More Money, More…Seaweed

Emma Mavis

IDS 702 Fall 2021

I. Summary

The global aquaculture sector has the potential to greatly reduce the effect that increasing greenhouse gas emissions have had on the Earth. Sustainable aquaculture practices can battle the phenomenon of ocean acidification (through carbon sequestration). Thus, investigating the countries that have had an aquaculture sector in place or have increased their production over the years can inform how to increase production amongst the rest of the world now and into the future.

I have shown a correlation between the amount of aquaculture that a country produces over time and demographic factors such as GDP per capita, population density, and the gender parity index in primary schools. There are of course many outside variables that have not been controlled for such as location, economic demand, and the divisions within a country's agriculture sector. My aim is simply to show that as these demographic indictors grow towards more equality and more wealth, so too does the production of aquaculture.

II. Introduction

In this research, I focus on how the rate of aquaculture production has developed globally since the 1960s, what qualities are attributed to countries with high rates of production, and what relationship can be determined between these demographic factors and production rates. The chosen demographics that are included in the model for aquaculture production are GDP per capita, carbon dioxide emissions, population density, urban population percentage, the years that people spend in primary school, and the gender parity index in primary school.

III. Data

Two data sets were used for this project. The first was gathered from the Food and Agriculture Organization of the United Nations and contains data on every species of global aquaculture production from 1950 until 2019. Using their data portal, FishStatJ, I subsetted out metric tons of total aquaculture production for each country since 1950 in their database. There were no missing values in this data set (of course many countries had 0 tons of production for some years).

The second data set was gathered from the United States Census Bureau and their World Development Index database. Because there were an enormous number of indicators recorded for almost every country since 1960, I initially filtered the data set down to just the indicators below a certain threshold of missing values. From that subset, a dozen target demographic variables were chosen to stay in the final data set, along with the country name and year.

After reshaping, the data was ready to merge on unique country-year identifiers. That way, for example, India in 1971 could have its tons of aquaculture production directly in line with information about its population that year, its GDP per capita, etc. The largest obstacle in the way of a clean merge was that each data set had some countries included that the other did not, and very many of the countries that did match were spelled differently in each set. Thus, I selected just the countries that each set both contained and changed all to a common spelling. The final data set contains 192 countries with 59 years of observations.

Summary Statistics of all Countries From 1990 - 2019

	Mean	Min	Median	Max
Aquaculture Production (tons)	71,561	0	56	8,101,298
Carbon Dioxide Emissions (kt)	86,276	18	4,976	4,688,373
Population	32,210,000	40,350	6,087,000	1,119,000,000
Population Density (/ km²)	101.8	1.2	43.1	1096
Urban Population (%)	44.4	3	42	96.3
Years People Spend in Primary School	5.9	3	6	7
GPI in Primary School (# females / # males)	0.87	0.15	0.96	1.23
GDP Per Capita	8,030	138	2,845	111,657

Summary Statistics of all Countries From 1959 - 1989

	Mean	Min	Median	Max
Aquaculture Production (tons)	361,119	0	2,263	66,135,059
Carbon Dioxide Emissions (kt)	147,016	0	11,380	10,313,460
Population	37,000,000	8,910	8,614,000	1,393,000,000
Population Density (/ km²)	130.1	1.8	71.4	7,953
Urban Population (%)	54.1	5.4	55.3	100
Years People Spend in Primary School	5.7	3	6	9
GPI in Primary School (# females / # males)	0.95	0.43	0.98	1.25
GDP Per Capita	11,317	163	4,224	87,981

IV. Exploratory Data Analysis

The response variable in this investigation is production of aquaculture and will be referred to from this point as "production." The distribution of its raw values was heavily skewed, so a log transformation was performed and greatly improved the normality of the distribution. Of course, upon a log transformation, all observations of 0 tons of production are lost, which is about 17% of the data. Since normality is an important assumption to hold moving forward, this 17% was dropped. The values that were 0 mostly occurred in the 1970s and 1980s when aquaculture was not as large of sector as it is now, and countries that still had 0 tons of production after 2010 were small island countries, which again, most likely do not have a large aqua/agriculture sector and rely mainly on tourism.

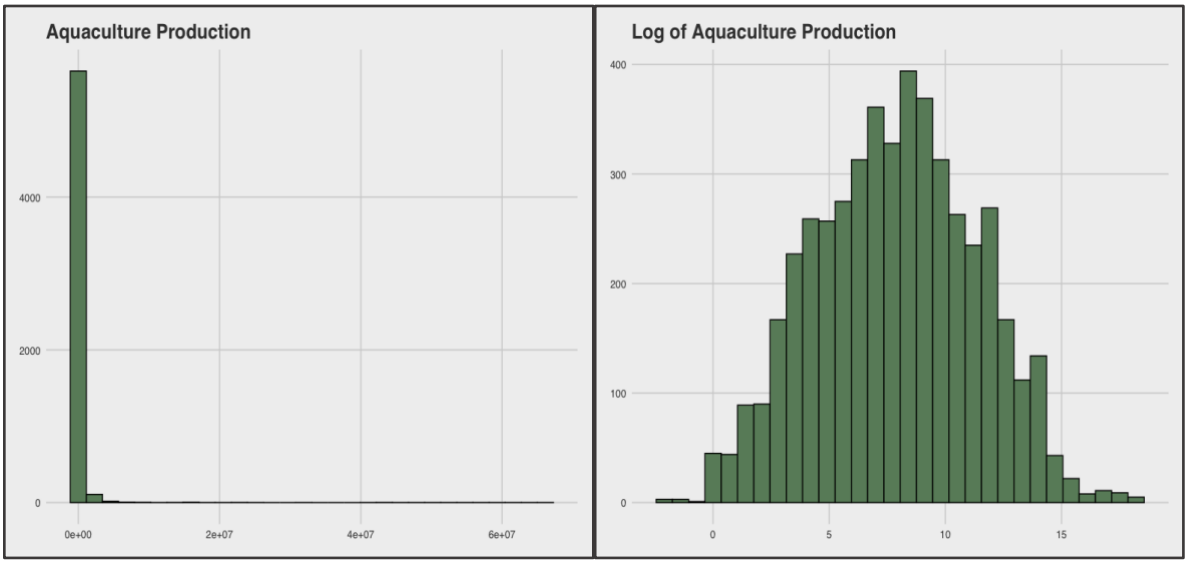

Figure 1: Response variable before and after logarithmic transformation

Initially plotting each potential predictor against the log of production was not encouraging towards a linear model; indeed, the most linear relationship that could be obviously seen was against year, as shown in *Figure* 2.

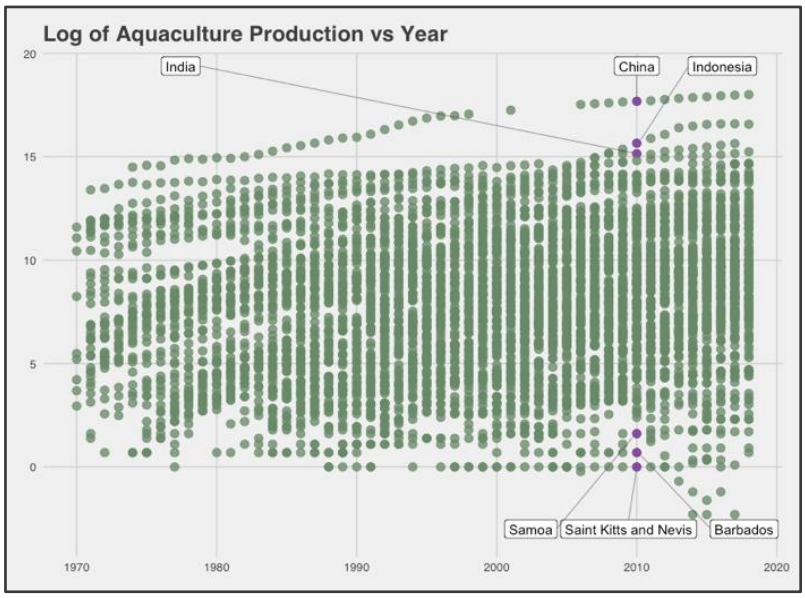

Figure 2: Purple points highlight several countries with relatively large and small aquaculture sectors in 2010

This issue does get addressed in modelling, in which further transformations of predictor variables greatly improve the overall linearity.

Of course, several of the predictors explored ultimately are significant in modeling the log of production, along with interactions with the two categorical variables. The first categorical variable is the number of years that people spent in primary school, which ranges from 3-9 years. The second is a grouped version of GDP per capita: a value of 1 for less than $5000, 2 for $5000 to $20,000, 3 for $20,000 to $40,000, and a value of 4 for any countries greater than $40,000.

As shown in *Figure 3*, a change in the linear trend does happen across the different GDP levels. Interestingly, those country-years with higher GDP per capita also tend to have greater proportions of urban population.

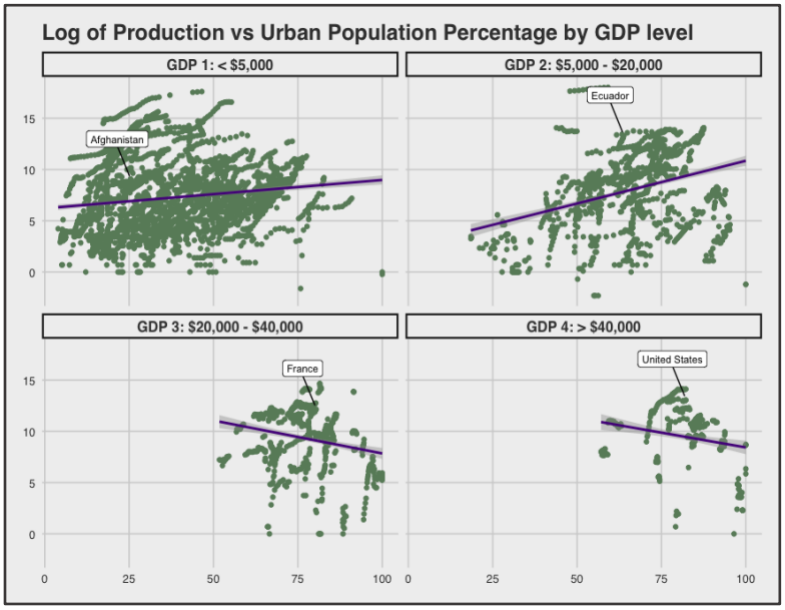

Figure 3: Displays the changing linear trend between production and urban population percentage across different GDP per capita levels. Each labelled country is at year 2018

After initial exploration of the predictors and interactions, I started model building with the following variables as main effects:

- Carbon dioxide emissions (continuous, centered)
- Population (continuous, centered)
- Population density (continuous, centered)
- Urban population percentage (continuous)
- Number of years people spent in primary school (categorical)
- The Gender Parity Index in primary schools (continuous)
- GDP per capita level (categorical)

and considered interactions between all continuous variables and the two categorical variables.

V. Model

The first model of all variables as just main effects showed that further transformations of the predictors were needed. Population, population density and CO_2 emissions were log transformed, and the improvement in linearity can be seen when comparing the plots of the residuals before and after, as shown in *Figure 4.*
After variable selection and model assessment, the final linear model is as follows:

$$\log(\widehat{Production}) \sim GDP\ Level * (\log(Population\ Density) + Urban\ Population\ Percentage) + Year + Years\ Spent\ in\ Primary\ School + GPI\ in\ Primary\ School + \log(CO_2\ Emissions)$$

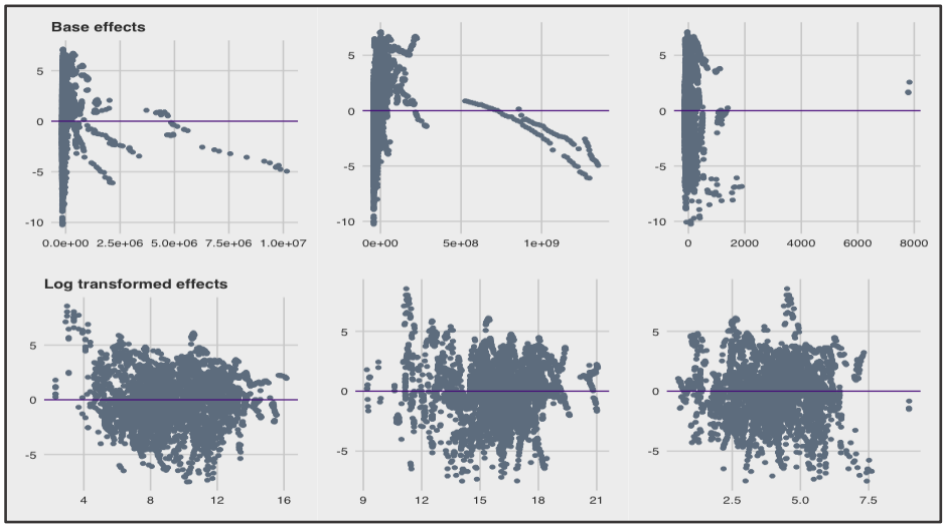

Figure 4: Residuals of first main effects model, before (top) and after (bottom) a log transformation of 3 predictor variables

The GDP level variable is interacted with both the log of Population Density and the Urban Population Percentage, which means that the relationships with the log of Production and these two latter variables change across different levels of GDP per capita.

Variable selection was performed through checking multicollinearity, nested F-tests with the interaction terms, and cross-referencing stepwise AIC and BIC model selection. Both AIC and BIC arrived at the same model, and analysis of the assumptions was performed, as seen in *Figure 5.*

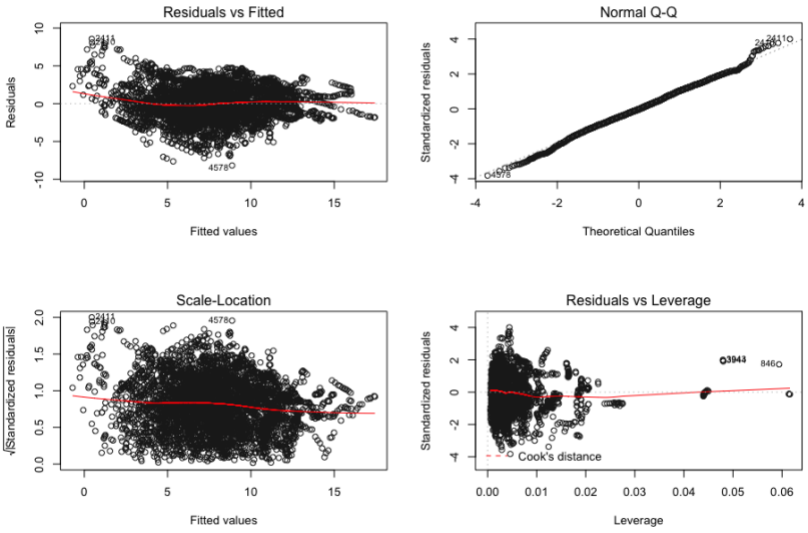

Figure 5: Plots displaying model's linearity, normality, constant variance, and outliers

Overall, the model assumptions are met fairly well. There are potentially some issues with linearity – there is a slight curved trend that could be remedied with additional transformations, but at the expense of interpretability. Normality is good, and the standard residuals show constant variance. There are some outlying points with relatively high leverage, but leverage values of 0.04 – 0.06 are not concerning. There is one point, El Salvador in 1997, that has leverage of one. No changes occurred to the model and its assumptions when the point was removed, so it stayed in the data set.

The values of the coefficients and the corresponding error bands are shown in *Figure 6.*

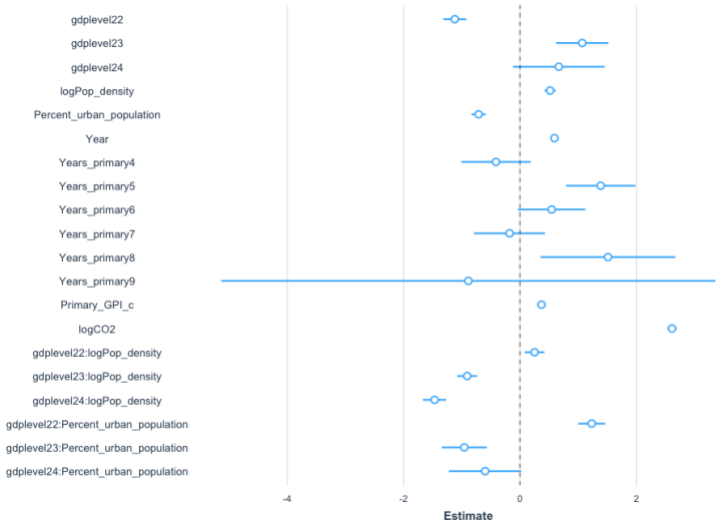

Figure 6: All model coefficients with 95% confidence interval error bands

All variables proved to be significant from the results of the nested F-tests and AIC/BIC model selection, but clearly some levels of *"Years_primary"* (number of years spent in primary school) have large standard errors (and therefore high p-values). That is acceptable though, since the variable proved to be necessary in this model. Additionally, it has an R^2 value of about 62%, and all variables apart from GDP Level and its interactions have VIF scores of less than 5, which shows that there are no serious multicollinearity issues.

In context, we have some interesting interpretations of these numbers (these interpretations are subject to "holding all other variables constant"):

- For each additional year that passes, countries increase their production by about 5%.
- A 10% increase in the gender parity score results in a 40% increase of production.
- Every additional kiloton of carbon dioxide that a country produces is correlated with an additional 3 tons of production.
- There is no apparent increasing or decreasing trend of production as the number of years that people spend in primary school increases.
- Compared to countries with GDP per capita less than $5000, countries with GDP per capita greater than $20,000 have a 500 – 900% increase in production.
- Each additional person added to the population density per square kilometer results in an additional 1.5 tons of production.

VI. Conclusions

Despite the many hidden confounding variables, this research has opened a window into how countries can increase their aquaculture production. Overall, greater equality, access to primary education, and higher wealth appear to be strong signifiers in how much a country produces. Of course, one of the biggest limitations in this analysis was missing data. Many of the World Development Index indicators were unusable due to a large proportion of missing observations. Additionally, the hidden variables limit how much can be interpreted. But they do not have to be a limitation going forward – additional geospatial information can be acquired to further inform how certain countries produce more aquaculture than others, and even chemical ocean acidification data can be gathered in those regions to determine if aquaculture does indeed have as great an impact at battling the effects of climate change as many proponents of the practice proclaim it does.

https://github.com/egmavis/Aquaculture_Global_Production