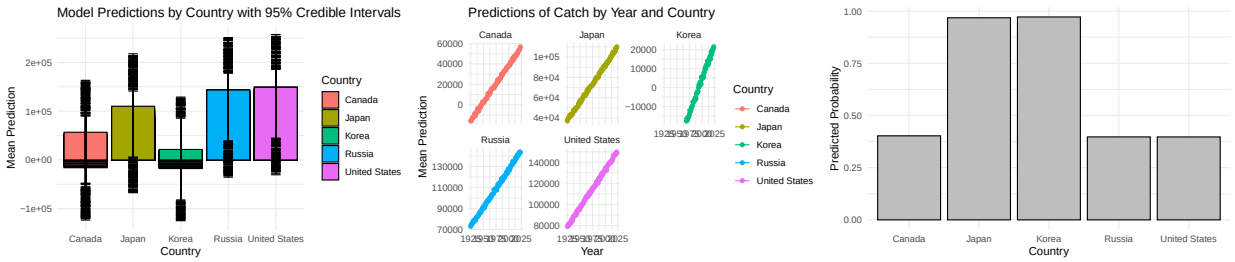
Metrics - Measured in Thousands of Fish as given in (NPAFC) (2024).

Training Data - Data from (NPAFC) (2024) was used to make/train this model.

Evaluation Data - We applied the model on the same data as the training set Figure 5b and we see that it is linear due to the fact that we applied the model on it self with the same rate.

Ethical Considerations - We are only using the observational data. Hence we are not actually changing the data themselves. The model does not take into account any personal data as the only data provided was the year, the number of fish caught and where they were caught.

Caveats and Recommendation - Use a data that has no sampling bias. This is because of the reasons discussed in Section A. To summaries it, it is because of the Simpson's and Berkson's Paradox that seems to arise.



(a) Bar plot with error bars for predictions and credible intervals by Country

(b) Plot predictions over time (by Year) for each Country

(c) Predicted Probability of Commercial Catch by Country

Figure 5: Quantitative Analysis

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Bürkner, Paul-Christian. 2017. "Brms: An r Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Fisheries and Oceans Canada. 2021. "Monitoring, Control and Surveillance." 2021. https://www.dfo-mpo.gc.ca/reports-rapports/regs/sff-cpd/fishery-monitoring-surveillance-des-peches-eng.htm.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Mountain House. 2024. "Most Popular Game Fish in North America." 2024. https://mountainhouse.com/blogs/hunting-angling/most-popular-game-fish-in-north-america.

North Pacific Anadromous Fish Commission. 2024. "North Pacific Anadromous Fish Species." https://www.npafc.org/species/#:~:text=Common%20anadromous%20fish%20include%20salmon,cherry%20salmon%20and%20steelhead%20trout.

(NPAFC), North Pacific Anadromous Fish Commission. 2024. "Statistics – North Pacific Anadromous Fish Commission." https://www.npafc.org/statistics/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Schauberger, Philipp, and Alexander Walker. 2024. *Openxlsx: Read, Write and Edit Xlsx Files*. https://CRAN.R-project.org/package=openxlsx.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." Journal Article. *Annual Review of Economics* 15 (Volume 15, 2023): 205–34. https://doi.org/https://doi.org/10.1146/annurev-economics-091622-010157.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.