Analysis of Indonesia's Fish Consumption with Regression Method using Go Language

Fabio Espinoza
Computer Science Department ,
Faculty of Computing and Media
Binu Nusantara University
Jakarta, Indonesia 11480
fabio.espinoza@binus.ac.id

Ravel Tanjaya
Computer Science Department,
Faculty of Computing and Media
Binu Nusantara University
Jakarta, Indonesia 11480
ravel.tanjaya@binus.ac.id

Nunung Nurul Qomariyah
Computer Science Department,
Faculty of Computing and Media
Binu Nusantara University
Jakarta, Indonesia 11480
nunug.qomariyah@binus.ac.id

***Abstract*—The study is made to predict the amount of fish consumption in Indonesia throughout the years, from the amount of fish production and catches. This study is conducted using the Go programming language to prove that even though Go is a general programming language that is rarely being used for data science, it can still be used to perform analytics and machine learning and out perform the languages that are usually used to do data science like Python and R. There are two datasets that is being used in this study, which is the fish captures data and fish consumption. Those two dataset will later be processed to a single file before being applied to linear regression and decision tree algorithms to predict Indonesia's fish consumption. The result of this study is that Go is a very capable language to be used for data science and additionally linear regression performs better than decision tree in the dataset that is being used.**

***Index Terms*—fish consumption, regression, fish capture, Indonesian ocean, machine learning**

I. Introduction

Indonesia's oceans have experienced many human intervention and have seen negative effects in species and sea plant diversity. The population of sea creatures have been plummeting due to over fishing and negligence.Fish Production seemed to peak in 1996, with global reported catches of 87 million tonnes. This number could even be as much as 130 million tonnes if discards, illegal, unreported or unregulated (IUU) catches are taken into account. Since then, catch has declined sharply, mainly due to the overfishing of many types of fish live-stocks. The resulting loss is reported to be more than one million tonnes every year [1, p. 62]. Coral reefs are also being damaged year after year. It is being degraded all over the world as the result of bleaching, cyclone intensity and human activites, this makes its number declining all over the world and it needs and immediate change to its current practices to prevent the complete destruction of the coral reefs. [2]. For this reason we strive to see what degree will human intervention have on the Indonesia's oceans in the near future and see how it effects the country's fish consumption. Eventually the fish consumption will catch up to the rate that our staple fishes can reproduce. According to the categories by WCU, most of the extinction that happens on all scle in the sea is directly attributed to 2 categories, which is exploitation that accounts 55% of all cause for extintion and habitat destruction or degradation that accounts for another 37% [3]. In order to see the rate that Indonesia's fish consumption has been increasing and how the country depends heavily around its marine life this project has collected datasets that involve fish consumption and fish catches. Further on in the paper the datasets will be analyzed deeper and show the relevance to how fish consumption may be increasing in a unsustainable rate.

Fish is a very crucial staple in world. People have been fishing since the very beginning of time and depend on their seas and local bodies of fresh water to survive. If our seas are not regulated or fishing boats do not have protocols to prevent undesired captures, then our fish consumption will be the factor that has to be regulated. Practices can be established and preventive measures should be made to circumvent the damage done on marine population and environment. In order to be able to establish responsible fisheries policies, FAO has been continuing to support the creation of fishers' related cooperatives and organizations, at the same time FAO has also continue to highlight the means of increasing production by reducing the post-harvest losses in fisheries that is categorized as small scale [4].

II. Related Work

A similar study conducted by Fesenfeld et al. [5] mentioned that employing a classical linear regression and advanced Bayesian Sparse Estimations to show the citizens prioritized political concern between meat/fish consumption and fossil-fuel car usage. This study focuses more on the political environmental factor of the citizens concern.

Another article aiming to see what fishes were the most popular to be consumed in Turkey was conducted by FST [6]. The results seem to be that Five of the most consumed fish species has accounted for 76% of the total fish consumption and three of the most liked and most consumed fish is Anchovies, git-head and sea bream [6].

While this project does not highlight the negative effects of over fishing too much, another study by the UN does. This study by Hazin states, There are a few ways how marine ecosystem can be affected with the capture fisheries such us the alteration the body-size and population size of a certain species that may lead to small individual organism size [4]. Although our project does state that the consumption rate and capture rate seems to be unsustainable, it is very important to know how these consequences are affecting marine life as well.

However what if the fish the people are consuming can be negative? The common conception is that fish is very healthy for you and will improve most health conditions. A paper done by Brookhaven National Library conducted research to find the correlation between fish consumption and human heart disease showing that fish contains high level of mercury. However, in all of the multivariate regression models, the effect found of the fish consumption was negative, although it was never statistically significant [7]. This shows that fish can ber harmful when over consumed. Another supporting and related work is an article by Circulation that states fish consumption is inversely linked with deadly CHD. The risk of negative side effects froms fish and CHD may be reduced by eating fish once per week [8].

Our project is more geared towards the future fish consumption of Indonesia and is objectively using multi variable regression in order to see the trend that Indonesia's fish consumption rate will be taking. The datasets are unique to Indonesia and are a clear depiction of our modern time since the datasets extend to 2017. The model will predict a clear trend of how Indonesia fish consumption rate will be increasing or decreasing and from that prediction there can be more support to regulate marine and aquaculture practices.

This study is special and different from the other studies as firstly we use the Go programming language. Go is commonly used as a programming language in cloud native application and command line application, but even though it is uncommon to be used in data science, we believe that it is can really suitable to be used for data science applications especially for those we are incorporating a huge amount of dataset processing as it has a really fast compilation and runtime speed, thus resulting in a more productive workflow. The other important selling point of this study is that rather than performing the study on a world-wide basis, we try to perform it on a more specific country to get a better view of what the consumption should be in the future.

III. Methodology

A. Datasets

For this study we are using these datasets provided. The first dataset is Fish Consumption[1] and the second dataset is Fish Catches[2]. These datasets were provided by public sources found online. Additionally the dataset is sourced from the UN Food and Agriculture Organization (FAO) and uses data from its Food Balance Sheets into a complete linear dataframe from 1961 to 2017. In the original FAO dataset, food supply data from 1961 to 2013 is stored under its 'old methodology' variable set. Data from 2014 to 2017 was stored under its 'new methodology' for food balance sheets. Most of these datasets take into account a global environment so many countries are included in these datasets. We used Indonesia which is a very good example of a country optimal for fishing, because it is the home country of this report and has one of the most highest fish consumption and fish captures out of all the countries in our dataset. Indonesia's fish consumption and fish capture rate should compare to other countries and resemble the same increase and decrease of data.

[1]https://bit.ly/3nfc1DH

[2]https://bit.ly/3IWW6To

TABLE I
Fish Catches Dataset Snippets

Country	Code	Year	Production	Catches
Indonesia	IDN	1969	105,690	1,129,110
Indonesia	IDN	1970	108,706	1,148,494
Indonesia	IDN	1971	114,121	1,159,181
Indonesia	IDN	1972	118,952	1,177,852
Indonesia	IDN	1973	129,830	1,164,611
Indonesia	IDN	1974	136,244	1,225,698

1) Fish Catches Dataset: This is a dataset that shows the amount of fish captured and produced in countries worldwide. This dataset is containing 5 columns and 14,675 rows, containing various countries spanning of all continents except Antarctica. This is one of our larger datasets that contain information dating back from the 1960 up until 2017. With this dataset we can actually compare how the rate of catches has increased and see the correlation between the fish boat escalation vs the fish catch escalation. The correlation between the two do not seem too strong if we take the earlier statistic for Indonesia. In a span of 7 years the rate of increase is 3.3% for fish boat production and the fish catch production on the other hand far exceeds 3.3% as seen on the visual below. The sample of the dataset is provided in Table I.

Indonesia is a prime country to examplify the fish production of a country in Asia. Regional fish production is the area where production increment is the most significant on, although there are a stable increase in world total production since the early 1960s, the highest regional development is observable in Asia [9]. Using Go we have managed to provide visual histograms based on the country fed to the Go program as can be seen in Figure 1. For this case we use Indonesia which is leading in fishing and aquaculture production. Our dataset provides information for many multitudes of countries since its a global dataset, but we use Indonesia since it is the main focus of this project.

2) Fish Consumption Dataset: Fish consumption dataset is a dataset that show the amount of consumed in worldwide from the year of 1961-2017. This dataset is containing 4 columns and 11,029 rows. To further clarify this dataset is showing how much fish consumption in kilograms will an average citizen eat in a year. Similar to our last dataset this one is quite large as well containing various countries of all continents excluding Antarctica. Luckily the data ranges all the way back from the 1960s to 2017 so the rate of increasement in consumption is more vividly shown. Fish consumption will be our target for the machine learning model and the

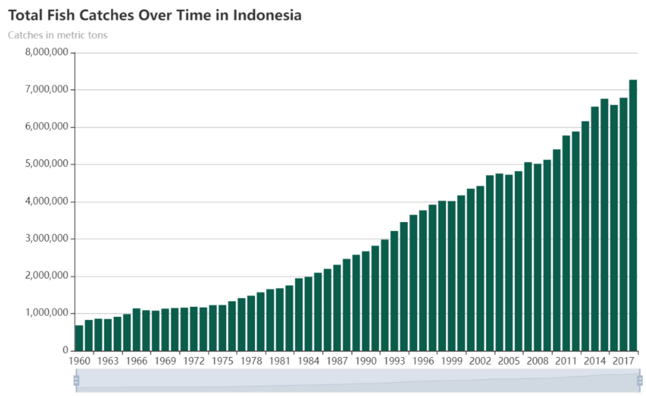

Fig. 1. Total Fish Catches Over Time in Indonesia

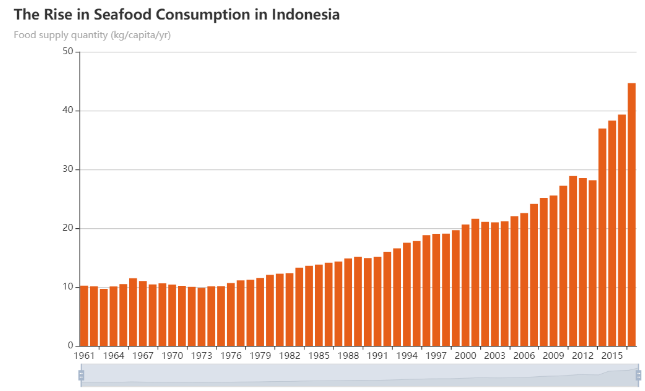

Fig. 2. Total Fish Consumption Over Time in Indonesia

independent variables will be shown later in the report. The reason for predicting fish consumption is because if there is no demand for fish then productivity will have no incentive to increase. Fish consumption is the root of the marine and aquaculture industry so we can assume that the prediction of fish consumption will determine if the slope of all the other variables will be negative or positive in the future. The sample of the dataset is provided in Table II.

The large increase of fish consumption may result in underlying health problems for Indonesia's population that may not have been expected. Two decades ago, epidemiologists discovered a low coronary heart disease (CHD) mortality rate among native Alaskan and Greenland Eskimos who ate a large amount of fish in their diets. The same situation was also seen among Japanese citizens [8]. Along with regulating the fish capturing methods for commercial boats the country will also benefit in advising the citizens about the dangers of consuming too much fish. Another study also shows a negative health effect when a sample population has eaten too much fish. Deaths tolls were to be speculated through a yearly timeframe to be 2000, and diet was assessed by means of a questionnaire at entry. Relative risks of death from various causes were assessed according to either fish consumption or intake of fatty acids [7].

This is no surprise that Indonesia is one of the leading countries in fish consumption. A surprising statistic is that a lot of predominantly European and Asian countries are the ones that do the most fish consumption globally. People of races other than Black and White actually have the highest fish consumption rates of all other race and ethnicity groups, with significant differences observed across all fish types [10].

Using Go's visualizing package, go e-charts, we have filtered the the fish consumption by country and provided a histogram to show the rise of consumption over 6 decades in each country in Figure 2. The only country in this project is Indonesia therefor that is the only country that will be displayed. The y-axis of the histograms is the food consumed by each citizen in kilograms and scales directly with the country so some other histograms may have a different upper limit than others. These histograms are intractable and can be scaled to only show less or more years for more pinpoint comparisons from year to year.

TABLE II
FISH CONSUMPTION DATASET SNIPPETS

Country	Country Code	Year	Fish
Indonesia	IDN	1961	10.24
Indonesia	IDN	1962	10.12
Indonesia	IDN	1963	9.68
Indonesia	IDN	1964	10.1
Indonesia	IDN	1965	10.49
Indonesia	IDN	1966	11.51

B. Data Preparation and Preprocessing

The data preparation and processing part is fairly minor since the datasets are already having the required shape to be used for our model/visualizations. Cleaning the data was done with Go Dataframe packages that allow us to find null indexes within our dataframes. They also allow us to drop unwanted columns and rename columns so the dataframes can be merged later on with the same column names. Once the datasets were all acquired, they were ready to be transformed into dataframes for further analysis and filtration. The visualization of the data processing can be seen in Figure 3

Go enables our CSV files to be read and parsed into a dataframe using the GoCSV and Dataframe package. Afterwards the dataframes were analysed for any potential merging and correlation. The only two dataset which is the Fish Consumption and Fish Catches will then be merged into a single struct in Go and then they will be used produced a dataset with 9019 records. The Go packages mentioned earlier also allow writing the dataset into a CSV file for exporting purposes. This is crucial because the only way to access the merged dataset is from a CSV file. Now the dataset is exported into our project directory and can be used for the machine learning model that we will implement.

C. Model and Techniques

For our model we use a supervised learning algorithm for continuous data known as Linear Regression and Dicision Tree

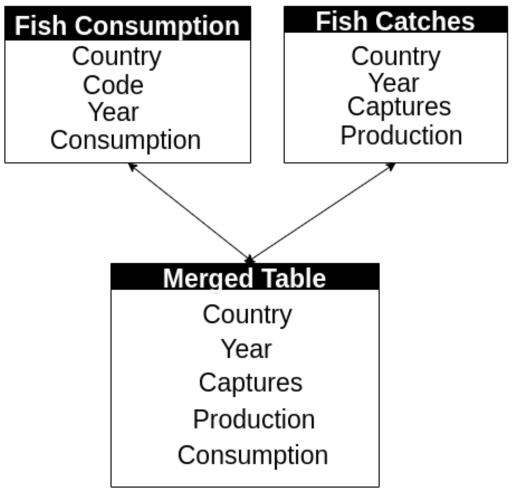

Fig. 3. Data Merged from 2 Different Dataset

Regression. Since our model will be predicting continuous data like fish consumption, it is optimal to use a regression algorithm compared to a classification algorithm. Discrete data plays a role in classification algorithms but none of the datasets in this project contain discrete data. Linear Regression works in a way where it finds the best fit line to determine the pattern our data is showing over time. The fish consumption over time is exactly what we are trying to successfully predict in this model. Regression works well with a lot of data points throughout a wide range of an independent variable. Our independent variable will be time/years for our model. Additionally linear regression will require all of our inputs to be numerical so we will just have to transform non-numerical data into numerical keys. We can easily do this with Go and the use of the gota package. This means countries and continents need to be transformed into their own unique number correlating with their value. Merging of our datasets will be done based on related columns or otherwise known as primary/foreign keys. Once the dataset is formed for our model, a training and testing data split is generated by Go. In order for the model to be validated we have to use a method to measure the error. The most common ways of measuring the error of the model or how likely the best fit line will predict new data is MAE or RMSE. MAE (Mean Absolute Error) measures the level of magnitude the prediction's error is compared to the testing data. RMSE (Root Mean Squared Error) does the same thing but the difference between them is that outliers in the dataset are given a larger weight when RMSE is applied. Since our dataset is dealing with many countries around the world with varying economic and developing situations, it is more effective to use MAE for our model.

D. Tools

The tools that we use in this research includes:

- Go Programming Language
  We use Go for the main and only programming language of the study. We use Go because it is statically typed and its compiled its really fast, thus improving the overall development experience. Even though Go is not a really big player in the field of Data Science, more and more tools and library has emerged to help Go programmer to do data science related actions.
- Go CSV
  This is a handy package that allows us to export our manipulated data into a new csv file for later implementations.
- Go-Sajari-Regression[3]
  This is the library we use to do the prediction, as this library runs on Go it really outperforms it Python and R counterpart.
- Go E-Charts[4]
  Go E-chart is a web based data visualization tools that is used for the visualizations of our dataset.
- Go Learn[5]
  Go Learn Package is a package very similarly resembling SKLearn in Python. Provides our project with many practical machine learning functions and tools to create models.

E. Evaluation

To evaluate the model, we are using MAE, RMSE and NRMSE as the standard method of measurement. We also apply data splitting by 70% and 30% for the training and the testing of the machine learning model.

IV. Results and Analysis

After evaluating the result of the training, we also see that there is actually no problem regarding to overfitting and underfitting. We can see from the result that the machine learning model is completely independent of the data and can still perform well where the data is completely unrelated with the one that is being used for the training. Here is a visual representing the best fit line and at what rate Indonesia's fish consumption is increasing in. This visual can be overwhelming but the main focus is the line being show. The slope of the line represents the direction the fish consumption has seen in the past years. In this case it is an upward trend.

In the beginning our model was suffering from overfitting. Overfitting resulted in our model displaying unwanted behavior for our final prediction which resulted in a much higher MAE. The reason for this unwanted behavior was because the model is performing very well on our training data but generalizing very poor on our testing data. In the case of this project, the overfitting was happening because we were not using enough data for our training set. Once we set the training and testing split to 70% / 30% the overfitting seemed to be resolved. The increase in size of training data let the model be trained for more types of changes and patterns found within

[3]https://github.com/sajari/regression
[4]https://github.com/go-echarts/go-echarts
[5]https://github.com/sjwhitworth/golearn

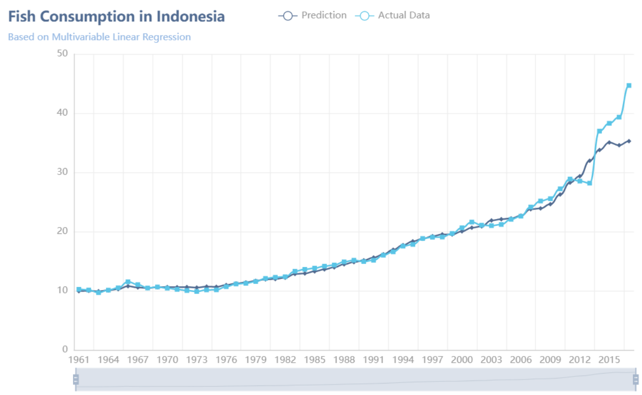

Fig. 4. Visualized Graph for Linear Regression Prediction

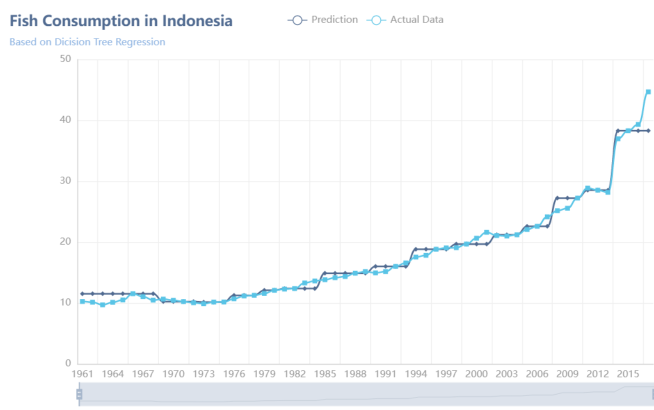

Fig. 5. Visualized Graph for Decision Tree Prediction

the training data and generalize a better prediction for our testing data.

Figure 4 shows the prediction of fish consumption in Indonesia using the multi variable regression technique, the MAE, RMSE and NRMSE of the technique is respectively 0.38, 0.48, and 0.01. We can also see in the plot as the point are very close to each other and have low variance.

Figure 5 shows prediction of fish consumption in Indonesia using the decision tree regression method, the MAE, RMSE and NRMSE of the technique is respectively 6.91, 10.07, and 0.30. which is also visible by variance in the line graphs.

These rises in Indonesia's fish consumption is self explanatory by the visual. The trend seems to be increasing at a considerable pace and if the country wants to continue to indulge in this fish consumption rate then there has to be regulations set in place. Considering there a new endangered marine species almost every year and marine habitats are being threatened by the over fishing techniques that many countries are implementing, the ocean cannot keep up with this pace in consumption. This project hopefully will bring to light how the seas should be regulated and marine life should be protected.

Table III show the final statistics of the prediction, we can see that Linear regression is the best performing algorithm of

TABLE III
MAE, RMSE, AND NRMSE OF LINEAR REGRESSION AND DECISION TREE

Algorithm	MAE	RMSE	NRMSE
Linear Regression	0.38	0.48	0.01
Decision Tree	6.91	10.07	0.30

the two.

V. Discussion

The conclusion of the project is represented in this regression formula generated by our model.

A. Best Model For Prediction

After taking a look at the graph and MAE, multivariable linear regression seems to be a better model to be used to predict the amount of fish consumption in Indonesia, this is proven by the model having a lower average MAE of only 0.38, RMSE of 0.48, and a NRMSE of 0.01.

B. Regression Formula

$$\text{Prediction} = 101.7895 + \text{Year} \times -0.0483$$

To translate this formula this means the predicted fish consumption in kg per year will be 101.7895 + Year × -0.0483. For instance in 2021 Indonesia's fish consumption will be around 21 kg for every human per year. The -264.3149 intercept shows what the value of fish consumption would be on the year 0. Since we are in the year 1950's and above this slope is more drastic than compared to a slope that accounts from the year 0 to now. Realistically since we cant provide data from the years before our data sets provide then we should not take them into account. Therefore the y intercept is just a result of our years starting around 1950. Lastly any input of any year in this formula from 1950 and above will give a reasonable output to the country's fish consumption. We can theoretically predict Indonesia's fish consumption for the year 2040 using this regression formula and we can get a realistic answer. From the model it is apparent that Indonesia's fish consumption will be rising and thanks to our datasets this model has given us a unique prediction for the country's consumption rate.

C. Go Environment

This project has shown the power that Go holds in the data science/artificial intelligence field. Although being still a young language in the field of Data Science, Go has been a really great general purpose programming language that has improved a lot through the years, and even though it is now mainly being used to do web/micro service development and Command Line Application development, Go has a vast amount of packages to help data scientists build and analyze data science related projects. Many packages are complete and some of them are in development but it is a matter of time until Go becomes a very considerable language to learn for building predictive models and visualizing big data.

Understandably the data field is dominated by Python, R, MATLAB, and other tools and languages. Using these tools is far easier as their dynamic and scripted nature and is a more established way of building data projects, however Go shows a lot of potential compared to these languages. Go, being a compiled and statically typed language will have a naturally more efficient and faster run time than these interpreted and dynamically typed languages. The cost is that Go will relatively require longer time for the project to complete, meaning more lines of code to be written compared to these other tools that have very minimal lines of code, but at the same time gives better run time speed and also better debugging capabilities. At the moment Go is a language to use for efficiency, optimal performance, and great structure for your data project. Go would really shine in creating a more complex and bigger project like a Neural Network perhaps. Since the speed of the program would far outmatch a Neural Network written in Python. Unfortunately in terms of productivity, Python would be beating Go because of the lower learning curve and available libraries that Python has available. We hope that our project inspres more data projects to be written in Go and that the Go ecosystem keeps improving to allow for more productive and innovative ways to develop data science projects in Go.

VI. Conclusion and Future Work

In conclusion, this study shows how linear regression's and decision tree's accuracy compares to each other when they are being used to predict fish consumption in Indonesia using the fish consumption and fish catches dataset. It also proves that Go programming Language is a very suitable programming language to be used when doing data science projects and can outperform languages like Python.

The result of the study also highlights how the linear regression is having better results in predicting the fish consumption in Indonesia by having an NRMSE of 0.01 while at the same time decision tree get the prediction done with NRMSE of 0.30. The result of both algorithms are also compatible with the real data that shows that the fish consumption in Indonesia will continue to increase in the upcoming years.

In the future, it is possible to improve this study by doing a comparison between how Go's is more efficient than the other common data science language such as Python and R in processing this dataset and doing the prediction. It is also possible to add more algorithms to predict the consumption of fish.

Supplementary Source Code

To see the source code please head to the following link: https://github.com/raveltan/fish

References

[1] D. Gascuel, "Overfishing and sustainable fishing: challenges for today and tomorrow," 2020. [Online]. Available: https://ocean-climate.org/wp-content/uploads/2020/01/8.-Overfishing-and-sustainable-fishing-scientific-fact-sheets-2019.pdf

[2] L. S. Van den Hoek and E. K. Bayoumi, "Importance, destruction and recovery of coral reefs," *IOSR Journal of Pharmacy and Biological Sci*, vol. 12, no. 2, pp. 59–63, 2017.

[3] D. Nicholas, "Extinction vulnerability in marine populations," 2003. [Online]. Available: http://www.dulvy.com/uploads/2/1/0/4/21048414/dulvy_etal_2003_faf.pdf

[4] F. Hazin, E. Marschoff, J. Rice, and A. Rosenburg, "Capture fisheries," 2016. [Online]. Available: https://www.un.org/Depts/los/global_reporting/WOA_RPROC/Chapter_11.pdf

[5] "The role and limits of strategic framing for promoting sustainable consumption and policy," *Global Environmental Change*, vol. 68, p. 102266, 2021.

[6] M. Ferit, A. Gunlu, and H. Yesim, "Fish consumption preferences and factors influencing it," *Food Science and Technology, Campinas*, vol. 35, p. 345, 2015.

[7] F. Lipefert and T. Sullivan, "Fish consumption, methylmercury, and human heart disease," 2005. [Online]. Available: https://www.bnl.gov/isd/documents/31076.pdf

[8] K. He, Y. Song, M. L. Daviglus, K. Liu, L. V. Horn, A. R. Dyer, and P. Greenland, "Accumulated evidence on fish consumption and coronary heart disease mortality," *Global Environmental Change*, vol. 109, p. 2711, 2004.

[9] Y. Ye, "Historical consumption and future demand for fish and fishery products," *FAO Fisheries Circular*, vol. 946, p. 2711, 1999.

[10] B. Jeffrey, R. J. Birch, and J. Rogers, "Estimated fish consumption rates for the U.S. population and selected subpopulations," 2014. [Online]. Available: https://www.epa.gov/sites/default/files/2015-01/documents/fish-consumption-rates-2014.pdf