Do We Need More Babies to Save the Planet?

A Study on Greenhouse Gas Emissions & Population Growth

MGT 6203 Progress Report - Team 64

Melissa Gibson, Stephen Kim, Nitya Nandagopal, and Jessica Petersen

Background:

The increase in greenhouse gases (GHG) since the industrial revolution has had an undeniably detrimental impact on our environment, with carbon emissions making up a majority of the GHG emissions. Over that last 40 years, “CO2 emissions have increased by about 90%, with emissions from fossil fuel combustion and industrial processes, contributing about 78% of the total greenhouse gas emissions increase from 1970 to 2011,” (EPA, 2023).

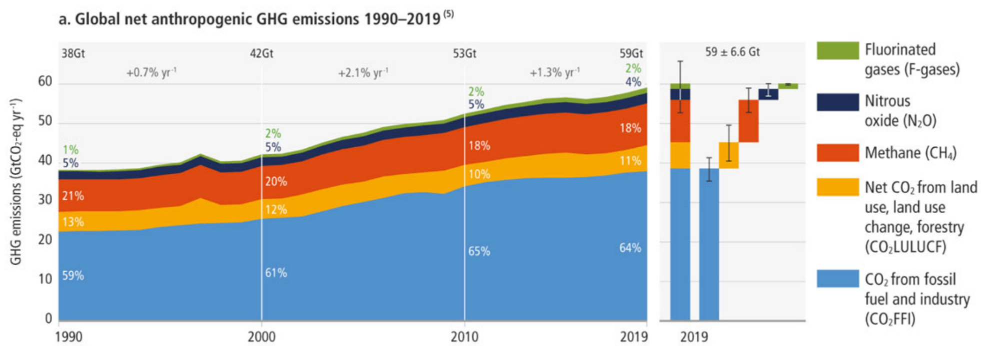

Figure A

Population growth is commonly thought to increase GHG emissions and negatively impact our global environment due to the strain on resources and increased production necessary to support a larger population. This strain on resources and threat of overpopulation in some countries resulted in population control measures imposed by various governments throughout the past 50 to 100 years. Now, however, global fertility rates have sharply declined as a result of these measures and other changing values of modern society, which in turn has caused the global population growth rate to slow over time. Many countries now face the threat of a considerable population decline by the end of the century. With the global population soon to be entering a decline, its effects on GHG emissions are still unknown.

Initial Hypothesis:

With the impending decline of the global population, there is now a question as to whether or not population significantly contributes to GHG emissions. If the world population shrinks considerably, and younger generations must support the larger, aging population, who will continue to innovate and improve the future health of the planet? Without a robust population to carry on the work and innovations of prior generations, how will the world continue to develop? The purpose of our investigation and analysis is to examine if and how population growth contributes to GHG emissions. Our hypothesis is that population growth does not significantly contribute to increased GHG emissions. In fact, we do not expect population to be the most contributing factor to the increase in GHG emissions.

Dataset Descriptions:

The primary datasets used by the team are Greenhouse Gases Emissions Data (1990 – 2019), which lists GHG emissions per country, and World Population Growth Data (1962 – 2021), which lists population data per country. These contain the two main variables we are trying to compare in our analysis. Additional datasets include those that show the age-dependency ratio, fertility rate, agricultural land %, birth rate, exports of goods and services, manufacturing value added, etc., for the countries listed in our primary dataset over the time-period of interest. A full list of the datasets the team is using for our analysis can be found in Appendix A.

Approach:

To determine which variables have the strongest effect on GHG emissions, we will use multi-linear and logistic regression. Our first step was preparing the data and completing exploratory data analysis (EDA) to learn more about and identify/visualize patterns in the data. During this process, we also identified what variables should be kept in the analysis or removed due to insufficient data available.

Looking forward, we anticipate building our models over the next week. We plan on training our model on a randomized 60% of our data. The remaining 40% of the data will be split equally into test and validation data sets. PCA, lasso and ridge regression will be used to reduce the number of independent variables involved in our model. We will then compare our models using fitting values such as R2 and adjusted R2. Experimenting with different regression models will allow us to determine the best fit for the data without introducing any additional biases.

Once we have a final model, we will use it to predict the GHG emissions for 2020 through 2022. If the model is moderately successful at predicting emissions for these years, we could use it to forecast further into the future. We can also use the population dataset to forecast population growth, and draw conclusions about the effects this could have on future GHG emissions.

Current Progress:

Data Preparation:

Our first step was to create a master dataset that we could use for EDA and modeling. We did this by merging the world population and GHG emissions data by country name and year as the primary key. We needed to find other independent variables outside of population that potentially contribute to the rise of GHG emissions (dependent variable) to enhance the current dataset.

Since our dataset looks at emissions on a per country level, we considered other factors at that level of granularity. For example, we theorized that the industrial production within a country could significantly affect its GHG emissions. To effectively measure production, we pulled in factors like a country's GDP, employment rate, land usage, exports of goods and services, and manufacturing volume. A full list of factors that the team identified and the data sources for them are listed in Appendix A. All of these additional factors were also combined into the master dataset by merging on the same primary keys.

Exploratory Data Analysis:

After a preliminary evaluation of our dataset, we noticed that there were several countries that had a lot of missing values. To deal with this issue, we removed the missing values and disregarded any countries with insufficient data for analysis. As a result, the universe of countries in our final dataset substantially decreased. We also noticed that the initial GHG dataset contained some region values that are not actual countries, such as the *"Rest of the World"* and *"South East Asia Region."* The histogram in Figure B shows these regions as outliers with extremely high population values. While these geographical breakdowns are interesting, they tend to be outliers and will not be incorporated into our regression analysis, further reducing the number of "countries" included from 350 to 270.

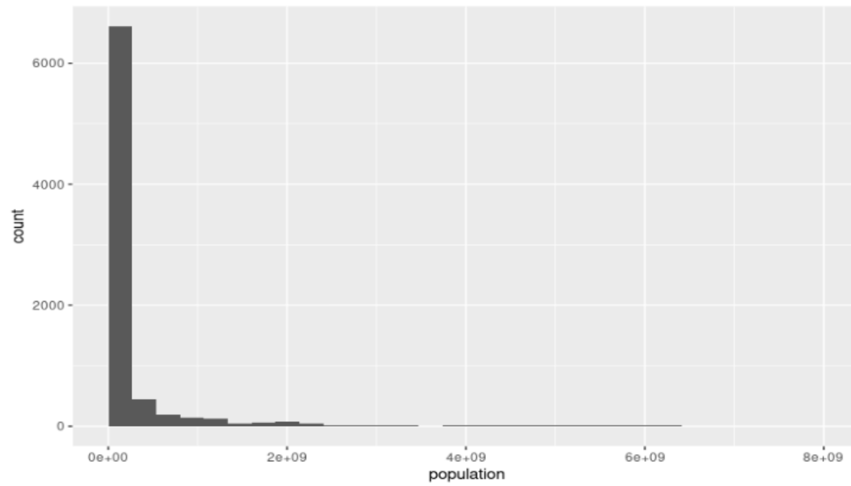

Figure B

In addition to finding countries that did not meet the team's requirements, some factors we considered only addressed a small fraction of the countries we wanted to analyze. We totaled the number of blank values per factor to identify what should be excluded due to insufficient data. The original exploratory master dataset had thirty factors (excluding country-identifying information). Out of these thirty factors, nine had data for less than half of the countries. In particular, vehicle-related data (e.g., the number of electric vehicles), yielded about 290 valid data points out of the total combined dataset value of around 7900. This is only 3% of valid data out of the whole dataset. As such, removing these factors was necessary as it did not address the global reach of our data. The table in Appendix A also shows the fraction of valid data each factor we considered had in relation to the overall working dataset.

Something interesting we noticed as we were going through our initial GHG emissions dataset was that some countries had negative emissions values. This can be better seen in Figure C on the next page. After doing further research, we realized that this dataset takes land use and forestry changes (LUCF) into account. Countries with negative GHG emissions values during certain years may have reallocated the usage of land or replenished forests in the country, which created an overall negative net emissions, as the benefit exceeded the costs. We were able to find another GHG emissions dataset that does not account for LUCF that we can also experiment with. As we are unsure about the necessity of land-use as a factor to include, we have decided to create two models with and without LUCF to determine which model has a better fit.

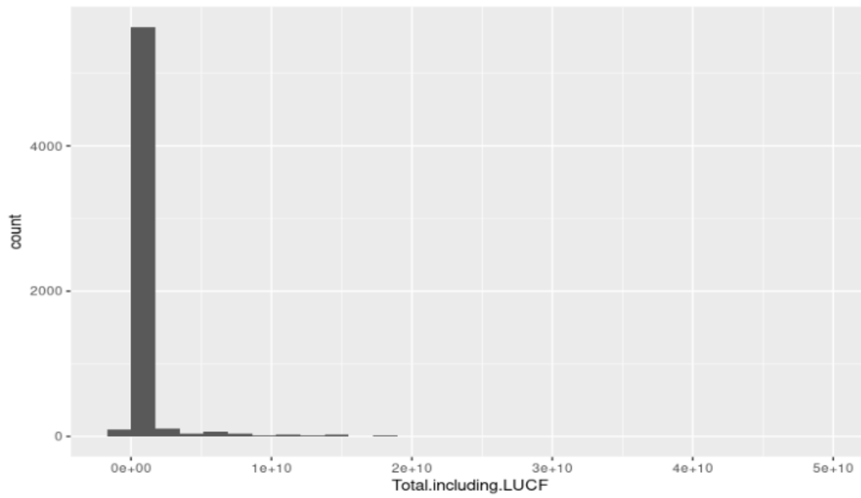

Figure C

As a baseline, the team has decided to use GHG emissions without LUCF, to keep all the data on the same scale. Since our hypothesis focused on analyzing the relationship between population growth and GHG emissions, we wanted to quickly visualize how the two factors relate to each other as they change year over year. To do so, we created a random sample of 15 countries and tracked the changes in GHG and emissions for them from 1990 to 2018. A quick animation was created in R that showed all 15 countries and where they fell on a GHG Emissions (without LUCF) vs Population graph and how their positioning changed over the years. Figure D below shows a snapshot of the animation of one such random sample of countries in 1990. On the next page, Figure E shows a snapshot of that same random sample of countries and where they fell in 2018.

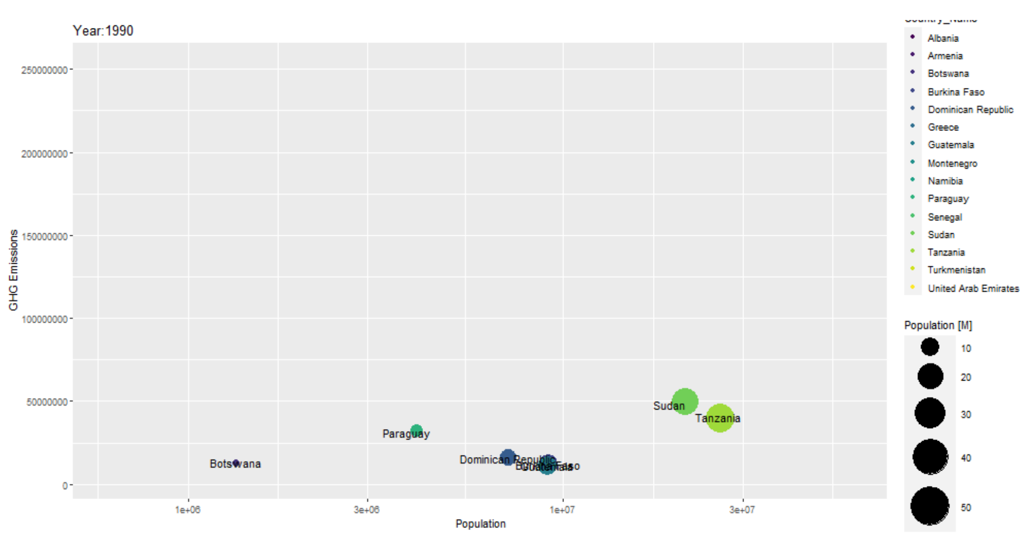

Figure D

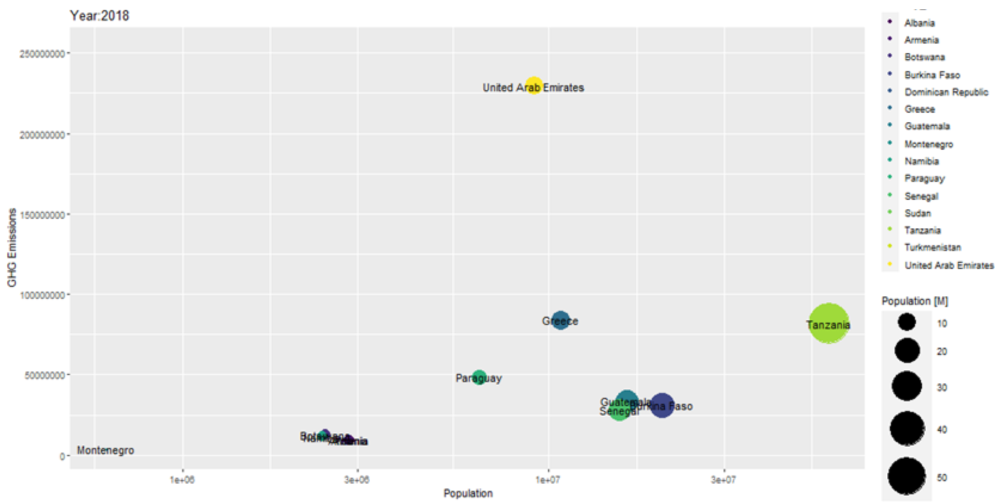

Figure E

These plots show that while an increase in population could result in an increase in GHG, there are also instances where a population of a specific country remains relatively level over the years while the GHG emissions significantly increase over a two to five year span. A good example of this is Tanzania and UAE respectively in the two figures above. This is something that the team would need to explore further.

Work That Lies Ahead:

While we have done the bulk of the data merging and pre-processing, we still have to complete the regression and assess the performance of the models. Once the regression is performed, we would need to check if the relationship is non-linear, if there are high leverage points, multicollinearity or if there is non-constant error variance that will need to be adjusted in our regression. Furthermore, we need to use certain metrics to compare the performances of our models that will help us finalize our conclusion. We have yet to determine whether our initial hypothesis was correct or incorrect.

A potential analysis our team is considering is to reduce the scope of countries being compared by shifting gears to analyze the top emitters of GHG to achieve a better comparison. This could provide a more targeted analysis for regions with more data available, but could also introduce bias with respect to identifying the significance of certain factors as it relates to emissions.

Potential Challenges:

Since our working dataset requires subjective judgment to decide what variables to include or exclude, there is a chance we could be leaving out data that would greatly affect our model. In addition, with the limited data available, we may simply be lacking the breadth of data needed to fully consider the issue from all angles. However, we believe we have found sufficient data to begin processing and analyzing, as we have incorporated up to twenty additional datasets as independent variables.

Works Cited:

Electric vehicles - worldwide. Statista. (2023, January). Retrieved March 25, 2023, from https://www.statista.com/outlook/mmo/electric-vehicles/worldwide

Tallungs, Kaj. (2022, November 16). *America's demographics over 100+ years*. Visual Capitalist. Retrieved March 25, 2023, from https://www.visualcapitalist.com/cp/animated-americas-demographics-over-100-years/

Union of International Associations (UIA). (2023). *Depopulation*. The Encyclopedia of World Problems &

Human Potential. Retrieved March 25, 2023, from http://encyclopedia.uia.org/en/problem/depopulation

United Nations (UN). (2022, April 4). UN climate report: It's 'now or never' to limit global warming to 1.5 degrees | UN news. United Nations. Retrieved March 25, 2023, from https://news.un.org/en/story/2022/04/1115452

United States Environmental Protection Agency (EPA). (2023, February 15). *Global Greenhouse Gas Emissions Data*. United States Environmental Protection Agency. Retrieved March 25, 2023, from https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data#:~:text=Global%20carbon%20emissions%20from%20fossil,increase%20from%201970%20to%202011

Quinson, Tim. (2022, December 13). *ESG Market in US Significantly Smaller Than Earlier Thought*. Bloomberg.com. Retrieved March 25, 2023, from https://news.bloomberglaw.com/esg/esg-market-in-us-significantly-smaller-than-e

Arlier-thought-1

Appendix A:

Data Sources for Factors Considered:

Variable	Data Type	Valid Values (%)	Kept or Removed from Dataset	Description	Source
Country_Code	Character	100	Kept	Unique identifier code for each country.	Our World in Data
Country_Name	Character	100	Kept	Country name.	Our World in Data
Year	Date	100	Kept	Year.	Our World in Data
GHG_with_LUCF	Numerical	74.31	Kept	Total greenhouse gas emissions from 1990-2019, including land use and forestry. Emissions are measured in carbon dioxide-equivalents.	Our World in Data https://ourworldindata.org/greenhouse-gas-emissions
Population	Numerical	85.1	Kept	World population data from 1962-2021.	WorldBank: https://data.worldbank.org/indicator/SP.POP.TOTL
Age Dependency Ratio	Numerical	82.3	Kept	Age dependency ratio of dependents - people younger than 15 or older than 64 to the working-age population.	WorldBank: https://data.worldbank.org/indicator/SP.POP.DPND
Agricultural Land	Numerical	78.8	Kept	Agricultural land (% of land area).	WorldBank: https://data.worldbank.org/indicator/AG.LND.AGRI.ZS
Air Transport (Freight)	Numerical	56.2	Removed - insufficient	Air transport (million ton-km): volume of freight, express, and diplomatic bags carried on each flight stage.	World Bank: https://data.worldbank.org/indicator/IS.AIR.GOOD.MT.K1
Air Transport (Passengers Carried)	Numerical	58.7	Removed-insufficient	Air passengers carried. Include both domestic and international aircraft passengers of air carriers registered in the country.	WorldBank: https://data.worldbank.org/indicator/IS.AIR.PSGR
Birth Rate	Numerical	79.3	Removed - highly correlated	Crude birth rate indicates the number of live births occurring during the year, per 1,000 people.	WorldBank: https://data.worldbank.org/indicator/SP.DYN.CBRT.IN

Life Expectancy at Birth	Numerical	77.6	Removed - highly correlated	Life expectancy at birth: the number of years a newborn infant would presumably live.	WorldBank: https://data.worldbank.org/indicator/SP.DYN.LE00.IN
Fertility Rate	Numerical	77.6	Kept	Fertility rate - births per woman.	WorldBank: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN
Exports of Goods and Services	Numerical	66.0	Removed - insufficient and highly correlated	Exports of goods and services represent the value of all goods and other market services provided to the rest of the world in current US$.	WorldBank: https://data.worldbank.org/indicator/NE.EXP.GNFS.CD
GDP_current	Numerical	76.1	Kept	Gross Domestic Product in current US$.	WorldBank: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
Manufacturing Value Added	Numerical	65.0	Removed - insufficient and highly correlated	Manufacturing value added in current US$.	WorldBank: https://data.worldbank.org/indicator/NV.IND.MANF.CD
Patent Applications	Numerical	36.5	Removed - insufficient	Patent applications, residents.	WorldBank: https://data.worldbank.org/indicator/IP.PAT.RESD
Natural Resources Rents	Numerical	75.4	Kept	Total natural resources rents are the sum of oil rents, natural gas rents, coal rents (hard and soft), mineral rents, and forest rents, in % of GDP.	WorldBank: https://data.worldbank.org/indicator/NY.GDP.TOTL.RT.ZS
GDP Growth %	Numerical	74.2	Removed - highly correlated	Annual percentage growth rate of GDP at market prices based on constant local currency.	WorldBank: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG
Unemployment	Numerical	43.3	Removed - insufficient	Unemployment (% of total labor force).	WorldBank: https://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS
Labor Force	Numerical	71.1	Kept	Labor force comprises people ages 15+ who supply labor for the production of goods/services. Total includes currently employed, unemployed but seeking work, and first-time job-seekers.	WorldBank: https://data.worldbank.org/indicator/SL.TLF.TOTL.IN

Energy Consumption per Capita	Numerical	79.5	Kept	Energy use per person measured in kilowatt-hours per capita. Not only electricity, but includes other areas (e.g heating, transport).	Our World in Data: https://ourworldindata.org/grapher/per-capita-energy-use?tab=table
Government Expenditure on Education	Numerical	39.8	Removed - insufficient	General government expenditure on education (current, capital, and transfers) is expressed as a percentage of GDP.	WorldBank: https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS
New Vehicle Registrations by Type	Numerical	3.7	Removed - insufficient	Based on new passenger vehicle registrations and for hybrid/battery electric vehicles.	Our World in Data: https://ourworldindata.org/transport#passenger-vehicle-registrations-by-type
Researchers in R&D	Numerical	19.5	Removed - insufficient	The number of researchers engaged in Research & Development (R&D), expressed as per million.	WorldBank: https://data.worldbank.org/indicator/SP.POP.SCIE.RD.P6
Deaths from Air Pollution	Numerical	80.1	Kept	Death rate from air pollution, 1990 to 2019: Estimated annual number of deaths attributed to air pollution per 100,000 people.	Our World in Data: https://ourworldindata.org/grapher/death-rate-from-air-pollution-per-100000?tab=table
GHG Emissions without LUCF	Numerical	74.3	Kept	Total greenhouse gas emissions from 1990-2019, excluding land use and forestry. Emissions are measured in carbon dioxide-equivalents.	Our World in Data: https://ourworldindata.org/grapher/total-ghg-emissions-excluding-lufc?tab=chart&country=~ROU