How do Oil Spills Affect Various Whale Populations Around the World?

CORE-200: Scientific Methods

Term Project Part 3a - Data Analysis and Findings

Spring 2024

Syed Mujtaba Hassan
Syeda Manahil Wasti

Methods

To test our hypothesis, we gathered datasets of global whale populations and amount of oil spilled in tonnes over the years globally. We found various datasets related to this. For whale population, our primary data set is from International Whaling Commission[1]. This dataset includes estimated population of various types of whales over the years in different regions of the world. Our other data set was a dataset of oil spills over the years that we found from Kaggle[2].

Now the first issue with our data set of oil was that a major amount of oil spills go unaccounted for. Often, oil corporations spill a large amount of oil but have no accountability for how much oil was spilled. As Deng Yuewen and Linda Adzigbli pointed out[3], that 80% of the oil spilled into the ocean goes unnoticed and unreported. Due to this, we had to discard oil spills for which the amount of oil spilled in tonnes was not known. Then, we added together the known amount of oil spilled for each year to obtain a data set of total oil spilled each year globally. We wrote a MATLAB code which cleaned our data i.e., removed spills where amount was unknown, summed up the total spill for each year, and plotted this data. Figure 1 shows a bar graph of total amount of oil spilled each year.

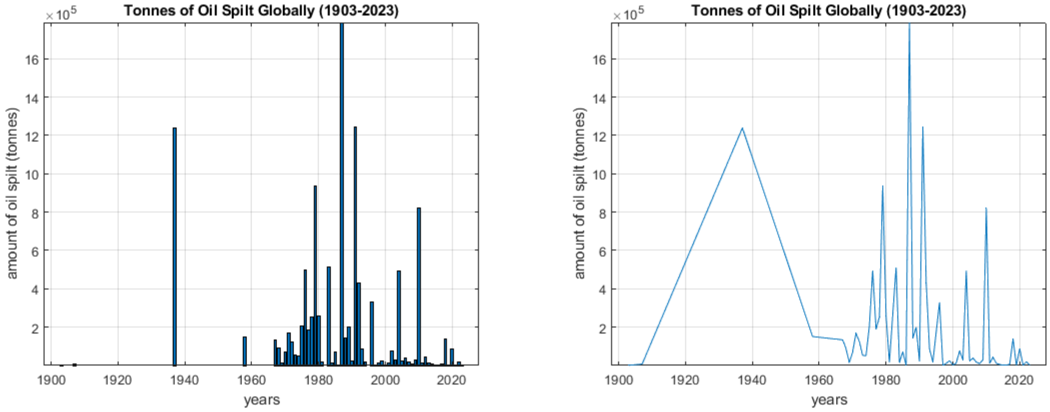


Figure 1: Tonnes of oil spilt globally per year

We had to compare the total known oil spilled per year globally with the global whale population for each species of whales per year and the data set we found for whales consisted of data of thirteen different types of whales. We chose six that we had a good amount of data for, namely: Blue whale, Fin whale, Gray whale, Humpback whale, Mink whale, and Right whale. The major issue with our data set of whale population was that for a lot of areas whale population was unknown for many years while we have whale population known for that year for some other location.So we extrapolate our known data points to extract estimated whale population of unknown years.

Let $Y \subseteq \mathbb{N}$ be the set of years where $Y = \{x \in \mathbb{N} | 1903 \leq x \leq 2023\}$. Let $A_s = \{a_{s1}, a_{s2}, \dots a_{sn}\}$ be the set of regions for which we have some known whale population for species s, let a_{si}^y denote the population of s in i^{th} region for year $y \in Y$. Then what we want is if for interval of years $[x, y] \subseteq Y$ we know a_{si}^x and a_{si}^y then we would like to also know a_{si}^z for all $z \in [x, y]$. So if we know a_{si}^x and a_{si}^y, and for each $z \in (x, y)$, a_{si}^z is unknown then we have that

$$a_{si}^z = \frac{a_{si}^y - a_{si}^x}{y - x} z + a_{si}^y - \frac{a_{si}^y - a_{si}^x}{y - x} y$$

We wrote a Python code that performed linear extrapolation to estimate our missing data points.

From this we obtained our data set of estimated whale population over the years for our six species of whales. With our MATLAB code we cleaned and plotted the data for global whale populations over the years for different types of whales. Figure 2 shows the global whale population over the years for Mink whale, Blue whale, Humpback whale, Fin whale, Gray whale and Right whale.

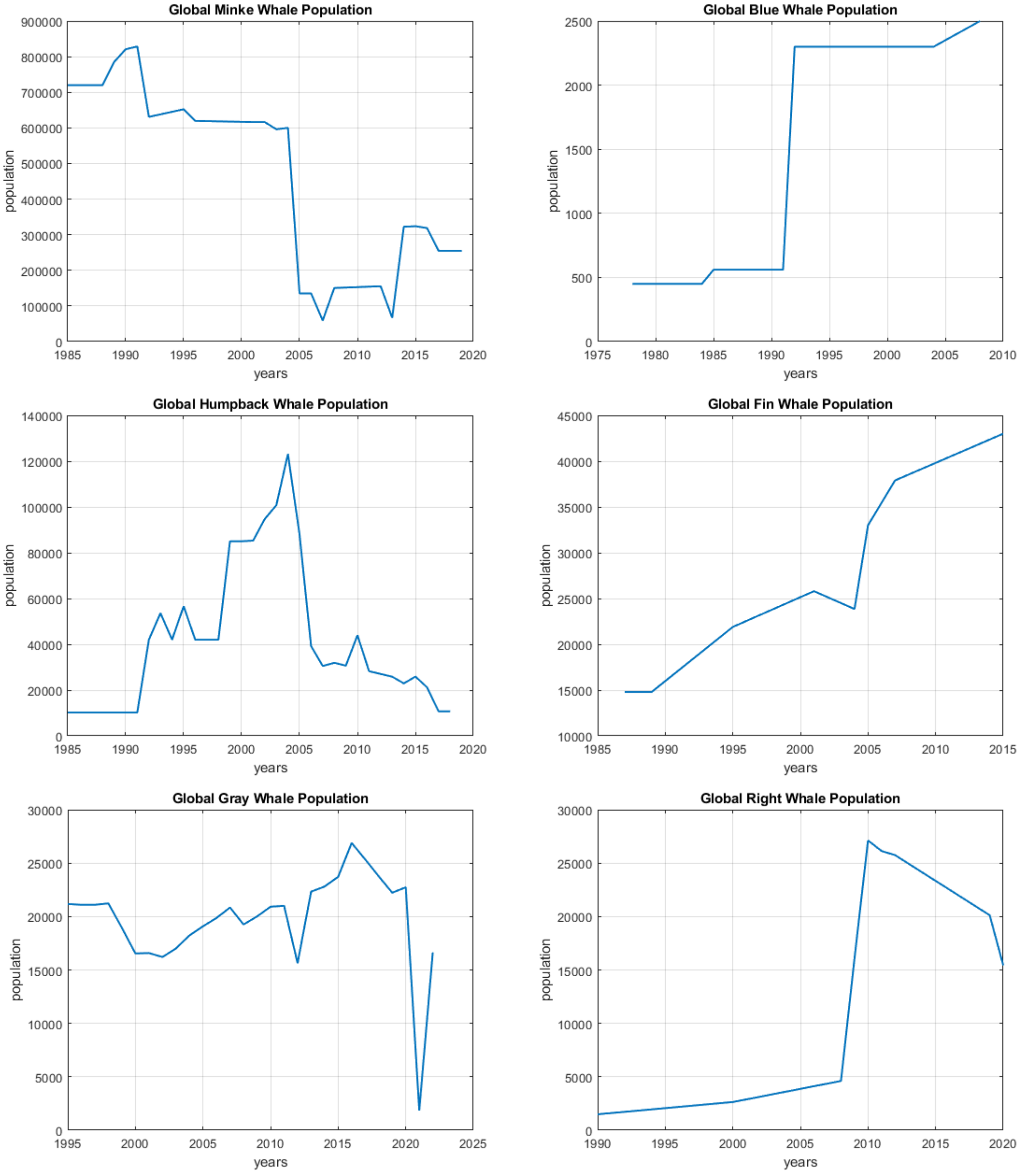


Figure 2: Global whale population over the years for different species

From this we performed linear regression on out data of oil spilled per year and whale population per year. We took amount of oil spilled in tones as our independent variable and the whale population for a specific species as our dependent variable. We wrote a Python code using the SciPy module, that we used to do our linear regression. All our code along with all our data files is available on our git repository `https://github.com/nitrodragonoid/Scientific-Methods-Project.git`.

Results and Discussion

We now analyse the results we obtained from our linear regression. Figure 3 shows the plots we obtained from our regression.

We first define our threshold of significance at 0.05 i.e., a 95% confidence interval and our p value has to be less than this threshold for us to confidently reject the null hypothesis which in our case is that global whale populations and the amount of oil spilled globally have a positive correlation with each other. Now we discuss our finding for each whale species from our data set. For Minke whales, the curve we fit on our data points is $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = 0.195x + 429000$ to three significant figures, where x is our amount of oil spilled and the $f(x)$ is the Minke whale population. Our r square value is 0.296 (to three significant figures) which is not good as its farther away from 1 meaning our curve doesn't fit the data too well. Our p value is 0.0840 (to three significant figures) which is a bit above our threshold of significance of 0.05, so we can't reject the null hypothesis in this case, so we can't rule out that any trend we see is due to mere chance. And so for Minke whales we can't surely say that the amount of oil spills globally per year effect the Minke whale population globally per year.

For Blue whales, the curve we fit on our data points is $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = -0.000856x + 1750$ to three significant figures, where x is our amount of oil spilled and the $f(x)$ is the Blue whale population. Our r square value is -0.377 (to three significant figures) which is not good as it is even more farther away from 1 as compared to Minke whales meaning our curve doesn't fit the data too well, one way to counter this is would be to try to fit a higher order polynomial on our data points. Our p value is 0.0400 (to three significant figures) which is below our threshold of significance of 0.05, so we can confidently reject the null hypothesis in this case, meaning the correlation between the global blue whale population and the amount of oil spilled globally is not due to mere chance. So we a negative correlation between the blue whale population and the amount of oil spilled. As the amount of oil spilled globally increases the blue whale population globally decreases.

For Humpback whales, the curve we fit on our data points is $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = -0.0122x + 43600$ to three significant figures, where x is our amount of oil spilled and the $f(x)$ is the Humpback whale population. Our r square value is -0.152 (to three significant figures) which is not good as its farther away from 1 meaning our curve doesn't fit the data well. Our p value is 0.3920 (to three significant figures) which is quite above our threshold of significance of 0.05, so we can't reject the null hypothesis in this case. And so for Humpback whales we can't surely say that the amount of oil spills globally per year effect the Humpback whale population globally per year.

For Fin whales, the curve we fit on our data points is $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = -0.00858x + 29800$ to three significant figures, where x is our amount of oil spilled and the $f(x)$ is the Fin whale population. Our r square value is -0.364 (to three significant figures) which is not good as its farther away from 1 meaning our curve doesn't fit the data well. Our p value is 0.0519 (to three significant figures) which is above our threshold of significance of 0.05, so we can't reject the null hypothesis in this case, however the the margin that our p value exceeded the threshold is quite small so there might still be some correlation between the amount of oil spilled globally and the total Fin

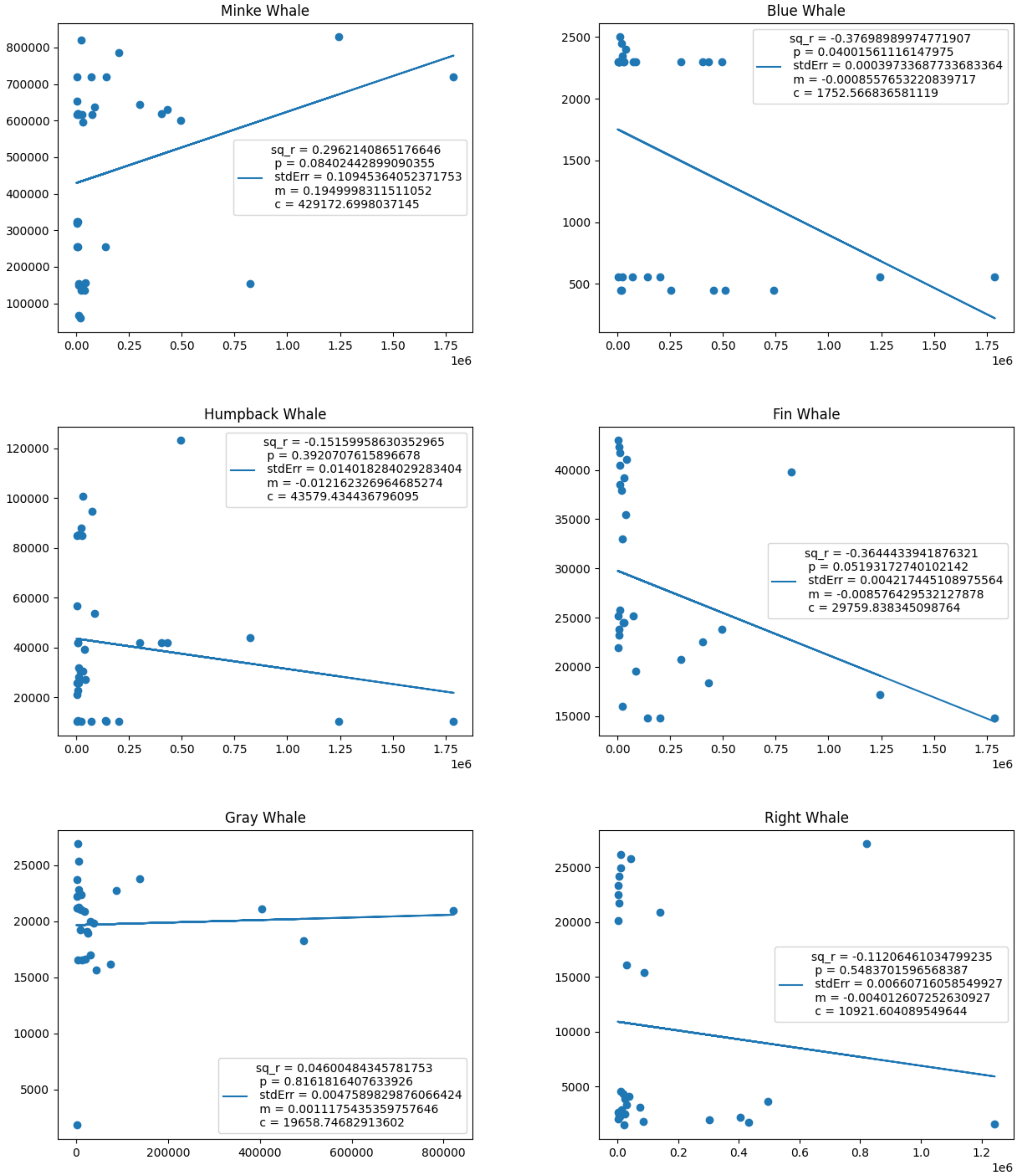


Figure 3: Linear regression on whale population and amount of oil spilled per year.

whale population per year. But as our p value $\approx 0.0519 > 0.05$, for Fin whales we can't confidently say that the amount of oil spills globally per year effect the Fin whale population globally per year.

For Gray whales, the curve we fit on our data points is $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = 0.00112x + 19700$ to three significant figures, where x is our amount of oil spilled and the $f(x)$ is the Gray whale population. Our r square value is 0.0460 (to three significant figures) which is not good as its farther away from 1 meaning our curve doesn't fit the data too well. Our p value is 0.816 (to three significant figures) which is above our threshold of significance of 0.05, so we can't reject the null hypothesis in this case, so we can't rule out that any trend we see is due to mere chance. And so for Gray whales we can't surely say that the amount of oil spills globally per year effect the Gray whale population globally per year.

For Right whales, the curve we fit on our data points is $f : \mathbb{R} \to \mathbb{R}$ where $f(x) = -0.00401x + 10900$ to three significant figures, where x is our amount of oil spilled and the $f(x)$ is the Gray whale population. Our r square value is -0.112 (to three significant figures) which is not good as its farther away from 1 meaning our curve doesn't fit the data too well. Our p value is 0.548 (to three significant figures) which is above our threshold of significance of 0.05, so we can't reject the null hypothesis in this case, so we can't rule out that any trend we see is due to mere chance. And so for Right whales we can't surely say that the amount of oil spills globally per year effect the Right whale population globally per year.

Conclusion

We see that for Mink whale, Humpback whale, Fin whale, Gray whale and Right whale, the p value exceeded our threshold of significance. And therefore we were unable to conclude if there was any correlation between the global whale population for these whales and the amount of oil spilled globally. There can be various reasons due to which correlation could not occur.

One reason for a higher p value can be that there is no correlation and the null hypothesis is indeed true. Another reason can be due to our data set. First, as noted earlier, a major percentage of the oil spills goes unaccounted for and we had to discard those oil spills for which there was no data. What could have happened was that the discarded data points accounted for a major amount of oil spilled, or since a lot of oil spills go unnoticed and therefore they would have accounted for a major part of the amount of oil spilled.

Another reason for these higher p values could be because of the missing data in the whale population. For whale population we had a a lot of missing data points, which we extrapolated from our known points. However, that could have given us incorrect and inconsistent points.

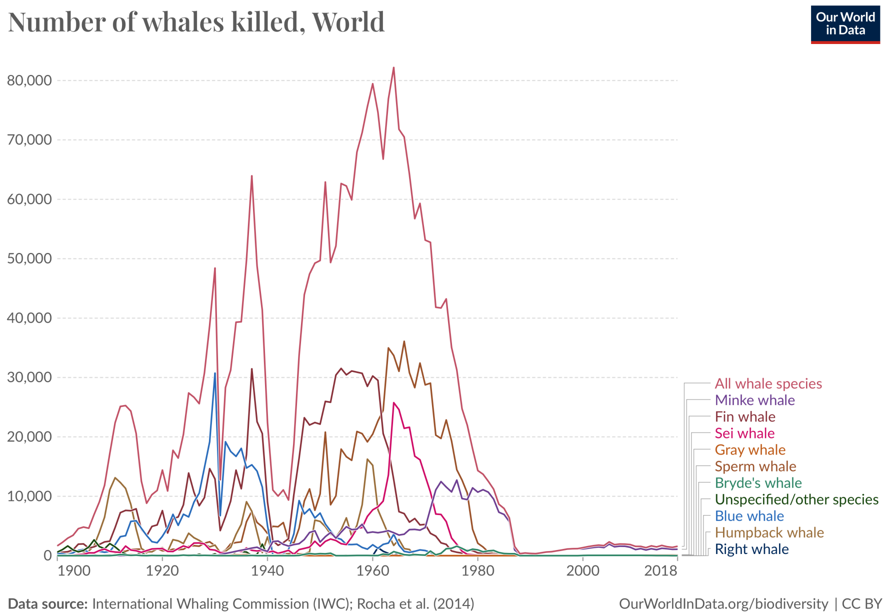
One other issue with the whale data was the non uniformity. The data we obtained was global and from different regions for different years. We filled the holes and tried to make it uniform by extrapolating our data points but it is not sufficient to make up for the missing data and for making it uniform.

The final reason possible could be some issue with our hypothesis. We compared whale population globally and with amount of oil spilled globally but this might not be a good comparison because as we observed, oil spilled in one part of the world does not seem to impact the whale population globally. This can be due to the vastness of the ocean and so maybe the oil spill impacts a smaller area. Another reason can be that some of the whales species usually are found much deeper in water and the oil spilled stays at the surface of the ocean for a longer amount of time.

We now look at the case where we did see a correlation between the amount of oil spilled globally and the whale population globally.

We saw a negative correlation in the case of blue whales. The reason we saw a correlation between our blue whale population and oil spilled was due to two major reasons. One, the data for blue whales was more uniform around the world and we had data from same location for the same years. Secondly, the location we had for the blue whale population seem to intersect with the location of the oil spills for a lot of data points. Due to this we saw a negative correlation.

We also have to take other factors in account. One major factor is the number of whales killed globally for various species of whales so, for various species of whales, the decline in whale population globally is also impacted heavily by whale killing. Figure 4 shows the trend. One good thing to notice here is that we see that the amount of whales killed is decreasing till now.


Number of whales killed, World

Data source: International Whaling Commission (IWC); Rocha et al. (2014) OurWorldInData.org/biodiversity | CC BY

Figure 4: Number of whales killed globally over the years[4]

From our analysis we can see that amount of oil spilled negatively affects the whale population in the case of blue whales, while for other species of whales we could not successfully conclude a correlation due to higher p values. As we saw a negative correlation in the case of blue whales, we can see that the oil spilled does effect some marine wildlife (it at least has been seen to affect blue whales) so we can conclude that the amount of oil spilled should be minimized to preserve the marine ecosystem. Moreover, we see a decrease in the amount of oil spilled globally in the last two decades along with the rise in global blue whale population, which is a good sign.

For future work we aim to find a larger amount of more uniform data. We will narrow our data to specific regions and see if this leads to a correlation.

References

1. *Population Estimates* Iwc.int, 2024. `https://iwc.int/about-whales/estimate` (2024).

2. Chauhan, C. *Oil spillage data — kaggle.com* `https://www.kaggle.com/datasets/warcoder/oil-spillage-data`.

3. Yuewen, D. & Adzigbli, L. Assessing the Impact of Oil Spills on Marine Organisms. *Journal of Oceanography and Marine Research* **06** (Jan. 2018).

4. *Number of whales killed — ourworldindata.org* `https://ourworldindata.org/grapher/whale-catch`.