10-2-2022

Analyzing the Production and Use of Fossil Fuels: A Case for Data Mining and GIS

Alejandro Conde
University of Denver

DOI
https://doi.org/10.56902/ETDCRP.2022.1

Analyzing the Production and Use of Fossil Fuels: A Case for Data Mining and GIS

Abstract

As technology progresses and data grows both larger and more complex, techniques are being developed to keep up with the exponential growth of information. The term "data mining" is a blanket term used to describe an approach to find anomalies and correlations in a large dataset. This approach involves leveraging data mining software to manipulate and prepare data, apply statistics to quantify trends and characteristics in the data from a high level, and potentially apply advanced techniques like machine learning to identify patterns that wouldn't be apparent otherwise. In this case study, data mining aided a GIS in displaying substantial amounts of oil, gas, and coal data to make observations regarding two groups: OPEC and the largest non-OPEC fossil fuel producers from 1980 to 2020. To make more sophisticated observations and apply additional context to the trends observed in the data, populations and GDP data for the same period were included in the analysis to enrich the hydrocarbon production and consumption data and to help explain how these valuable resources are traded and consumed. This case study will apply appropriate data mining methods to feed data to a GIS and showcase trends that wouldn't be apparent otherwise and will additionally identify topics for further research.

Document Type

Masters Capstone Project

Degree Name

M.S. in Geographic Information Science

Department

Geography

Keywords

Fossil fuel, Data mining, Oil, Gas, Coal, Organization of the Petroleum Exporting Countries (OPEC)

Subject Categories

Data Science | Environmental Sciences | Geography | Oil, Gas, and Energy | Other Geography | Physical and Environmental Geography | Social and Behavioral Sciences

Analyzing the Production and Use of Fossil Fuels: A Case for Data Mining & GIS

Alejandro Conde

University of Denver University College

GIS Capstone

Dept. of Geography & the Environment

10/02/2022

Table of Contents

Abstract

As technology progresses and data grows both larger and more complex, techniques are being developed to keep up with the exponential growth of information. The term “data mining” is a blanket term used to describe an approach to find anomalies and correlations in a large dataset. This approach involves leveraging data mining software to manipulate and prepare data, apply statistics to quantify trends and characteristics in the data from a high level, and potentially apply advanced techniques like machine learning to identify patterns that wouldn’t be apparent otherwise. In this case study, data mining aided a GIS in displaying substantial amounts of oil, gas, and coal data to make observations regarding two groups: OPEC and the largest non-OPEC fossil fuel producers from 1980 to 2020. To make more sophisticated observations and apply additional context to the trends observed in the data, populations and GDP data for the same period were included in the analysis to enrich the hydrocarbon production and consumption data and to help explain how these valuable resources are traded and consumed. This case study will apply appropriate data mining methods to feed data to a GIS and showcase trends that wouldn’t be apparent otherwise and will additionally identify topics for further research.

Relevant Definitions

Algorithm: A set of rules followed in a calculation or problem-solving operations.

Artificial Intelligence (AI): the development of computer systems to perform tasks that would normally require a human.

Data Mining: the process of applying statistics and other processes to discover patterns in a database.

Database: a structured set of data, maintained within a computer.

Decision Tree: a decision support model that highlights the decisions in an algorithm, forming a tree-like display.

Esri ArcGIS: Proprietary GIS mapping software utilized in this study to leverage mined data for display.

Fossil Fuel: a naturally occurring fuel, created by geologic processes from the remains of once living organisms.

Fuzzy Clustering: a form of clustering where each point can belong to more than one cluster. One kind of Data analysis that is intended to identify clusters.

Genetic Algorithm: a model for solving problems based upon natural selection (the process behind natural evolution).

Graph: a diagram used to display the relationship between two variables on a pair of axes.

GDP: The monetary measurement representing the total market value of all goods and services in a country.

Matrices: environments where something develops.

Microsoft Access: The proprietary database software utilized in this study to both contain and provide structure to the study data.

Microsoft Excel: The proprietary spreadsheet software utilized in this study to manipulate and graph the study data.

Neural Network: a computer system modelled after the human brain; a method of AI.

OPEC: The Organization of the Petroleum Exporting Countries.

Orange software: The open-sourced Data mining software applied in this study.

Spatial Data: Data with a spatial component used to visually represent an object in 3D space.

Tensors: a mathematical object that represents an array of components; analogous to a vector.

Vector Data: a quantity displaying magnitude and direction of an object in space relative to another.

Widgets: a simple application or component of a user interface that enables a user to perform a function.

Introduction

Data mining is the process of developing patterns and making observations from large data sets by applying statistics and other data processing functions to a dataset; this data is usually contained within a database. This analytical method is not new, as there have been professional data mining companies formed as far back as the 1990's to help develop market research from other companies' data warehouses. These warehouses were large databases containing transactional data for millions of customers and proved too large for traditional statistical analysis (Clifton, 2022). One of the earliest applications in this industry was credit card fraud detection (Bhattacharyya et al., 2011). By detecting patterns in customers purchasing habits, anomalies become rapidly apparent and were flagged. Through this revolution in the industry, data mining aided in larger scale anomaly detection by grouping customers together and evaluating their behavior to fine-tune the anomaly detection algorithm, improve false positive detection, and reduce declined purchases. The application of robust computational power enriched the existing field of fraud detection and highlights what an optimized data mining application can do. This case study is meant as an exploratory effort to delve into more sophisticated methods of "cleaning" data by means of basic algorithm development and producing detailed graphics for examination. As for the data being employed, fossil fuel use data were chosen because of its sufficiently large size, temporal span, and the progressively stringent regulations around fossil fuel use. This exercise combines GIS and data mining practices to showcase how the combination of these disciplines can be leveraged to answer the question; how do the OPEC countries and non-OPEC countries compare regarding fossil fuel trends?

Literature Review

The text *Data mining: Practical Machine Learning Tools and Techniques* provided a full summary of the topic of data mining (Witten et al., 2017). Topics ranging from general concepts and algorithms to decision trees and data transformations were discussed at length. Other concepts (such as deep learning) were omitted from the initial literature review, as it was considered a bit too advanced for the scope of this study. This text served as an exceptional foundation because most journals found mostly provided a brief overview and did not elaborate on the subject on such a low, detailed level.

In *Comprehensive study of applications of spatio-temporal data mining in GIS*, the concept of "mining" data by processing and applying statistical based analytical methods to produce graphics that display sophisticated trends based on the refined data was explored (Malik & Nandal, 2018). This example of data mining included the integration of refined temporal data into a GIS, which relates directly to this study, and establishes a president for this analytical process.

In *Data mining for Credit Card Fraud*, another example of data mining is present, but this study focuses the previously mentioned subject of data mining with respect to credit card fraud (Bhattacharyya et al., 2011). The key takeaway from this study is that data mining is an excellent method to extract information from large datasets and can be applied to industries such as banking to identify patterns and detect instances of fraud. In this example, a regression analysis and two mining approaches (referred to as random forests and vector machines) were employed to better detect credit card fraud by identifying transactions that were atypical of the associated customer.

The main dataset contributing to this analysis was a breakdown of fossil fuel production (oil, gas, and coal) by each country, for each year from 1980-2020. To better understand how oil production affects trade, the paper by Akman & Bozkurt, (2016) was incorporated into this study to shed light on the subject. In this study, it was observed that the increase of the price of oil lowers the exports of oil-exporting countries, which is contrary to their assumption of an increase in exports during times of high demand. The supposition that there may be several factors driving this inverted relationship was put forth: the increasing import prices for commodities due to the surcharged oil prices, higher costs of imported goods discourage oil-exporting countries from purchasing foreign goods, and that the motivation for increasing savings for new investments in different sectors to increase domestic production all contribute to why a country may choose to lower oil exports. A final observation pointed out this unexpected effect is present for countries such as the U.S. and Korea but is not present in countries such as China, India, and France. This paper outlines trends in oil producing countries and explains that high oil prices do not enable a strong trade balance in the long-term. This insight into the complex relationships between major oil producers is important for a thorough understanding of the information present in the data being processed by this study and aided in the interpretation of these massive datasets.

Casey & Galor (2016) provide evidence for lower fertility rates being correlated to increased income per capita and lower emissions. This study was included in the review since population is a key factor in the production and consumption of hydrocarbons for energy use. Since population numbers over time are included in the statistical analysis, a link between the population and the trends needed to be established, and this journal provided a framework for establishing this connection.

In *Application of Data Mining to Production Operation and Control System in Oil Field* the concept of data mining with respect to petroleum production was introduced (albeit at the much lower level of oil production at the field level, not the larger international level of production). This paper provides basic theories on the application of data mining such as the neural networks, the decision tree method, genetic algorithm methods, fuzzy theory, and others. It shows the principal of the use of data mining with respect to oil and shows that this methodology has been applied to this field in the past (Tong Wang et al., 2012).

Another example of data mining applied in the oil field is in the paper *Evaluating Geothermal Energy Production from Suspended Oil and Gas Wells by using Data Mining* (Moussa & Dehghanpour, 2022). This paper highlighted the successful use of data mining to solve a real-world petroleum related problem by identifying the trends in data at a high level and sought to answer the question, “Can suspended oil and gas wells be retrofitted to generate geothermal power”. A supervised fuzzy clustering algorithm was used to estimate the geothermal potential from a dataset consisting of wells. Impressively, the results not only showed that this was possible, but it showed which parameters were optimal for retrofitting. Specifically, the optimal kind of borehole was identified by sifting through data related to the diameter of the pipe, the tolerance of depth and other factors, all within the tabular data. The data revealed both the possibility that existing holes can be used for geothermal purposes and the identification of the optimal dimensions of pipe to use. This work provides insight into the preferable characteristics of borehole and might enable a company to provide an immense cost savings for nearby residents by retrofitting the most optimal wells and tapping into geothermal energy for their home. If this venture were pursued further, a new industry of recycling old wells might become an economically and environmentally sound way to utilize existing drilling byproducts as source of power.

The study “Data mining, GIS and multicriteria analysis in a comprehensive method for bicycle network planning and design” displays how GIS and data mining can be utilized to solve the dilemma of identifying bike paths on a very large scale: in the cities of Brazil. While the task being solved (that of planning bicycle networks) does not relate to oil production or consumption, the methods being implemented do. This study provides an excellent example as to how a data mining analysis should be structured, and this study does a great job of determining a methodology to design and compare cycling networks by the characteristics of the data. The three-step process included choosing a dependent variable and employing decision tree data mining techniques to identify potential real users, then further distinguishing their patterns and probability of cycling from massive spatial and tabular datasets. Next, by relating the data from the first step to a disaggregated network dataset, the compatible road sections with the greatest possible number of overlapping routes within the required criteria were queried for further use. Finally, a comparison between the obtained networks and the roads currently maintained by the municipal government was performed to verify if the existing configuration could be optimized. The result was a more precise method then was previously utilized by the Brazilian government and provided a more sophisticated and detailed alternative for the local government to identify local roads to build upon.

Spatio-Temporal Data Mining: A Survey of Problems and Methods is another example of data mining applied to spatial data over time (Atluri et al., 2019). This article includes a broad survey in many areas of expertise. This survey includes data which pertains to current and historic oceanic and atmospheric conditions in the climate science field. With respect to neuroscience, data was mined to identify the governing principals and determine potential disruptions to normal conditions with respect to Functional Magnetic Resonance Imaging (fMRI), Electroencephalogram (EEG), and Magnetoencephalography (MEG) data. Next, the recording of health care records in epidemiology/healthcare were connected to spatial data and mined to track the spatial trends in the spread of disease. For Heliophysics, the events that occur on the sun and their effects on the solar system were mined to study the patterns in these events. This study goes over each area of study and provides a framework for analyzing each data type found in the study. The methodologies put forth here provided inspiration in the analytical phase of this case study, as this article provided the most wide-ranging application of data mining methodologies observed.

Zheng (2015) speaks on data mining with respect to trajectory (vector data focused upon the movements of objects in space). This study conducted a survey on the research into the topic itself and explains methods on converting this spatial data into other data formats (like graphs, matrices, and tensors). Carefully manipulating spatial data before processing is critical, and this study suggests several approaches to convert the data before data mining and machine learning techniques can step in and further refine these data for final export. It drives the point home that data mining requires proper structuring and processing before analysis can begin. To put things concisely, set-up is key.

Data Resources

In this case study, hydrocarbon data, worldwide gross domestic product (GDP), emissions by country and population data were sourced from their publicly available databases and were related in local database software. These interconnected data form the foundation from which the data mining software builds upon and identify the trends autonomously. These data were gathered from separate organizations but are all connected as they share two dimensions of commonality: geography and time. Each fossil fuel dataset represents both the production or consumption of coal by ton, or gas and oil by cubic meters for each fuel producing country from 1980 to 2020. The Gross Domestic Product (GDP) data represents the monetary value of all final goods and services in billions by every country from 1980 to 2021. The carbon dioxide emissions data represents the total amount of emissions from all sources by metric tons and spanned from 1980 to 2021. Finally, the world population data extends from 1973 to 2021, but was constrained from 1980 to 2021 to be comparable to the other data. OPEC membership status was the final dataset, and was created for this analysis. This dataset connected directly to the OPEC dataset in Microsoft access and constrained the data for these countries by their years of membership and quickly produced the subsets of the many datasets by this active membership attribute. To compare relevant fossil fuel producing countries to this group, the thirteen largest non-OPEC fossil fuel producers were selected and queried from each dataset. This step of constraining the datasets was critical in reducing “noise” in the data, as each dataset contained thousands of entries associated with every country that produces coal, oil, and gas over a forty-year period.

Software

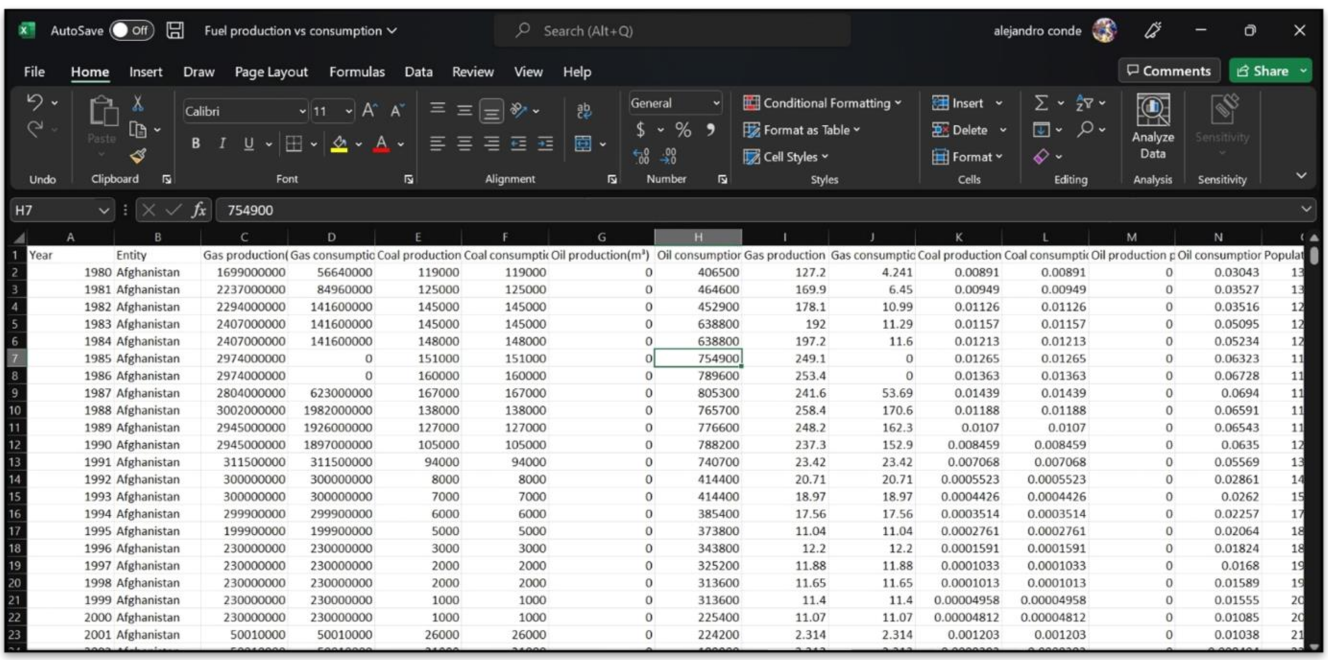
Data were collected and prepped with Microsoft's Access database software prior to the data-mining process. An open-source data mining program called "Orange" was chosen to spearhead this in-depth analysis of the data. This program was chosen because it is open source and free to use, is analytically powerful, and can produce data compatible with a GIS. Orange proved effective in producing a simple, widget-based algorithm with the capability of refining thousands of data records quickly and the process could be repeated by merely importing each dataset and exporting the results. Additionally, Orange proved to be a very feature rich platform that allowed for various statistical applications: decision trees, algorithm structuring, machine learning applications, regression analysis, neural networks, and a plethora of other data manipulation options. However, not all statistical applications are appropriate for every study. To start this study, Orange widgets (as showcased in the flowchart) were employed leveraged to manipulate each dataset and produce graphics that showcased a variety of summarized trends in the data. To further enrich the analysis, a series of line chart functions were added to graphically display the data over time, which allows for an additional visualization of the countries with respect to one another over the entire temporal span of the data. This was where GIS's mapping capability excels: leveraging spatially attribute data to generate sophisticated maps and graphics for further review.

To further display temporal and statistical trends as they evolve, an ESRI-based Geographic information system was created with the sole purpose of leveraging the previously mentioned data outputs from the data mining software, connect them to their corresponding countries using their geographic component in the data and create maps that showcase their trends in a four framed time series map for each production characteristic, GDP, and population. These maps displayed bars to indicate the magnitude of both the mean and standard deviation of the characteristic being modeled in that decade. This sophisticated cartographic prowess eclipsed anything the non-GIS software could produce and proved to further aid in the conveying of massive amounts of information.

These two software platforms were applied to the pre-processed data in Microsoft Access and formed a cohesive analytical ecosystem that allowed for powerful analysis to be made in a manner far faster than what is possible processing the data manually. ArcGIS could have handled much of the data manipulation and statistical analysis. However, the methodology must be repeated by hand and would leave more room for human error. This is where Orange was employed to process this data and remove the previously mentioned human error from the cleaning process. After delving into the various widgets and functions, it was noted that Orange does have a mapping extension with associated widgets. They're inferior when compared to the robust mapping abilities of ArcGIS, so they were not considered for the algorithm. Each software possessed corresponding strengths and made up for the other software's shortcomings.

Methods

To describe data mining with respect to this study; data mining is the application of statistics to large datasets managed in a database, with the express purpose of identifying trends and drawing conclusions from the connected data. For this case study, large datasets were gathered from reputable sources with the express purpose of connecting to and enriching one another. Four independent datasets were processed through this study, with production and consumption data taken from one and converted into six related datasets for coal, oil and gas production and consumption (*U.S. Energy Information Administration - EIA - independent statistics and analysis,* 2022). Emissions data and world population were obtained from the same site, and global Gross domestic product (GDP) by country was taken from the world



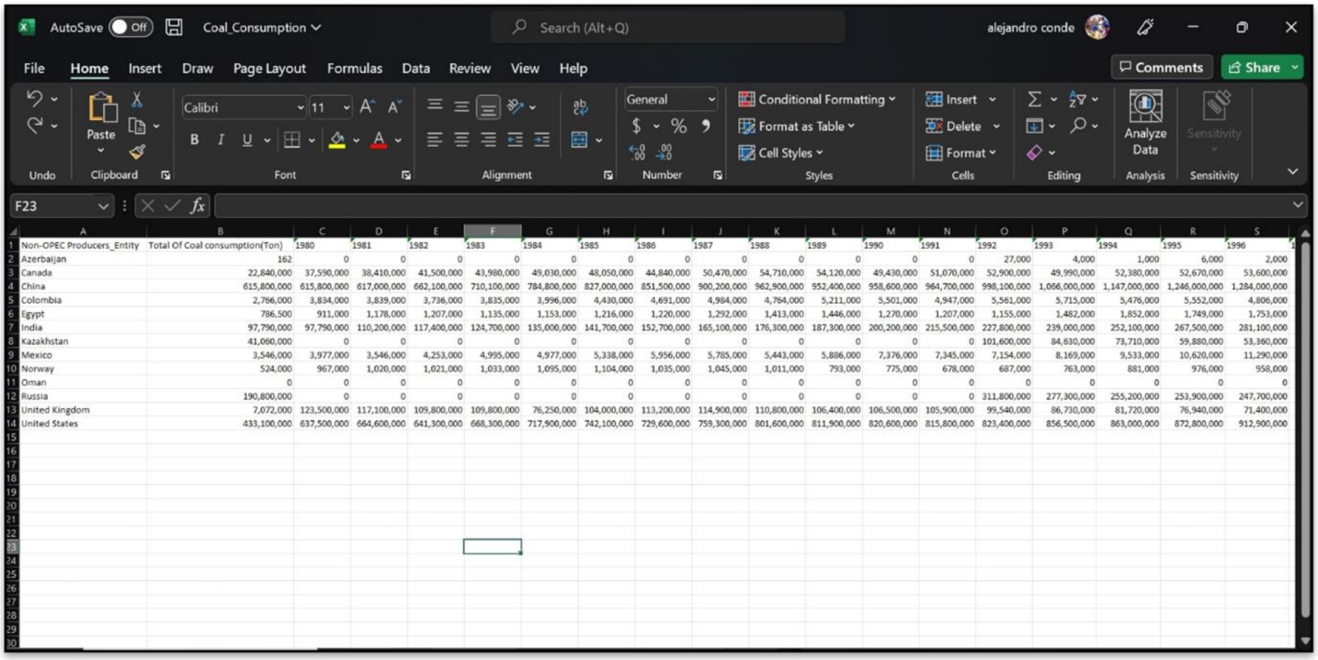
Year	Entity	Gas production(	Gas consumptic	Coal production	Coal consumpti	Oil production(m³)	Oil consumptior	Gas production	Gas consumptic	Coal production	Coal consumpti	Oil production p	Oil consumptior	Populat
1980	Afghanistan	1699000000	56640000	119000	119000	0	406500	127.2	4.241	0.00891	0.00891	0	0.03043	13
1981	Afghanistan	2237000000	84960000	125000	125000	0	464600	169.9	6.45	0.00949	0.00949	0	0.03527	13
1982	Afghanistan	2294000000	141600000	145000	145000	0	452900	178.1	10.99	0.01126	0.01126	0	0.03516	12
1983	Afghanistan	2407000000	141600000	145000	145000	0	638800	192	11.29	0.01157	0.01157	0	0.05095	12
1984	Afghanistan	2407000000	141600000	148000	148000	0	638800	197.2	11.6	0.01213	0.01213	0	0.05234	12
1985	Afghanistan	2974000000	0	151000	151000	0	754900	249.1	0	0.01265	0.01265	0	0.06323	11
1986	Afghanistan	2974000000	0	160000	160000	0	789600	253.4	0	0.01363	0.01363	0	0.06728	11
1987	Afghanistan	2804000000	623000000	167000	167000	0	805300	241.6	53.69	0.01439	0.01439	0	0.0694	11
1988	Afghanistan	3002000000	1982000000	138000	138000	0	765700	258.4	170.6	0.01188	0.01188	0	0.06591	11
1989	Afghanistan	2945000000	1926000000	127000	127000	0	776600	248.2	162.3	0.0107	0.0107	0	0.06543	11
1990	Afghanistan	2945000000	1897000000	105000	105000	0	788200	237.3	152.9	0.008459	0.008459	0	0.0635	12
1991	Afghanistan	311500000	311500000	94000	94000	0	740700	23.42	23.42	0.007068	0.007068	0	0.05569	13
1992	Afghanistan	300000000	300000000	8000	8000	0	414400	20.71	20.71	0.0005523	0.0005523	0	0.02861	14
1993	Afghanistan	300000000	300000000	7000	7000	0	414400	18.97	18.97	0.0004426	0.0004426	0	0.0262	15
1994	Afghanistan	299900000	299900000	6000	6000	0	385400	17.56	17.56	0.0003514	0.0003514	0	0.02257	17
1995	Afghanistan	199900000	199900000	5000	5000	0	373800	11.04	11.04	0.0002761	0.0002761	0	0.02064	18
1996	Afghanistan	230000000	230000000	3000	3000	0	343800	12.2	12.2	0.0001591	0.0001591	0	0.01824	18
1997	Afghanistan	230000000	230000000	2000	2000	0	325200	11.88	11.88	0.0001033	0.0001033	0	0.0168	19
1998	Afghanistan	230000000	230000000	2000	2000	0	313600	11.65	11.65	0.0001013	0.0001013	0	0.01589	19
1999	Afghanistan	230000000	230000000	1000	1000	0	313600	11.4	11.4	0.00004958	0.00004958	0	0.01555	20
2000	Afghanistan	230000000	230000000	1000	1000	0	225400	11.07	11.07	0.00004812	0.00004812	0	0.01085	20
2001	Afghanistan	50010000	50010000	26000	26000	0	224200	2.314	2.314	0.001203	0.001203	0	0.01038	21

Figure 1. Raw Data Example

bank's website (World Bank 2022). These datasets were downloaded in a comma separated format (Excel files with a ".csv" extension) compatible with both ESRI's ArcGIS software and Orange.

With these datasets secured, the first round of prep work was performed using Microsoft Access. The production and consumption data were located within the same dataset and needed to be queried and pivoted (rearranged) to have each dataset structured to have the countries in the rows and the columns to showcase years. These were further constrained into two data sets each: one by OPEC members and the other by non-OPEC members. These parsed datasets supported an initial round of graphic production that showcase the trends in hydrocarbon production and consumption temporally as series of bar graphs. A similar approach to emissions, world population and GDP were leveraged to produce similarly structured graphics for visual comparison.

As previously mentioned, the open-source data mining software was leveraged here because it is both intuitive to learn and is extremely powerful in terms of analytical possibilities. After a few tutorials, a basic data processing structure was formed. At first, the data was mostly reshaped, queried and employed to create more detailed graphs. Specifically, line graphs to better display the trends over time, and produce correlograms to visually showcase the randomness in the dataset or correlation between all the data. The data manipulation process required intermediate steps like transposing variables and changing domains (such as coding OPEC & Non-OPEC text to “1” and “0” respectively and changing these from text to numeric values) in order to be ingested by other widgets. The temporal based widgets were critical to transforming the time-based data to a form that could display the data as a correlogram and line charts.



Non-OPEC Producers_Entity	Total Of Coal consumption(Ton)	1980	1981	1982	1983	1984	1985	1986	1987
Azerbaijan	162	0	0	0	0	0	0	0	0
Canada	22,840,000	37,590,000	38,410,000	41,500,000	43,980,000	49,030,000	48,050,000	44,840,000	50,470,000
China	615,800,000	615,800,000	617,000,000	662,100,000	710,100,000	784,800,000	827,000,000	851,500,000	900,200,000
Colombia	2,766,000	3,834,000	3,839,000	3,736,000	3,835,000	3,996,000	4,430,000	4,691,000	4,984,000
Egypt	786,500	911,000	1,178,000	1,207,000	1,135,000	1,153,000	1,216,000	1,220,000	1,292,000
India	97,790,000	97,790,000	110,200,000	117,400,000	124,700,000	135,000,000	141,700,000	152,700,000	165,100,000
Kazakhstan	41,060,000	0	0	0	0	0	0	0	0
Mexico	3,546,000	3,977,000	3,546,000	4,253,000	4,995,000	4,977,000	5,338,000	5,956,000	5,785,000
Norway	524,000	967,000	1,020,000	1,021,000	1,033,000	1,095,000	1,104,000	1,035,000	1,045,000
Oman	0	0	0	0	0	0	0	0	0
Russia	190,800,000	0	0	0	0	0	0	0	0
United Kingdom	7,072,000	123,500,000	117,100,000	109,800,000	109,800,000	76,250,000	104,000,000	113,200,000	114,900,000
United States	433,100,000	637,500,000	664,600,000	641,300,000	668,300,000	717,900,000	742,100,000	729,600,000	759,300,000

Non-OPEC Producers_Entity	1988	1989	1990	1991	1992	1993	1994	1995	1996
Azerbaijan	0	0	0	0	27,000	4,000	1,000	6,000	2,000
Canada	54,710,000	54,120,000	49,430,000	51,070,000	52,900,000	49,990,000	52,380,000	52,670,000	53,600,000
China	962,900,000	952,400,000	958,600,000	964,700,000	998,100,000	1,066,000,000	1,147,000,000	1,246,000,000	1,284,000,000
Colombia	4,764,000	5,211,000	5,501,000	4,947,000	5,561,000	5,715,000	5,476,000	5,552,000	4,806,000
Egypt	1,413,000	1,446,000	1,270,000	1,207,000	1,155,000	1,482,000	1,852,000	1,749,000	1,753,000
India	176,300,000	187,300,000	200,200,000	215,500,000	227,800,000	239,000,000	252,100,000	267,500,000	281,100,000
Kazakhstan	0	0	0	0	101,600,000	84,630,000	73,710,000	59,880,000	53,360,000
Mexico	5,443,000	5,886,000	7,376,000	7,345,000	7,154,000	8,169,000	9,533,000	10,620,000	11,290,000
Norway	1,011,000	793,000	775,000	678,000	687,000	763,000	881,000	976,000	958,000
Oman	0	0	0	0	0	0	0	0	0
Russia	0	0	0	0	311,800,000	277,300,000	255,200,000	253,900,000	247,700,000
United Kingdom	110,800,000	106,400,000	106,500,000	105,900,000	99,540,000	86,730,000	81,720,000	76,940,000	71,400,000
United States	801,600,000	811,900,000	820,600,000	815,800,000	823,400,000	856,500,000	863,000,000	872,800,000	912,900,000

Figure 2. "Pivoted" Data Example

After this initial model creation, additional branches to this mining tree were added to enrich the functionality of the model. A heat map widget was added to the original dataset, as it produces a compact graphic that presents the magnitude of the greater values by the saturation of red vs. the lack of value being represented by white. While this is simplistic, the amount of data packed into this simple visual representation surpasses the bar or line charts in terms of the ability to convey information quickly, and without the need to zoom in or scrutinize the data at a low level.

Next, a basic linear regression was added to the model to attempt to correlate all the data to OPEC membership. The previously mentioned coding and domain change for OPEC status was what enabled the program to format the data for use by the linear regression widget. While this is simple method of performing a linear regression, this method provides repeatability, as well as robust functionality (with the widget allowing the user to adjust the alpha value, regression strength and type of regression). This allowed for the export of a list of regression coefficients for each variable (one for each year and each country).

With the data mining model set up to process the first dataset, the remaining datasets were also processed, and exports were saved by switching out each CSV Excel sheet and allowing the program to process the sheet and save the outputs. After this was performed, integration into a GIS was the next step. ArcGIS was chosen as the platform to develop this GIS as the additional statistical capabilities would be critical to gain additional spatial insight from the data as it can be connected to geography. First, a map for the OPEC and Non-OPEC countries was created as a key from which the other maps would be compared to identify the geographic orientation of the countries. The data were "rolled-up" or aggregated to produce means and standard deviation for each decade (1980's, 1990's, 2000's and 2010's). These data were leveraged to produce a quadruple time-series map for each decade, recurring for each data type. This produced series of maps for each fossil fuel (production and consumption), Gross Domestic Product, population, and emissions. It visually displays the data's mean value for each OPEC and Non-OPEC country in their respective group, as well as the standard deviation for the data being modeled in these countries. The intent here was to use basic statistics to highlight the average values for that decade as well as display the spread in the data with the intent of comparing the variability of the decade's data on the same map. GIS allowed for the user to convey a massive amount of data here: geographic spread of the data, color coding the OPEC status of those countries, as well as average values for the decade, and how wildly these values varied over each decade all in one cohesive graphic. This level of detail is not possible with the data mining's mapping widgets, which can only display geographic location of data at best.

Flowchart

The templates created in the data mining software produce nine series of charts that visually highlight the trends as they exist in the data. The charts will be discussed in the following order: the initial charts representing overall population & emissions trends between OPEC & non-OPEC members produced by Excel, the production & consumption trends between OPEC & Non-OPEC members produced in Orange data mining software. Last, the maps produced by the addition of GIS into the workflow are discussed. The algorithm developed in Orange provides a convenient visualization of the workflow.

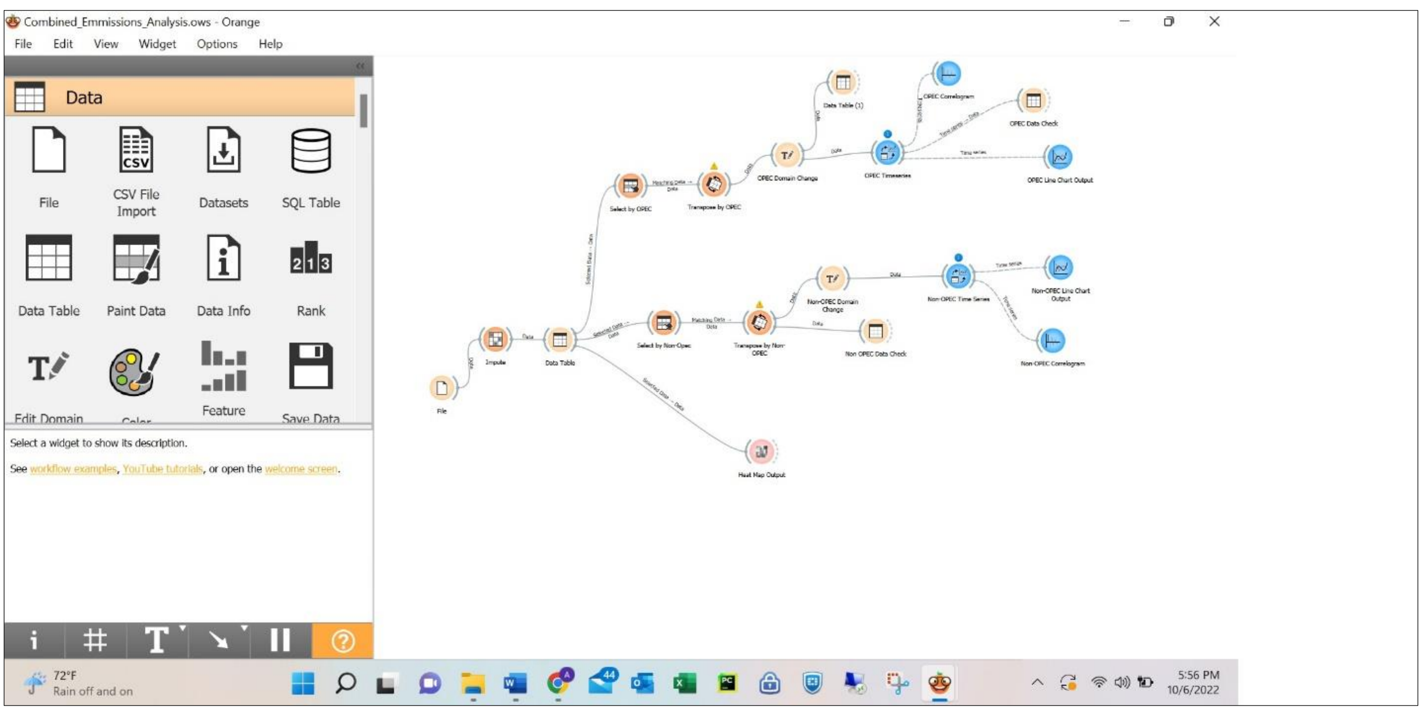


Figure 3. Orange data mining software example

Results

Note: the data being analyzed were exhibited at a high-level (being at the international level), as a more low-level approach (country to country) was considered out of the scope of this study but would be an area for additional research. The intent was to compare the data relative to each country and separately analyze the two groups (OPEC & Non-OPEC producers) over a forty-year period, but these groups are a small subsection of the overall population, and the possible assemblages are limited only by the creativity of the analysts examining the data. Note that, though the data dips towards the end of the absolute value charts (which can be explained by the program showcasing that there are no data for the next year, and these values are effectively "0"), the relative charts do not share this characteristic and displays the true end to each series and does not search for data in the next year. Effectively, when there is a dip at the end of the relative chart, this is significant. Also, the data mining software does not provide options to edit the scale in the exported graphics at the time of this publication, so for the time being, these charts are mostly used for relative analysis. If quantitative analysis is required, other data mining software could be leveraged, or the data at a specific time can be looked up. Also, a Linear regression analysis was processed for each data series but was not included here because it could not be verified with a proper third-party statistics software to be sure. This would be another area of additional research that would enrich this study and provide more weight behind the observations.

Overall population & emissions

Emissions data were included in this study, as the byproducts of the combustion of fossil fuels directly correlates the real usage for each country. The data displays how emissions have increased as time progresses across this forty-year period. For the OPEC members, the data was queried to represent only values for when the countries were active members of OPEC. For example, countries like Indonesia are not included for the entire forty-year period, despite having produced fossil fuels for the entire span of the study. Indonesia was a part of OPEC in 1980, left OPEC in 2009, rejoined in 2016, and left the same year. All related data will reflect these variations in OPEC membership by removing data for years of non-OPEC membership.

In this graph, the trends are positively increasing as time progresses. Likely, if for nothing else, this is due to population increasing and globalization leading to more travel and trade, so this does not necessarily translate to a lack of awareness, concern, or action by either group. However, in terms of overall numbers the non-OPEC numbers are far higher compared to their OPEC counterparts. These numbers are being skewed by two of the largest consumers: China and India. If not for these two, the remaining large non-OPEC countries would be better poised

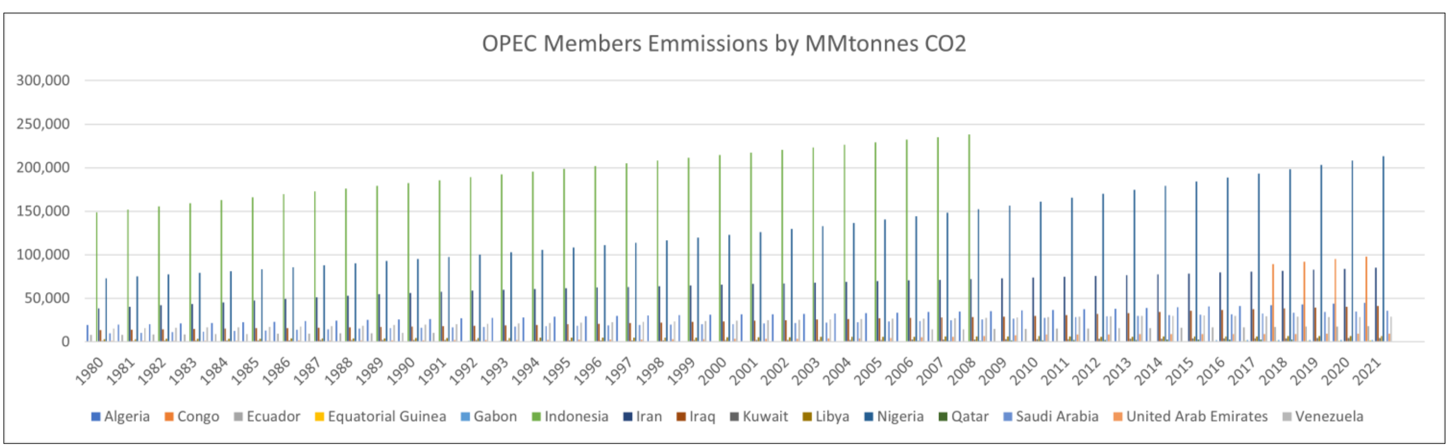


Figure 4. OPEC emissions

to directly compare to the OPEC countries. The United States are consistently higher in their CO_2 emissions compare to any OPEC members. The next largest non-OPEC emitter (Russia) lies well below the 200,000 MMtonne mark (which is within the range for direct comparison to OPEC emitters). This preliminary comparison of the data suggests there is a larger spread in the non-OPEC data, which will be observed with the statistical analysis in the GIS.

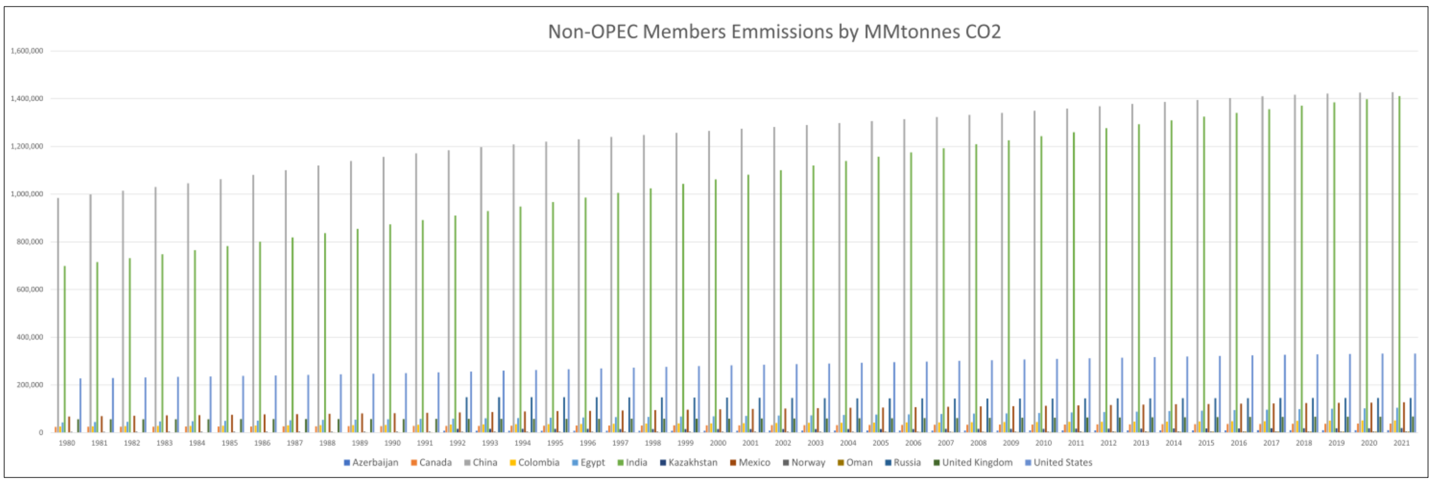


Figure 5. non-OPEC members emissions

In the data mining software, there are many widgets which provide options for manipulating and processing outputs that allow for meaningful interpretations to be derived from a dataset. For this study, the heat map was chosen as it is the most powerful graphic produced in the entirety of the exports. The heat map presented in this study is intended to highlight magnitude rather than placing the focus upon exact numbers. This graphic shows the relative relationship of all the data in a series and aggregates these data by country and OPEC status. This information is conveyed by increasing the saturation and intensity of color and includes a scale of the numeric magnitude at the top of the chart. Note: detail is lost when comparing similar countries with respect to this scale. Instead, this chart is intended to (and excels at) highlighting the relative information between both these countries and their OPEC status. The graphic produced is one of the most powerful means of identifying the trends in the data without the clutter associated with bar and line graphs.

This heat map displays the emissions for OPEC and Non-OPEC countries. It does an excellent job of illustrating just how great the difference is between China, India, and the next greatest polluter: the United States. The OPEC countries certainly produce far less emissions when compared to the non-OPEC polluters.

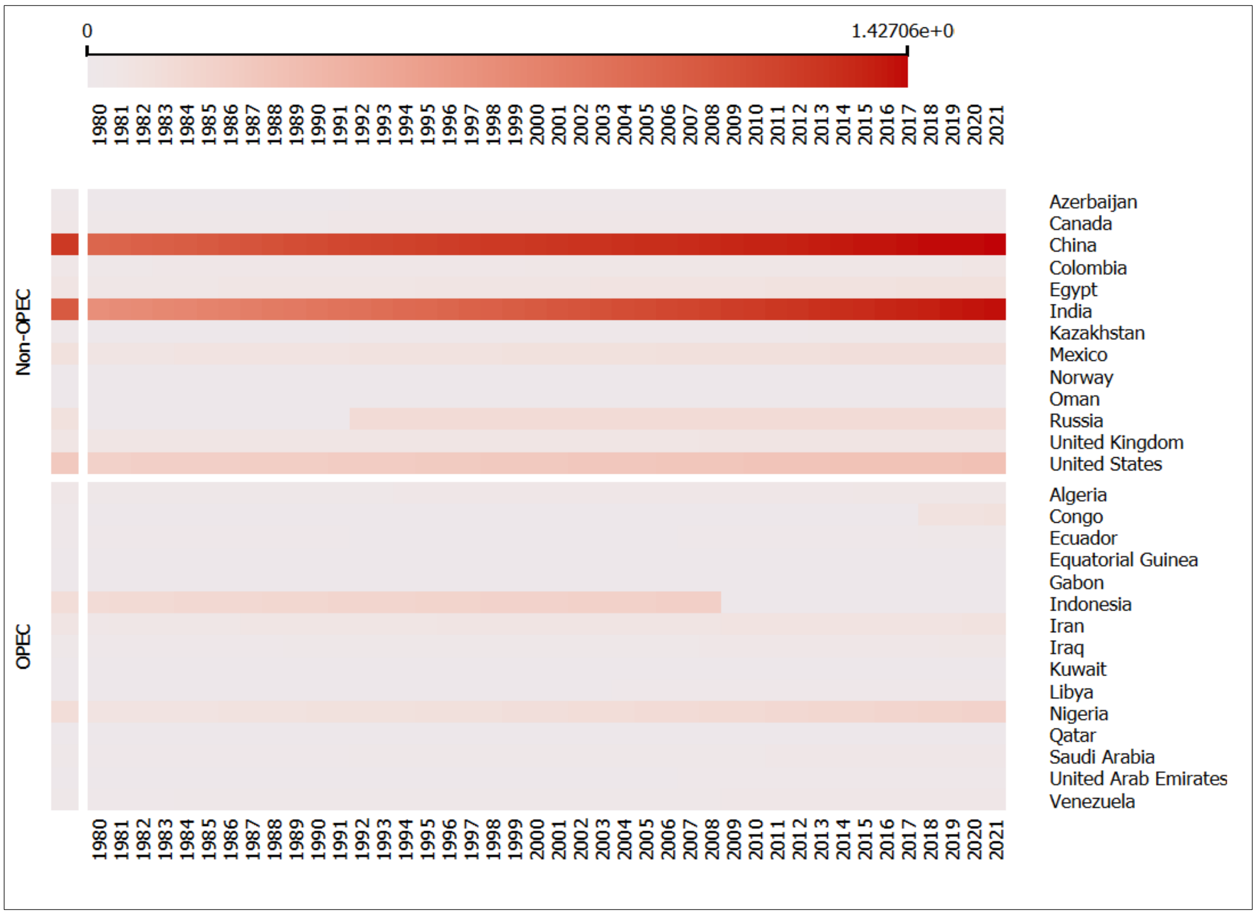


Figure 6. Emissions Heat Map

Global Population & GDP

The population data were queried from the main dataset for the fifteen OPEC nations and the thirteen non-OPEC nations. The presupposition for this dataset was a relationship to hydrocarbon usage, with larger populations likely translating to more emissions from oil, gas & coal consumption for energy, propulsion, shipping, etc. For OPEC trends, Indonesia is consistently the country with the highest population. Nigeria is the relatively close second, with the country exceeding 150 million people in 2009. The final larger populated country is Iran; however, the remaining OPEC countries never exceed fifty million people in 2021.

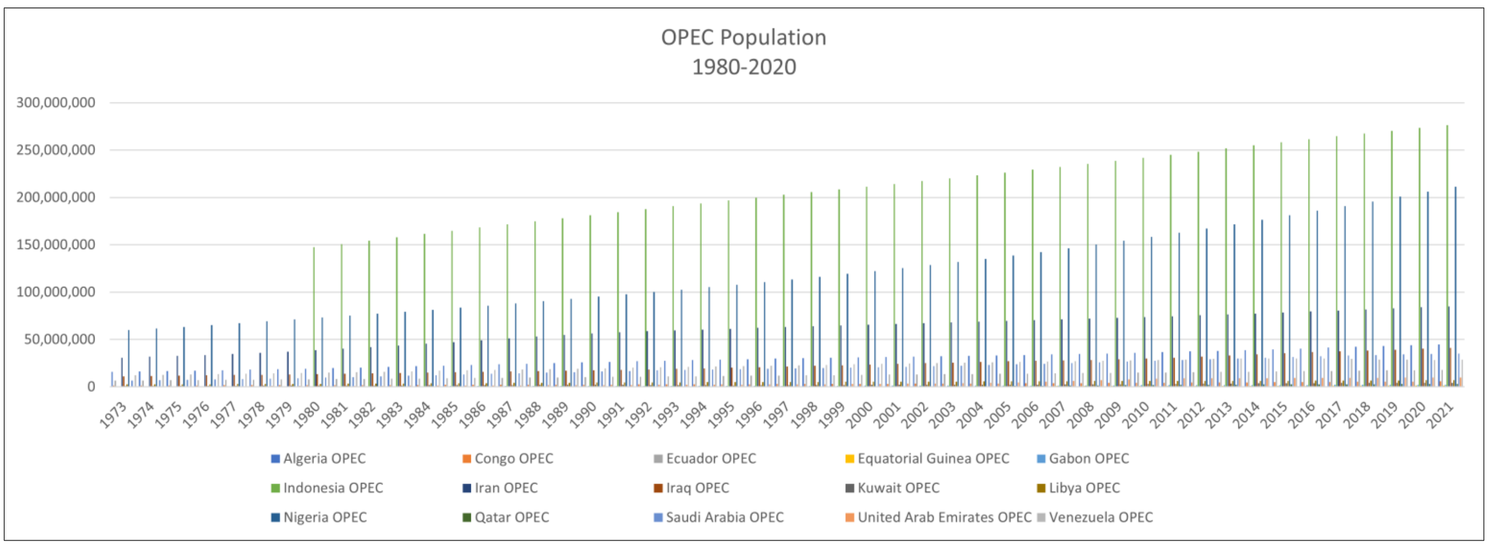


Figure 7. OPEC Population

As for the Non-OPEC countries, the trends are similar to the emissions data: India and China are by a wide margin the largest populated countries in the non-OPEC members, and also have the highest populations of people in the study overall. By identifying connections from disconnected emissions and population data, connection between people and emissions was confirmed, but does not correlate precisely. An additional correlation test could be performed for each country to measure the extent to which these population data correlate to emission. As for the OPEC and Non-OPEC members at the extremes, the non-OPEC members are far worse in terms of hydrocarbon usage, but a finer analysis is needed to remove the outliers and produce a more equivalent comparison.

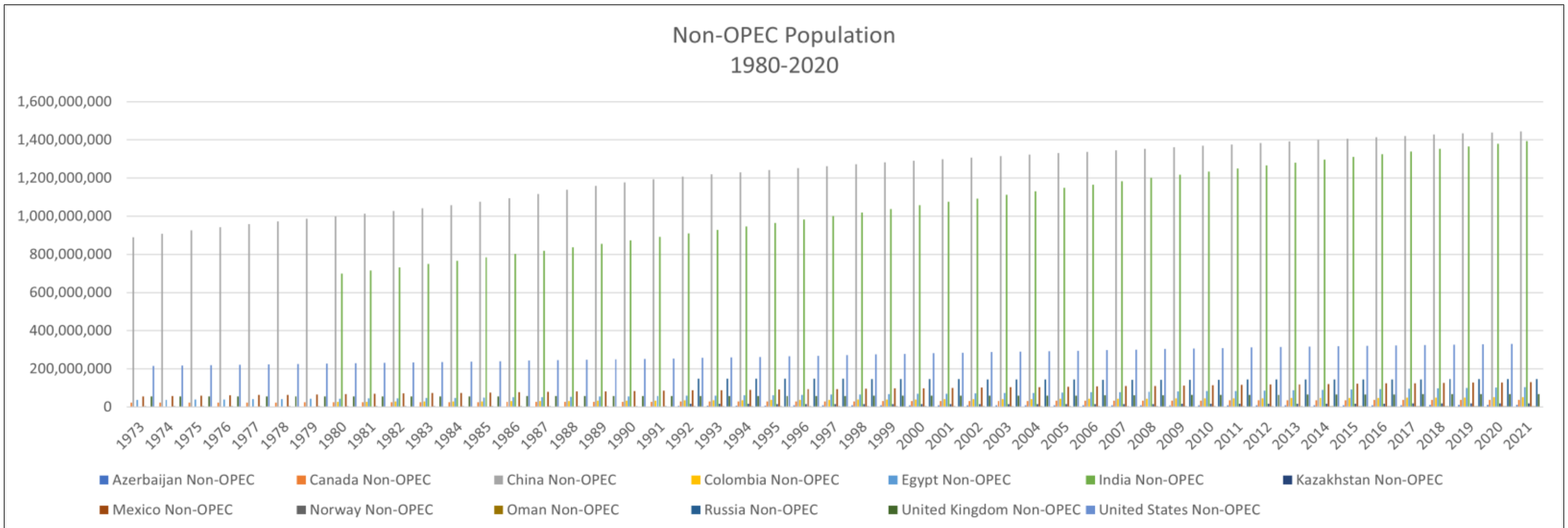


Figure 8. non-OPEC Population

Gross Domestic Product (GDP) is a measurement of the monetary value of all goods and services for a country in a given year. These data were included in the analysis because it points to how much a country produces monetarily and suggests that a country would have produced more resources for use and trade. While this can be insightful, it should be noted that a detailed analysis into the proportion of GDP being hydrocarbon related should be performed to expand on this analysis. The GDP numbers do not provide an analysis of the specific trade trends and do not track the flow and movement of fossil fuel throughout the world. This level of detail was considered outside the scope of this study.

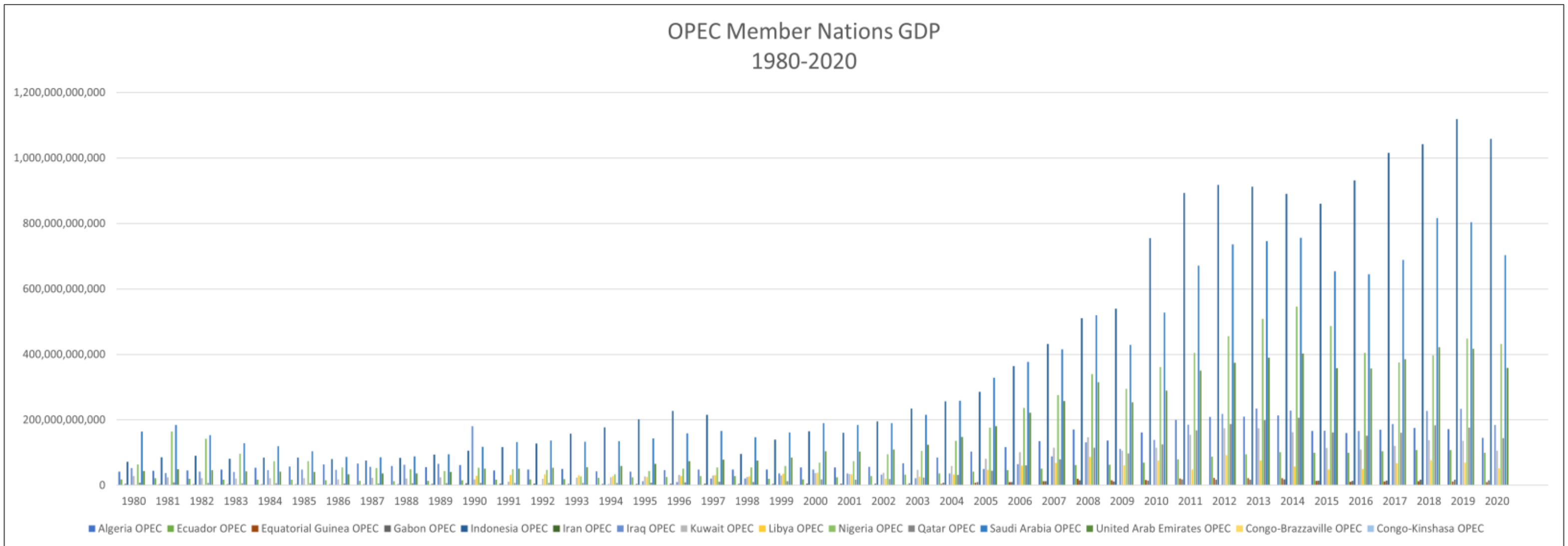


Figure 9. OPEC GDP

For the OPEC members, a surprising trend permeated: massive GDP values for the Congo. This was caused by the Congo crisis in the 1960's which split the country into two (Congo-Brazzaville & Congo-Kinshasa). The problem in the data that arose was that the Congo GDP values effectively displayed the GDP for two countries. The issue here is that the other datasets were not bifurcated, and the GDP values eclipsed other OPEC nations, which were not accurate. For this analysis, the Congo values were split into their respective countries for more accurate numbers, and the trends are showcased in this chart. Note: the Congo is represented as a single entity in the remaining charts for the remainder of the study, as these data were not split into the two countries.

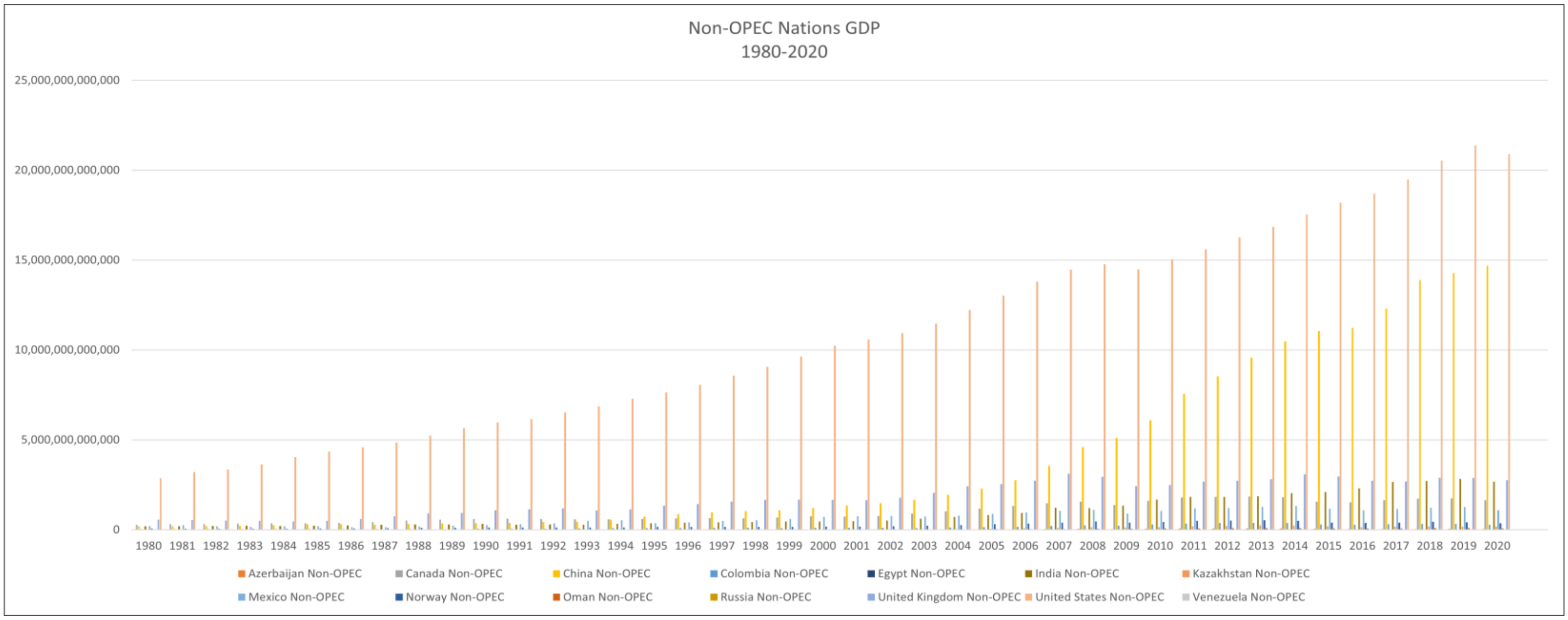


Figure 10. non-OPEC nations GDP

The OPEC GDP numbers suggest that Indonesia has the largest GDP numbers, followed by Saudi Arabia. The numbers for OPEC members are lower overall than their non-OPEC counterparts, with some variability in trends. The United States, despite having a smaller population and emissions values, has the highest GDP. China seems to be catching up rapidly, with the remaining countries staying under five trillion dollars annually. More research could be done to identify the impact of fossil fuels on each countries' economies. This would highlight the magnitude of fossil fuels' impact on global trade.

Coal Production & Consumption Trends

As fossil fuels go, Coal seems to be the least popular, likely due to its' high carbon content compared to the other fossil fuels, as well as it contains many impurities, heavy metals, and other chemicals. First, a line graph was produced from the data, with the graph widget containing two options for graphs: relative values and absolute values. While these graphs are helpful in terms of identifying trends in the data, the data mining software does not allow for the same malleability in terms of changing the formatting of graphs compared to programs like Microsoft Excel. These graphs are good visual observations, but do not allow for changing informaiton used by scales or changing the formatting of numbers in the axes.

This graph displays the absolute values for non-OPEC coal production. Noted trends here were the fact that China produces more coal than any other country by an extreme proportion. The United States comes in second and outpaces India by a slight margin until India overtook the U.S. in the late 2010's. Russia is the last coal producer of note, and the remaining non-OPEC countries dwindle in their coal production comparatively. The relative graph for coal production was also exported using a logarithmic scale to display the data relative to one another and allows for every country's data to be clearly seen rather than losing much of the visibility of the data because of the massive outliers such as China or India.

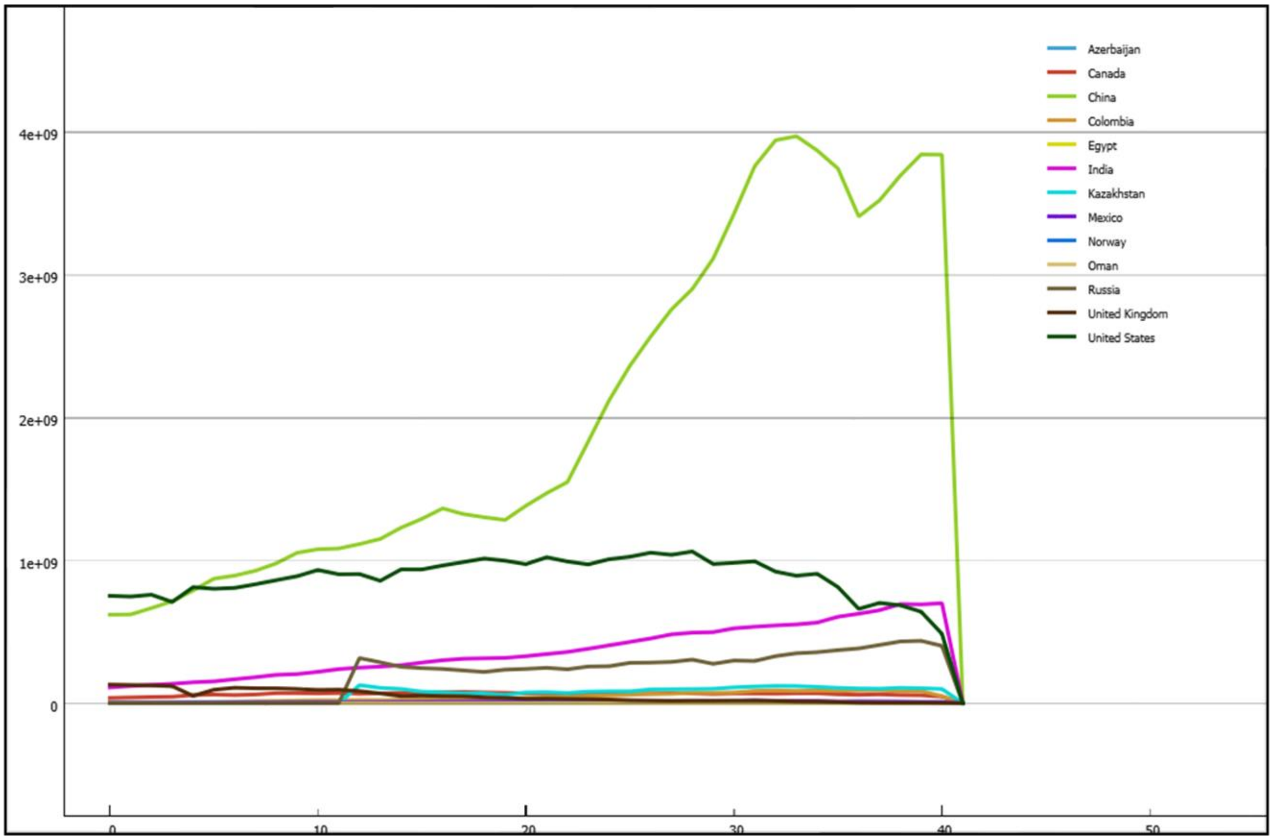


Figure 11. non-OPEC Coal Production graph

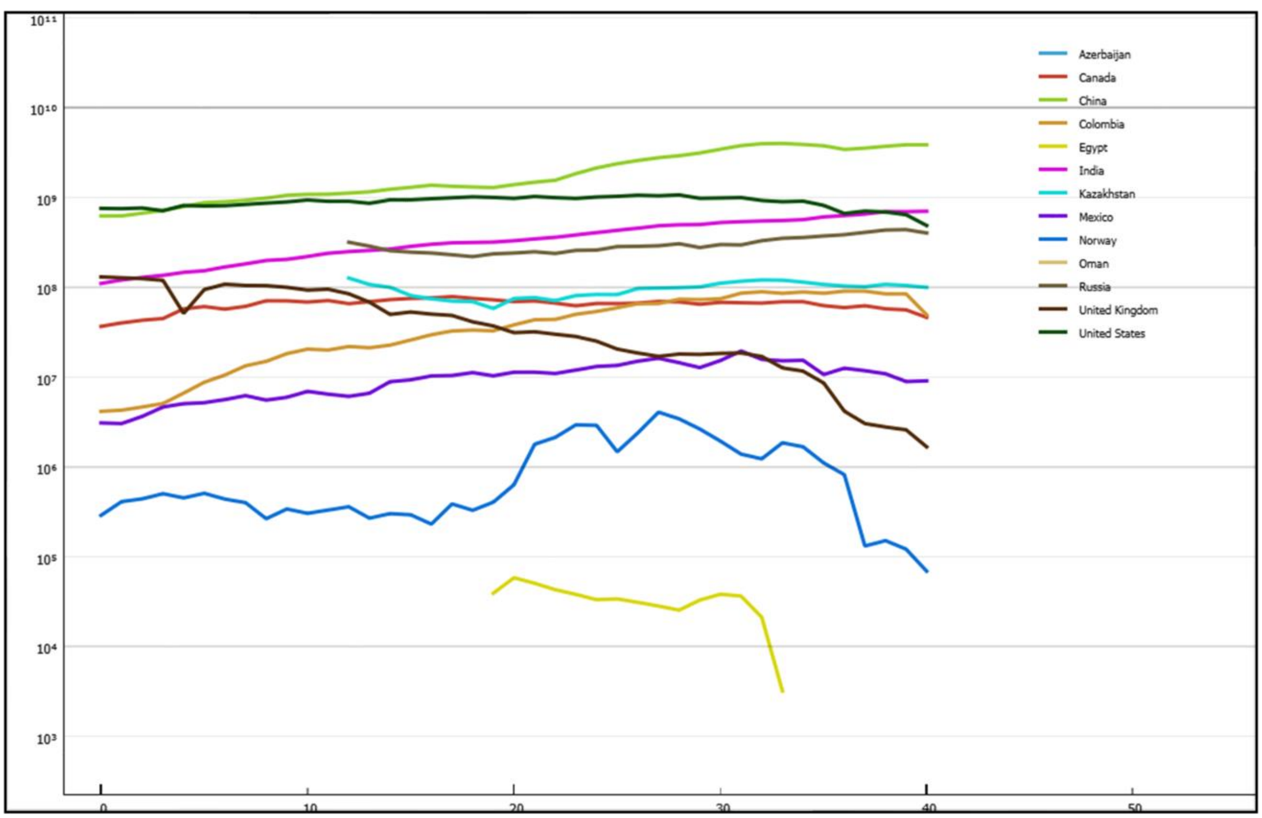


Figure 12. non-OPEC Coal Production Relative Graph

Finally, a correlogram was produced to highlight the correlations and randomness of this dataset. This was, like the heat map, produced by an Orange data mining widget and provided additional insight to how the data behaves relative to one another. This displays correlation coefficients, the overlapping lines highlight coefficients that share a close relationship, and the horizontal dotted lines represent the 95% confidence intervals. The bars that exceed these intervals suggest a statistically significant dataset and the overlapping bars imply a relationship between the non-OPEC countries involved. For example, this chart suggests a relationship between Columbia & Kazakhstan, the United States & Russia and Kazakhstan, as well as the United Kingdom and Norway having non-random data.

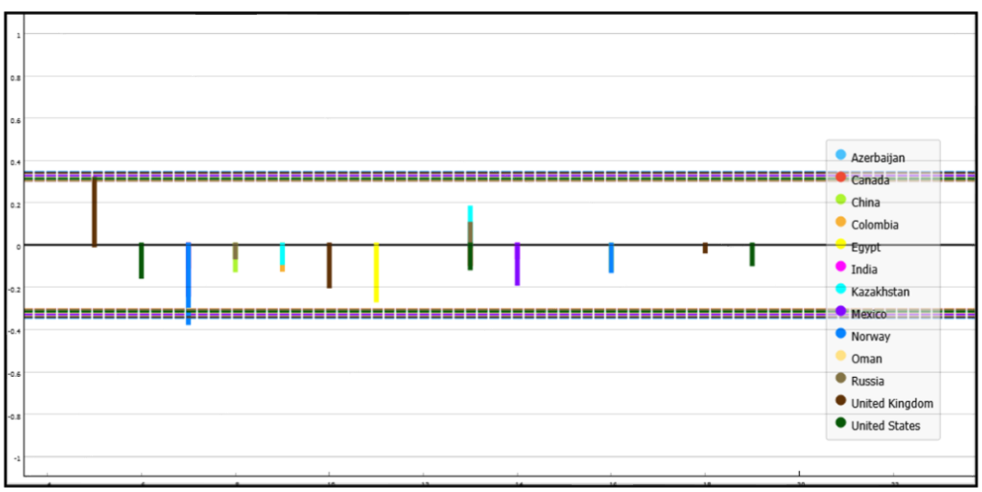


Figure 13. non-OPEC Coal Production Correlogram

Coal production for OPEC members is displayed in the same manner: an absolute graph, relative graph, and correlogram. The data does not seem to be as extreme as the non-OPEC members. It would seem at first glance that the OPEC countries do not focus on coal production as much, focusing mainly on other fossil fuels. The relative coal production graph displays little informaiton, compared to their non-OPEC counterparts, due to a lack of production. The OPEC data correlogram displayed that some data was likely statistically. This includes a correlation with Iran and the United Arab Emirates, as well as potentially another with Algeria and the United Arab Emirates.

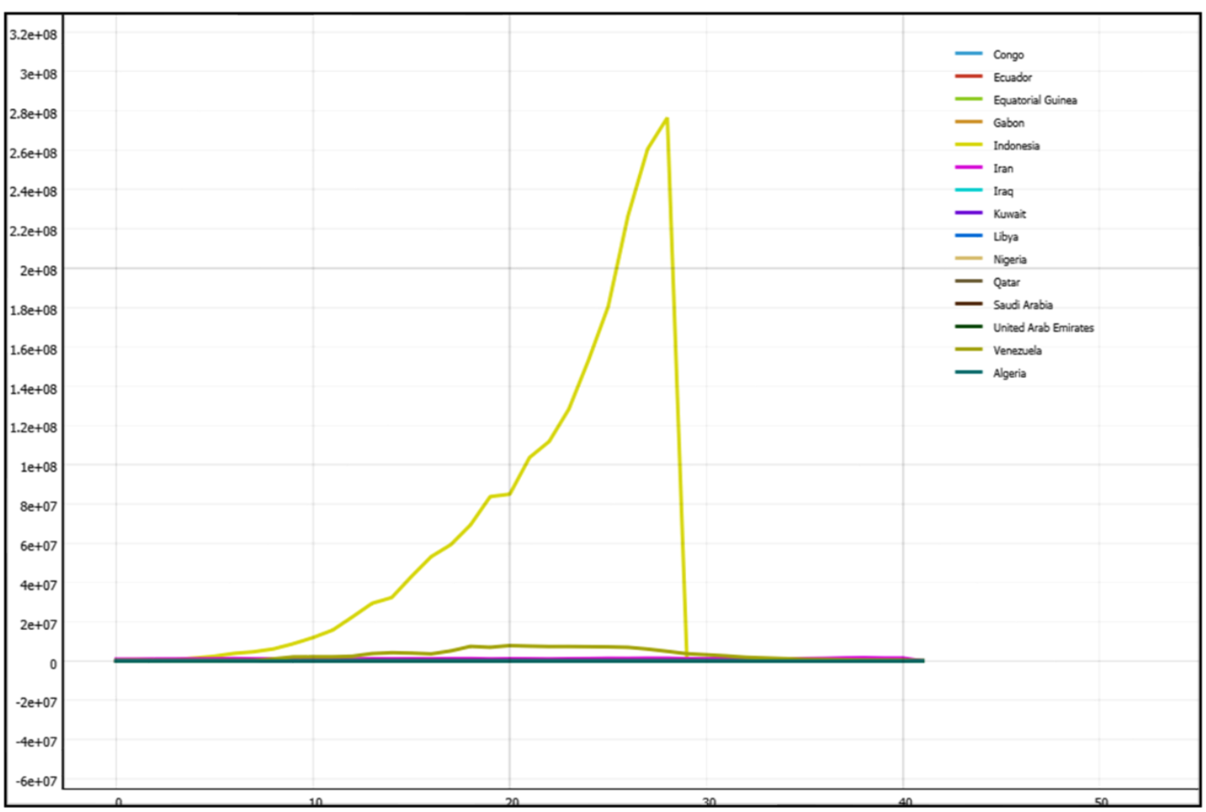


Figure 14. OPEC Coal Production Graph

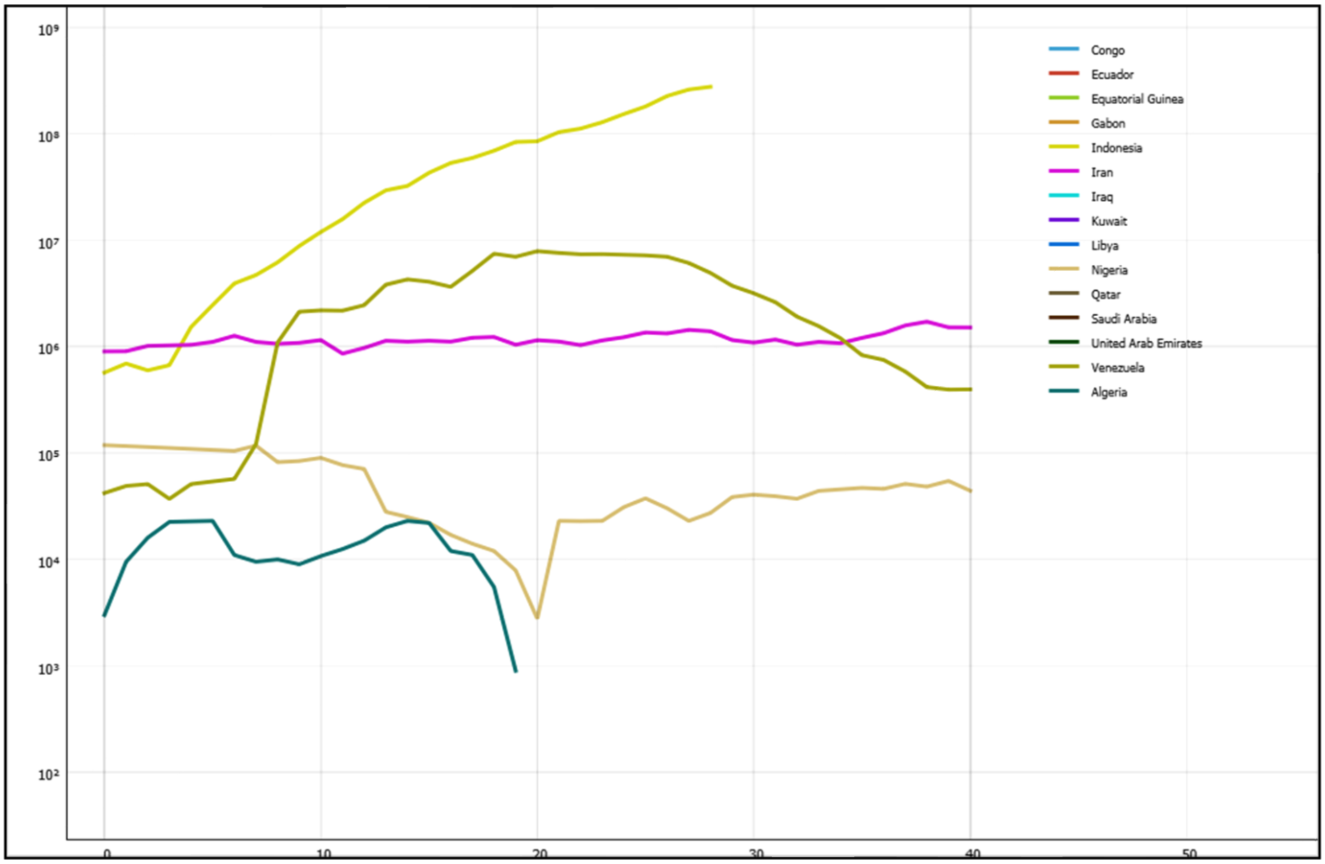


Figure 15. OPEC Coal Production Relative Graph

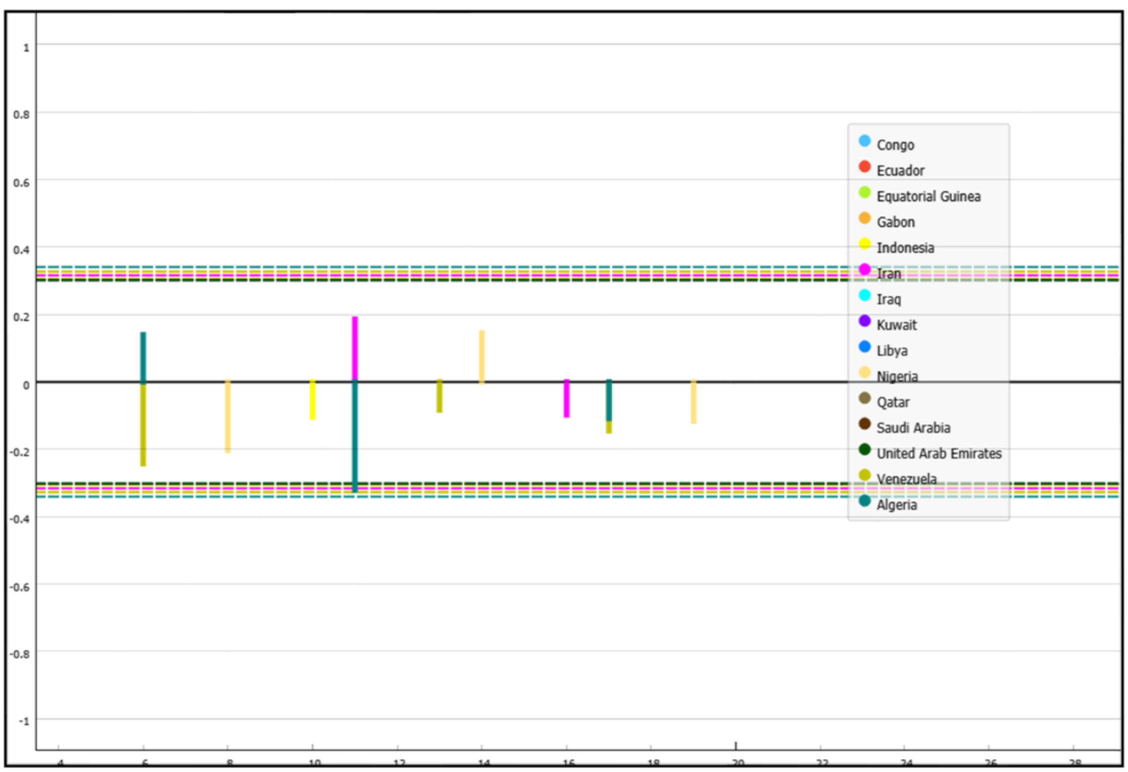


Figure 16. OPEC Coal Production Correlogram

A heat map was produced for coal production, and it showcases the trends in the data in a much clearer manner. The most immediately apparent observation is that the most intense production from OPEC members is hardly visible on the graph relative to the non-OPEC group. China is the overall highest coal producer; however, it should be noted that they have slowed their production since about 2013 onward. At first glance it would seem that the response to climate change is resonating through the data and can be seen here. More research can be done to verify China's response to climate change, and this may explain why China's coal production has declined.

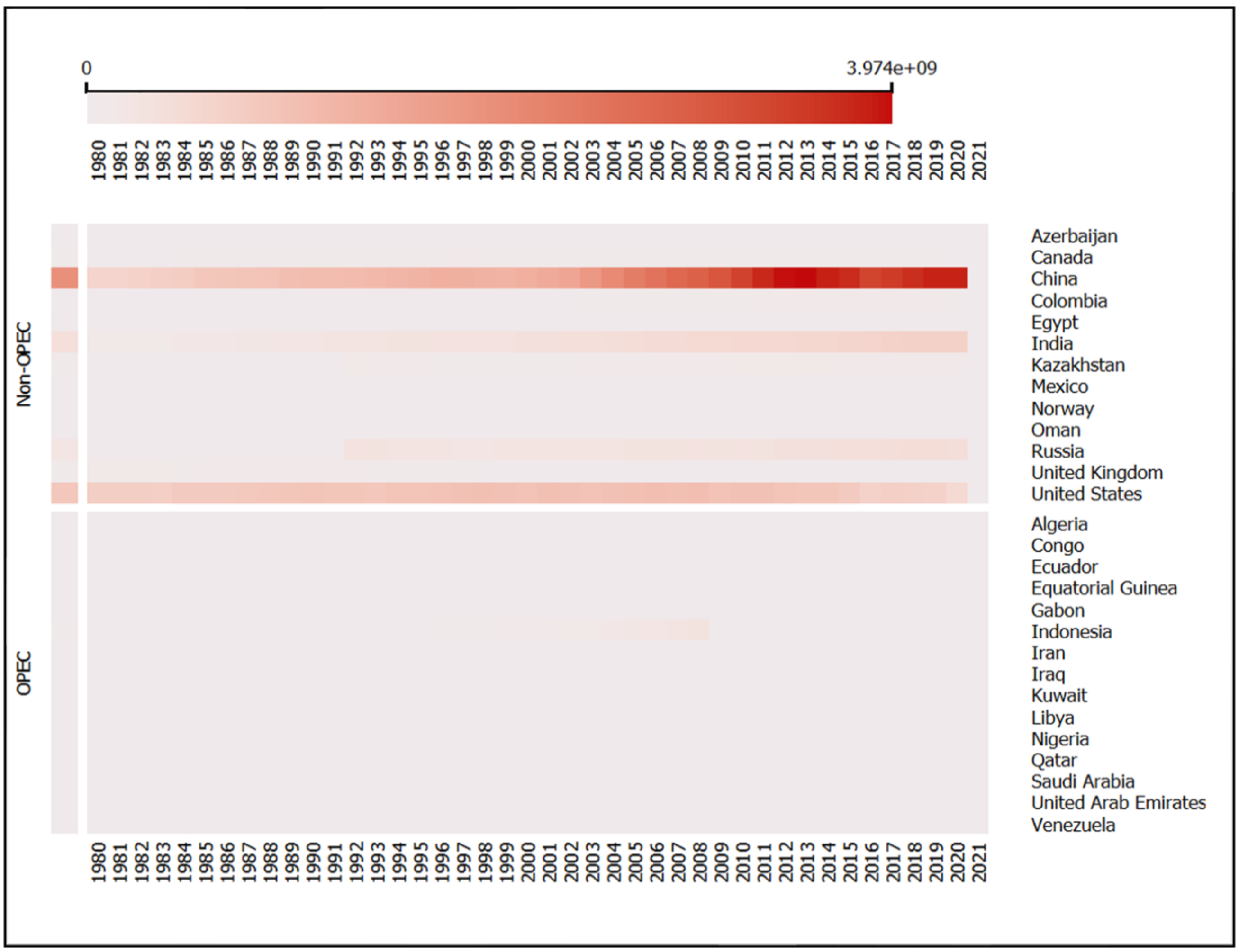


Figure 17. Coal Production Heat Map

As for coal consumption for non-OPEC countries, China is by far the largest consumer of coal, followed by the United States, until India overtook the US for second place. Russia ramped up their consumption and has remained the fourth largest consumer of coal since the early 1990's. Specifically, this trend seems to coincide with the collapse of the Soviet Union. The relative graph confirms these observations and highlights the other countries' trends in greater detail. For the non-OPEC correlogram, there seems to be less randomness in this dataset, as well as a significant correlation in the leftmost value (Oman & the United Kingdom) as this value approaches the 95% confidence interval.

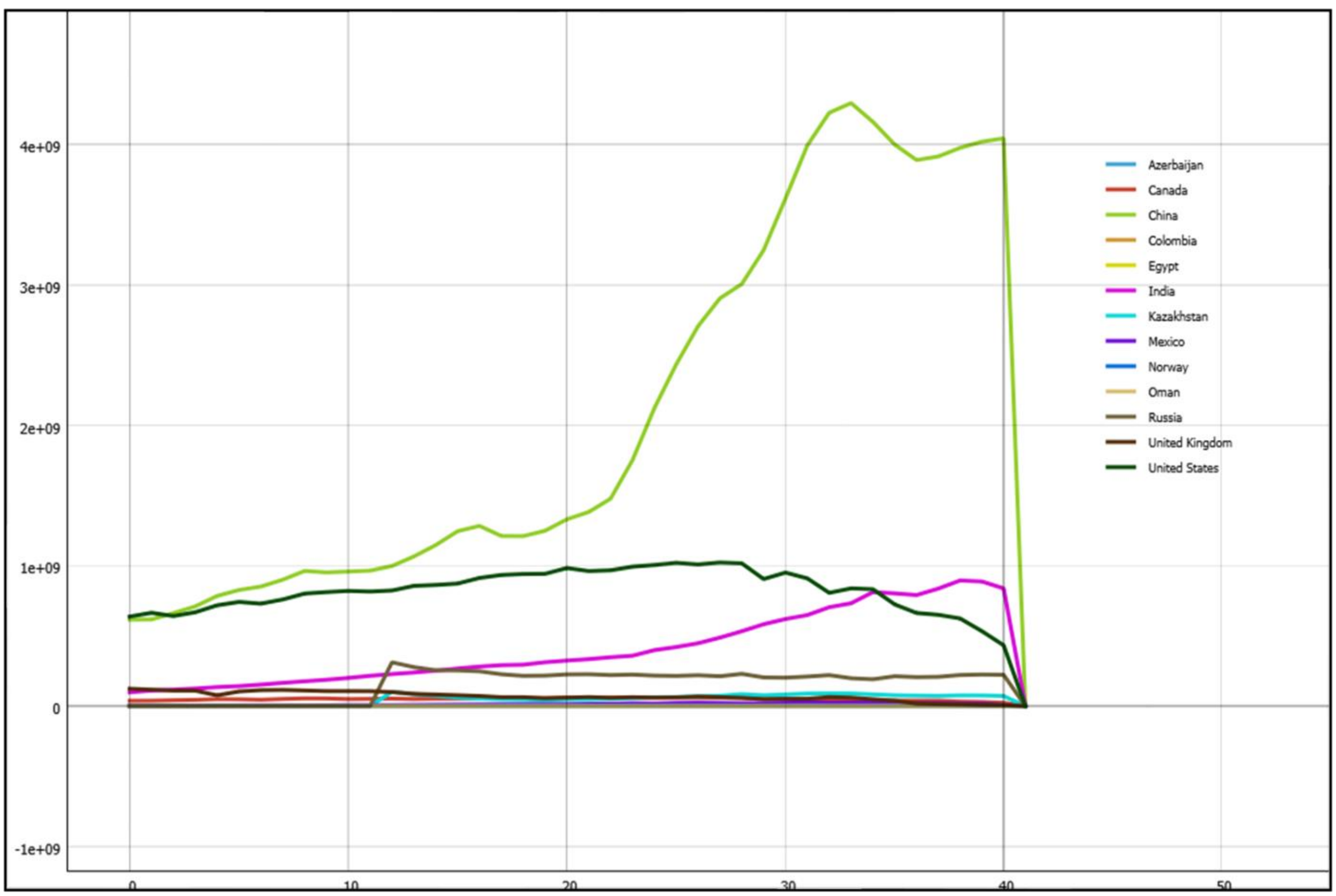


Figure 18. non-OPEC Coal Consumption Graph

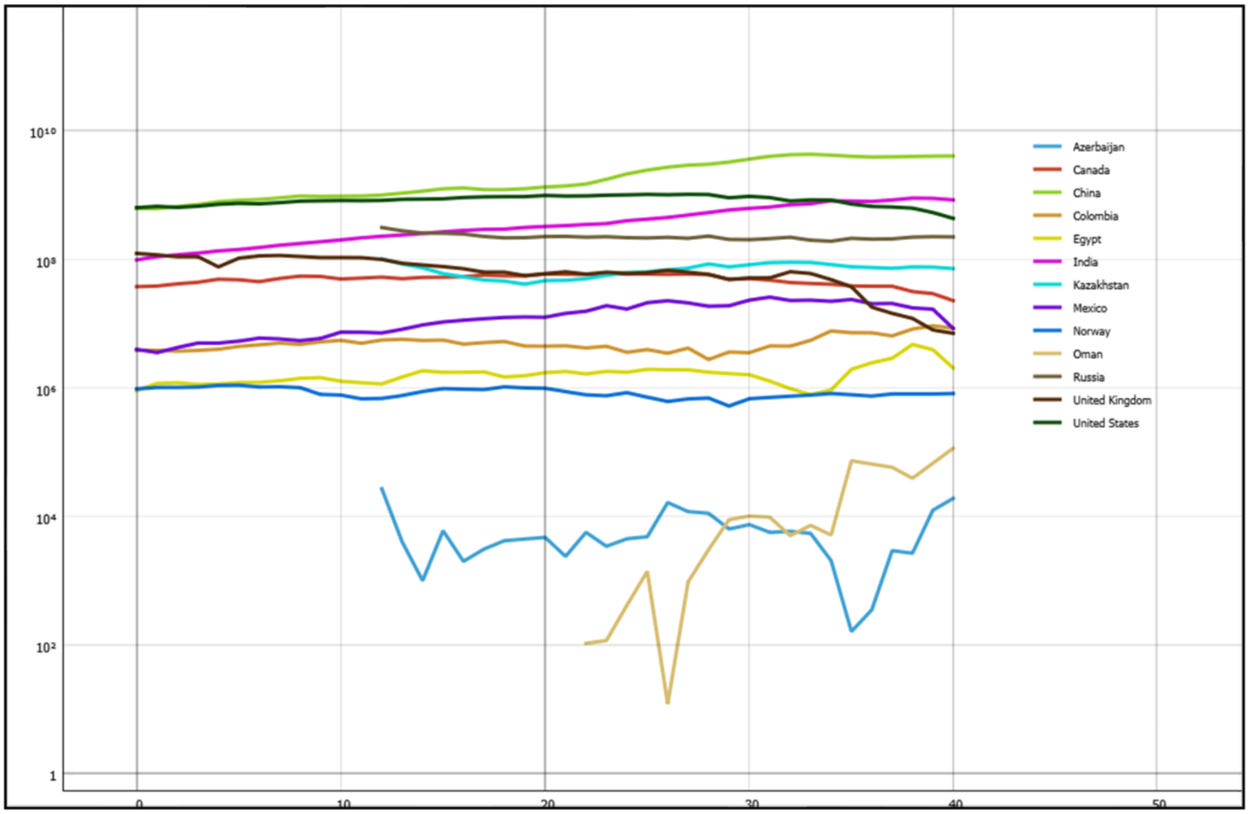


Figure 19. non-OPEC Coal Consumption Relative Graph

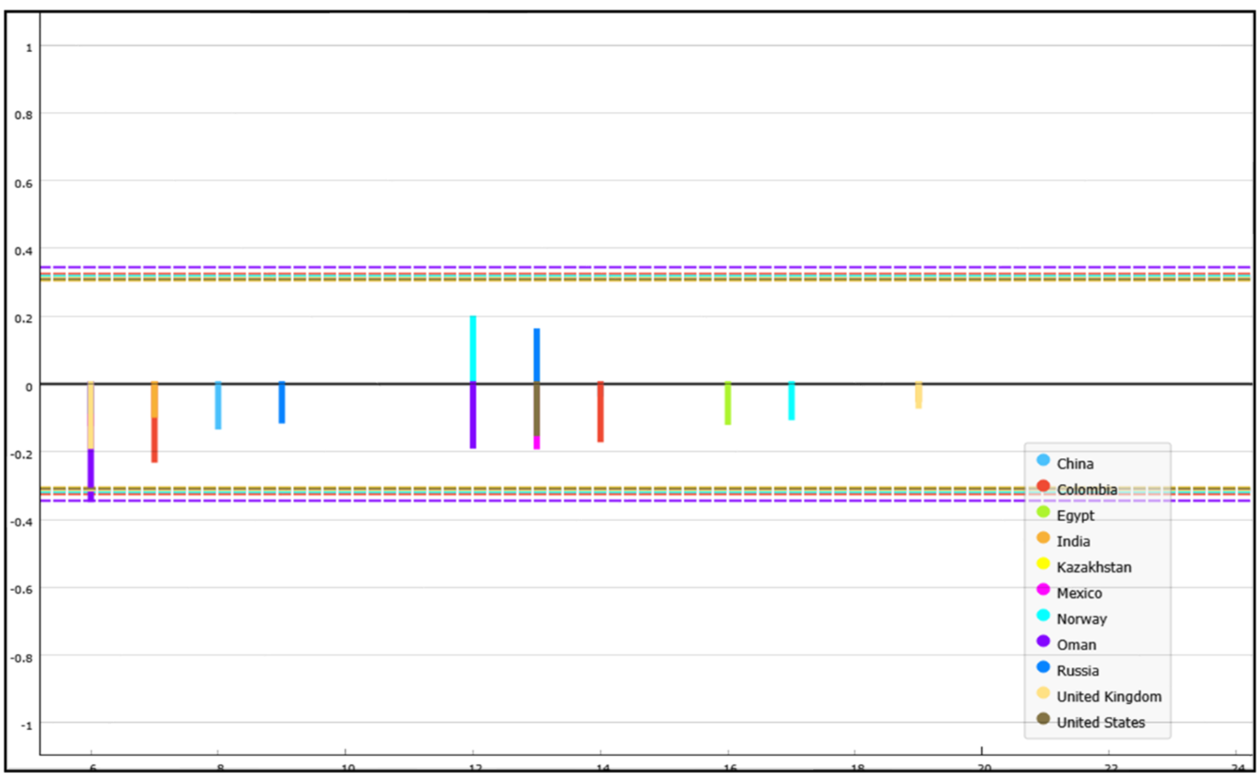


Figure 20. non-OPEC Coal Consumption Correlogram

The OPEC nations also seem to abstain from consuming coal, save for Indonesia. In order to derive better conclusions from this data, the relative graph will need to be consulted to view the other OPEC nations data. The other consumers seem to stay relatively consistent in their coal consumption, except for Kuwait, which seemed to ramp up their consumption and have rocketed up to the fourth largest consumer.

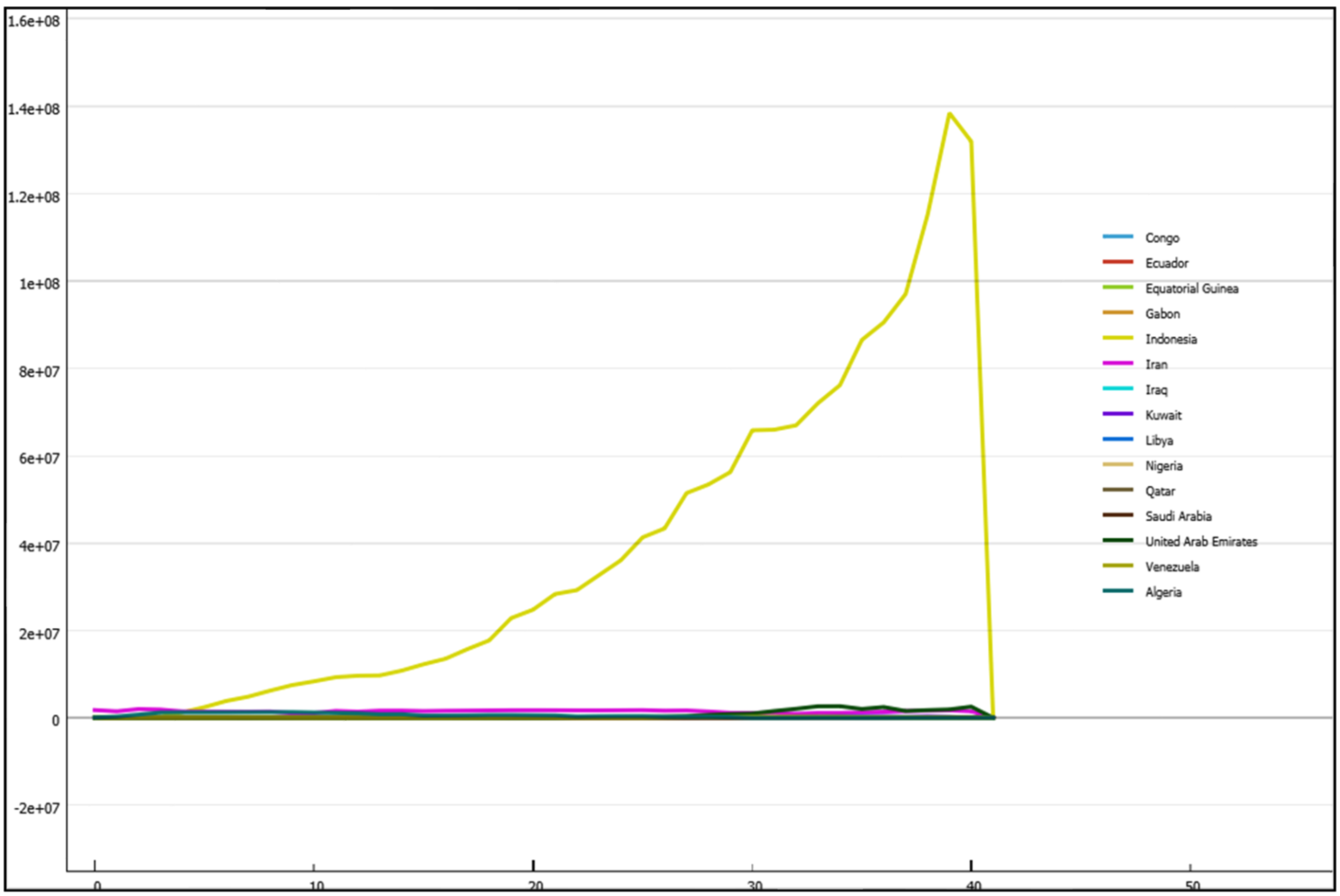


Figure 21. OPEC Coal Consumption Graph

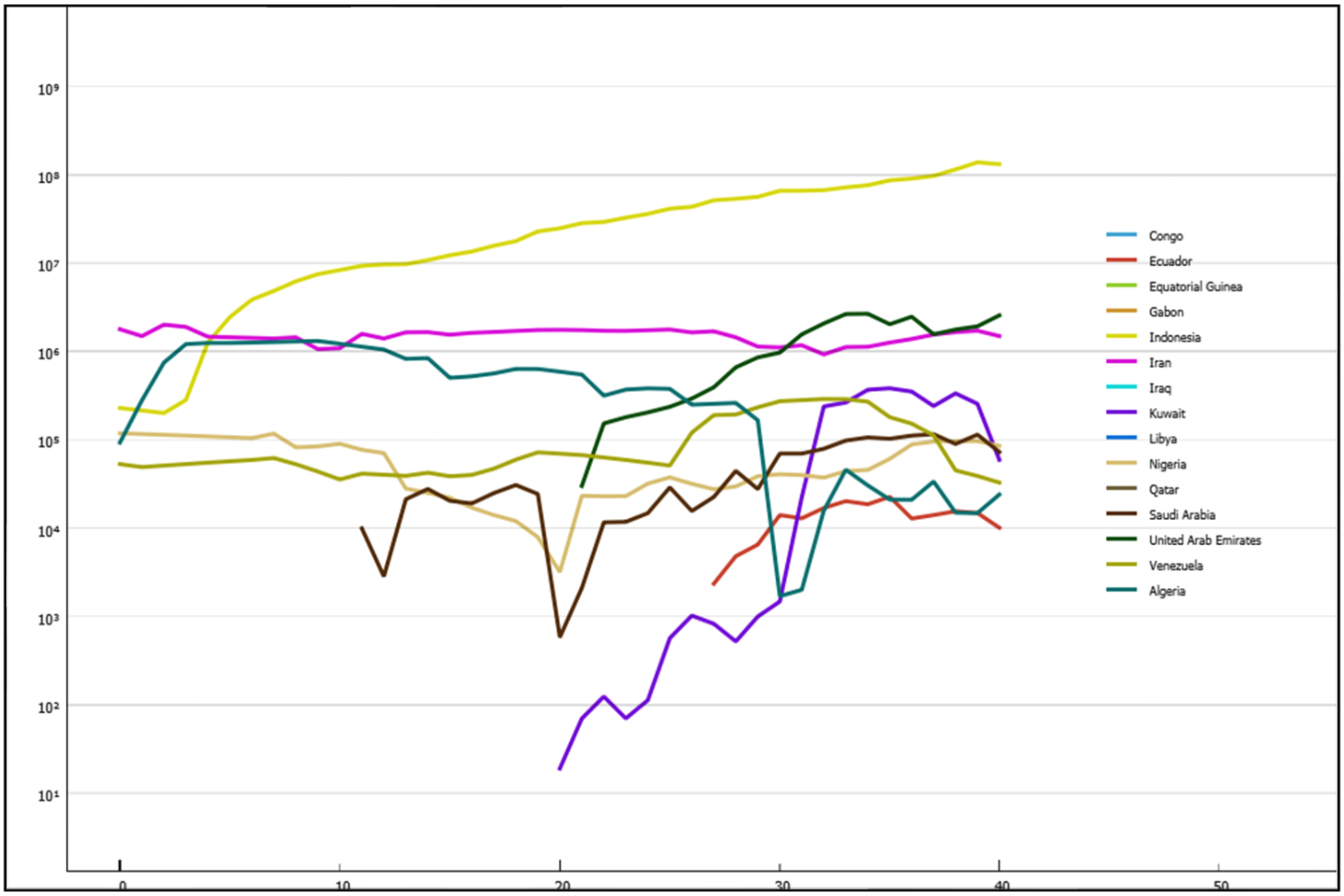


Figure 22. non-OPEC Coal Consumption Relative Graph

As for the correlogram, there are several statistically significant correlations in the data (including Iran, Nigeria, Saudi Arabia, and Indonesia), and most everything else exists well below the 95% confidence level, save for Venezuela, Indonesia, and Saudi Arabia.

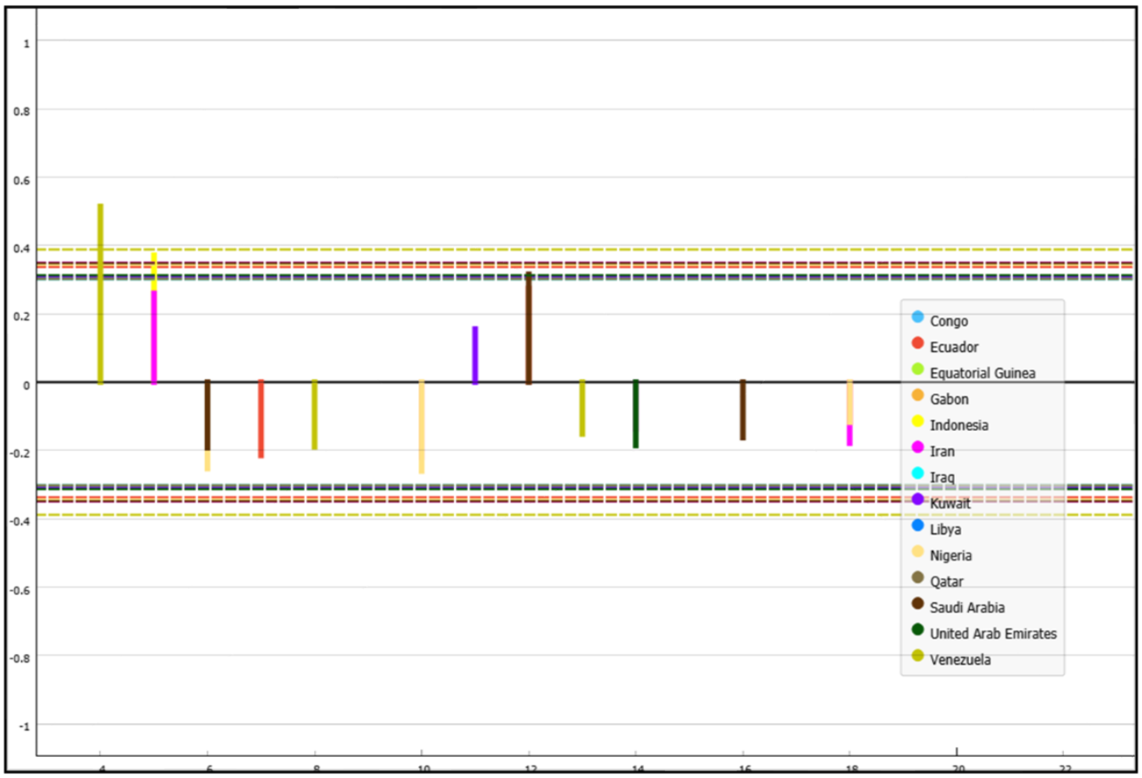


Figure 23. non-OPEC Coal Consumption Correlogram

As for the coal consumption heat map, it displays the trends in the same manner as previous charts, but also displays just how different the two groups can be: the coal usage among the non-OPEC nations is so great that the OPEC usage does not even register in the chart. The clear takeaway here is that China, India, and the United States are consuming the most coal, and the OPEC members use hardly any of the fuel.

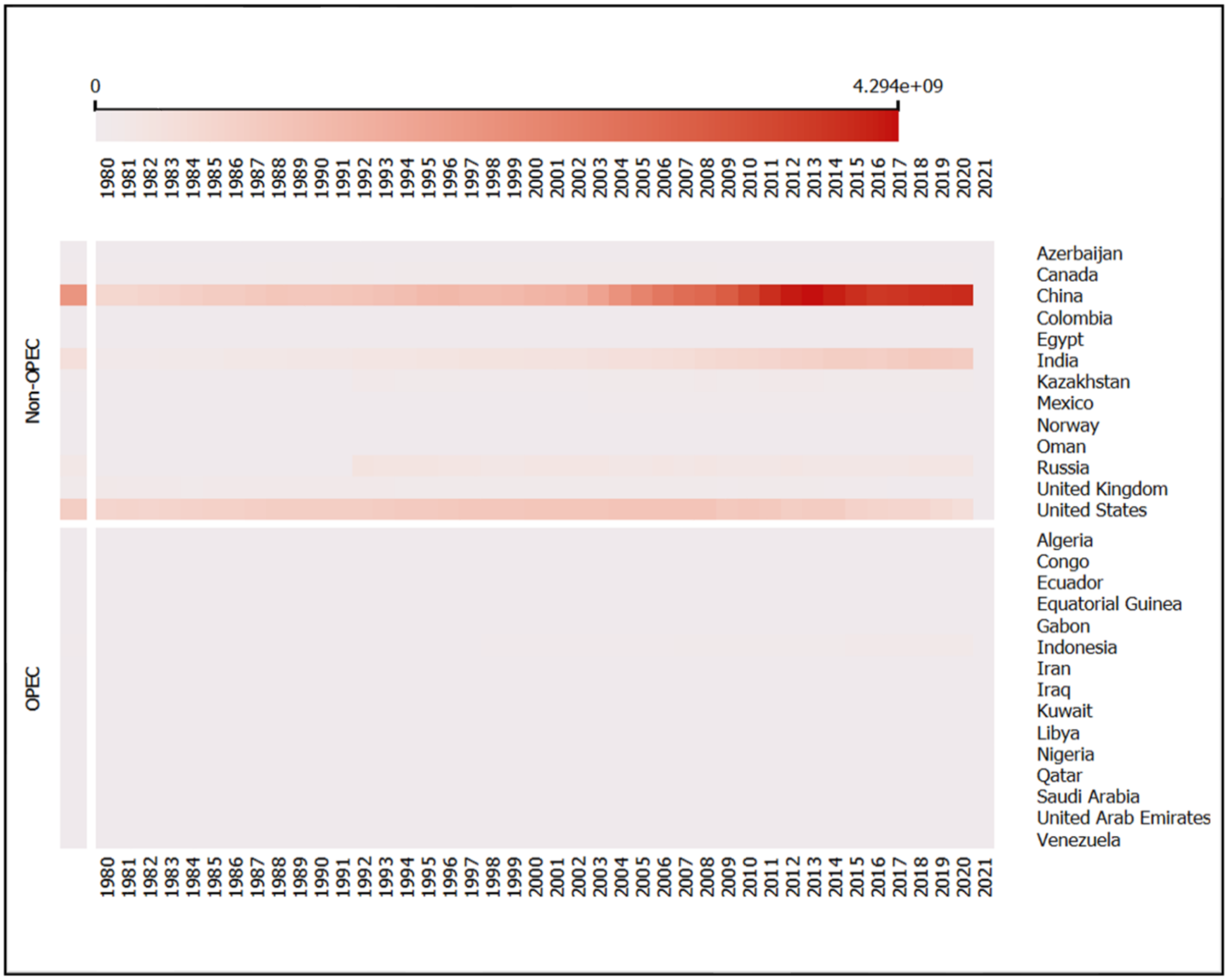


Figure 24. non-OPEC Coal Consumption Heat Map

Natural Gas Production & Consumption Trends

The trends that exist for gas production are not at all like those of the coal production trends. As for the primary non-OPEC country producer, the United States and Russia have alternated between the primary producers, with the United States taking the lead and remaining there in recent years. In order to view the trends of the clustered countries, the relative chart was utilized.

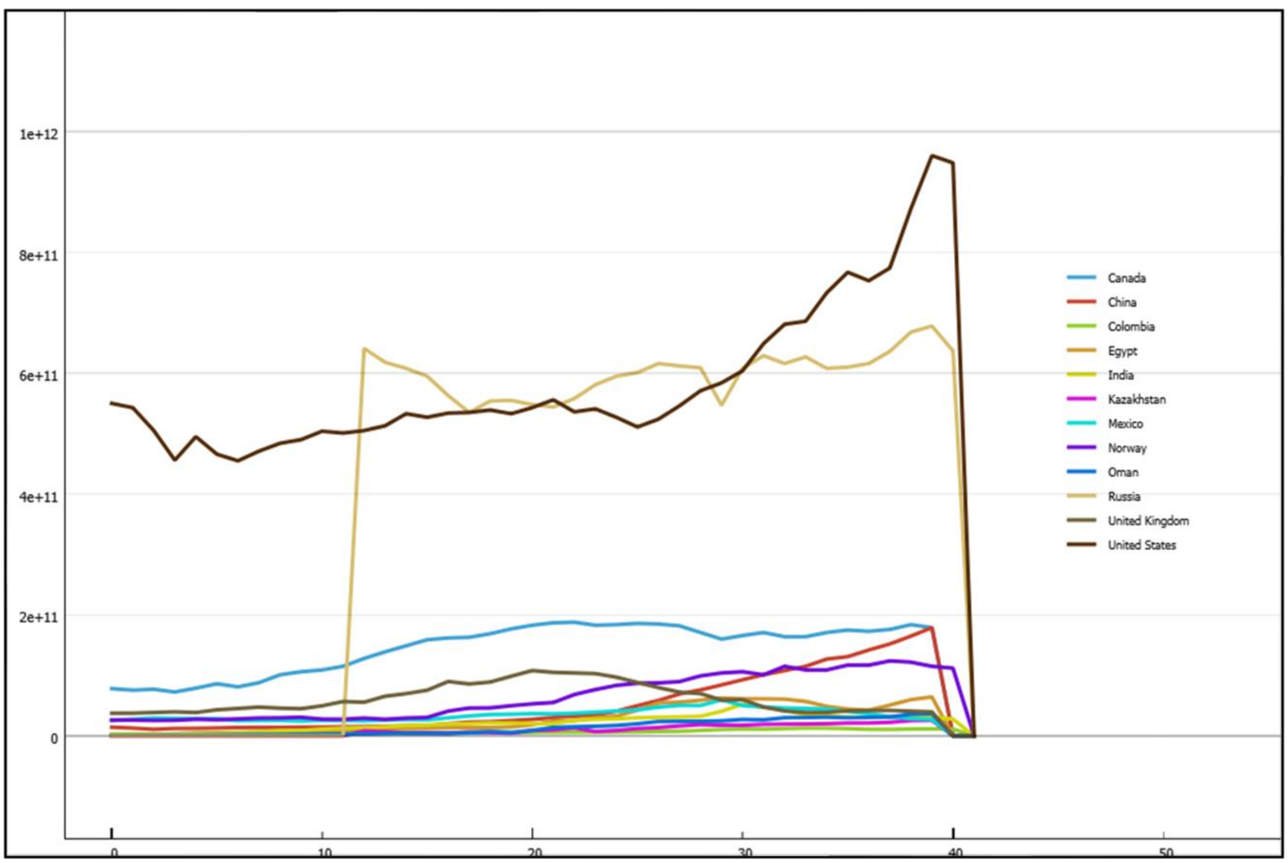


Figure 25. non-OPEC Gas Production Graph

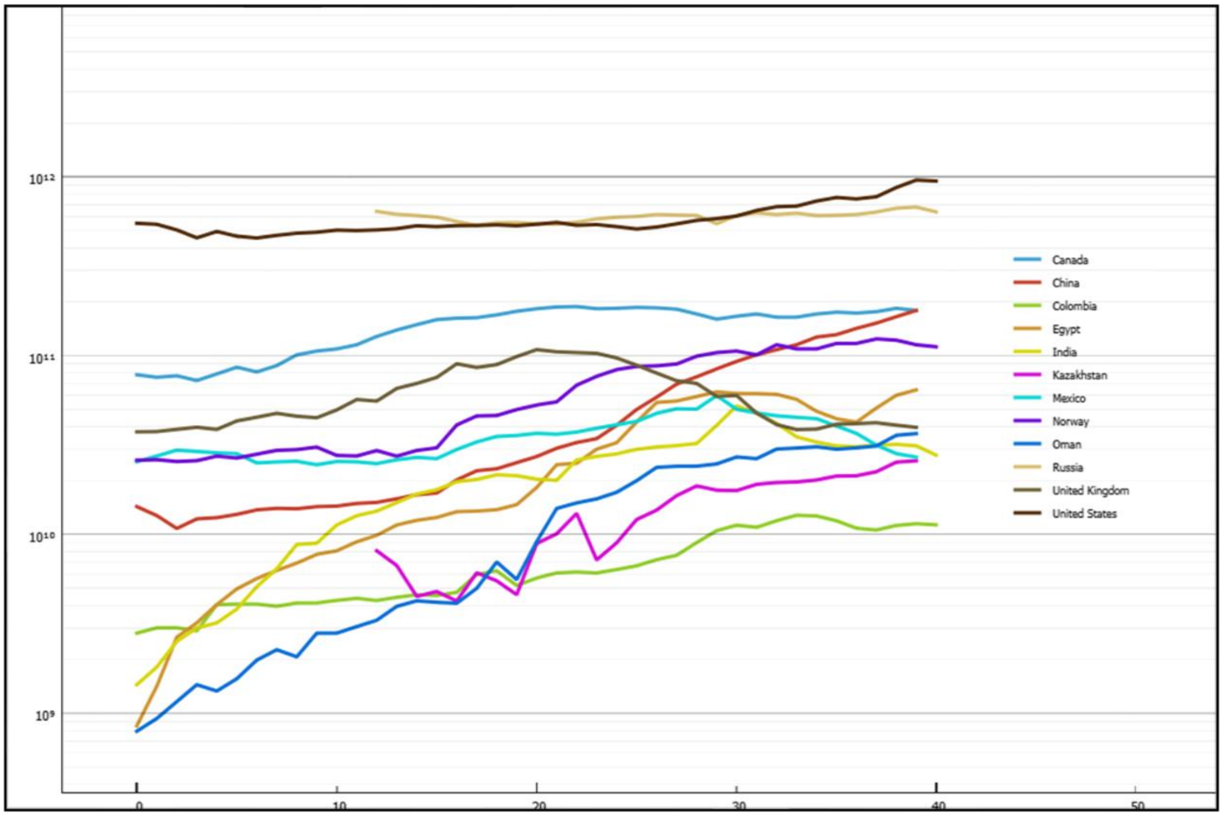


Figure 26. non-OPEC Gas Production Relative Graph

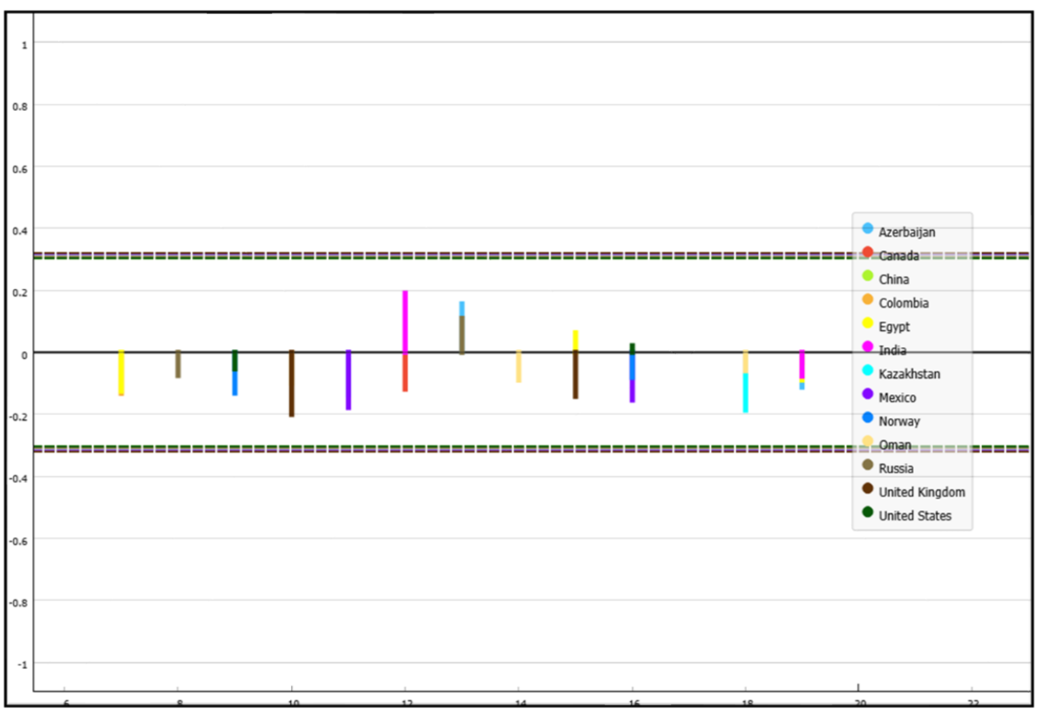


Figure 27. non-OPEC Gas Production Correlogram

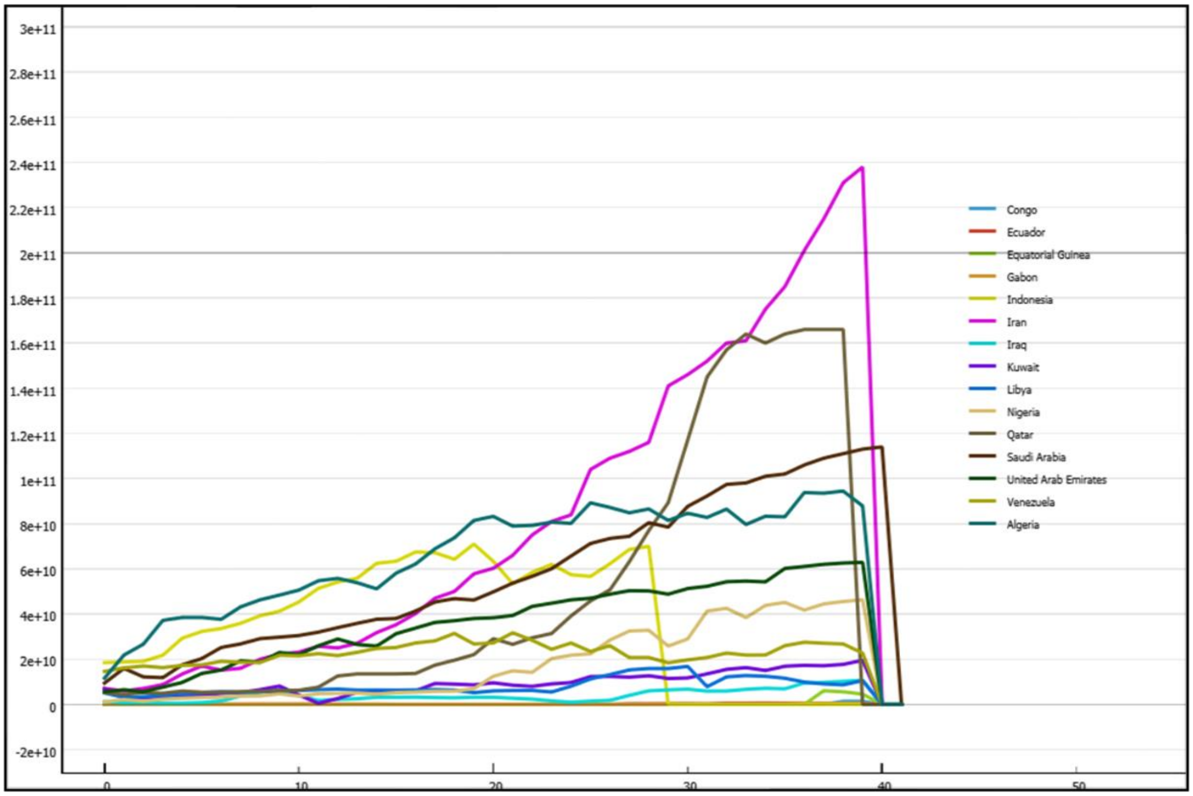


Figure 28. OPEC Gas Production Graph

The most shocking trend noted was that China and India are present among the lower producers, it may be due to coal mining being a more economically viable option then gas or Oil for the country. While it's unwise to assume that these two countries would remain the top producers for all kinds of fossil fuels, the idea that India is reducing its production of gas is not something that would be expected necessarily, given their population and growth. As previously mentioned, these trends in the data open new avenues for research to explain why they exist rather than merely observing. The correlogram yielded no statistically significant correlations or nonrandom data.

The OPEC trends were more evenly distributed across the production numbers. Iran was consistently the highest producer, Qatar followed in recent years, but Algeria was second place until Qatar and Saudi Arabia overtook it. After consulting the relative scale chart, several holes were present in the data. This is because countries like Gabon and Ecuador left OPEC during various years, and these data were removed to ensure that all data included was only for OPEC members. Apart from periods of time where countries like Iraq or Kuwait's production fluctuated, these trends seem to gradually rise.

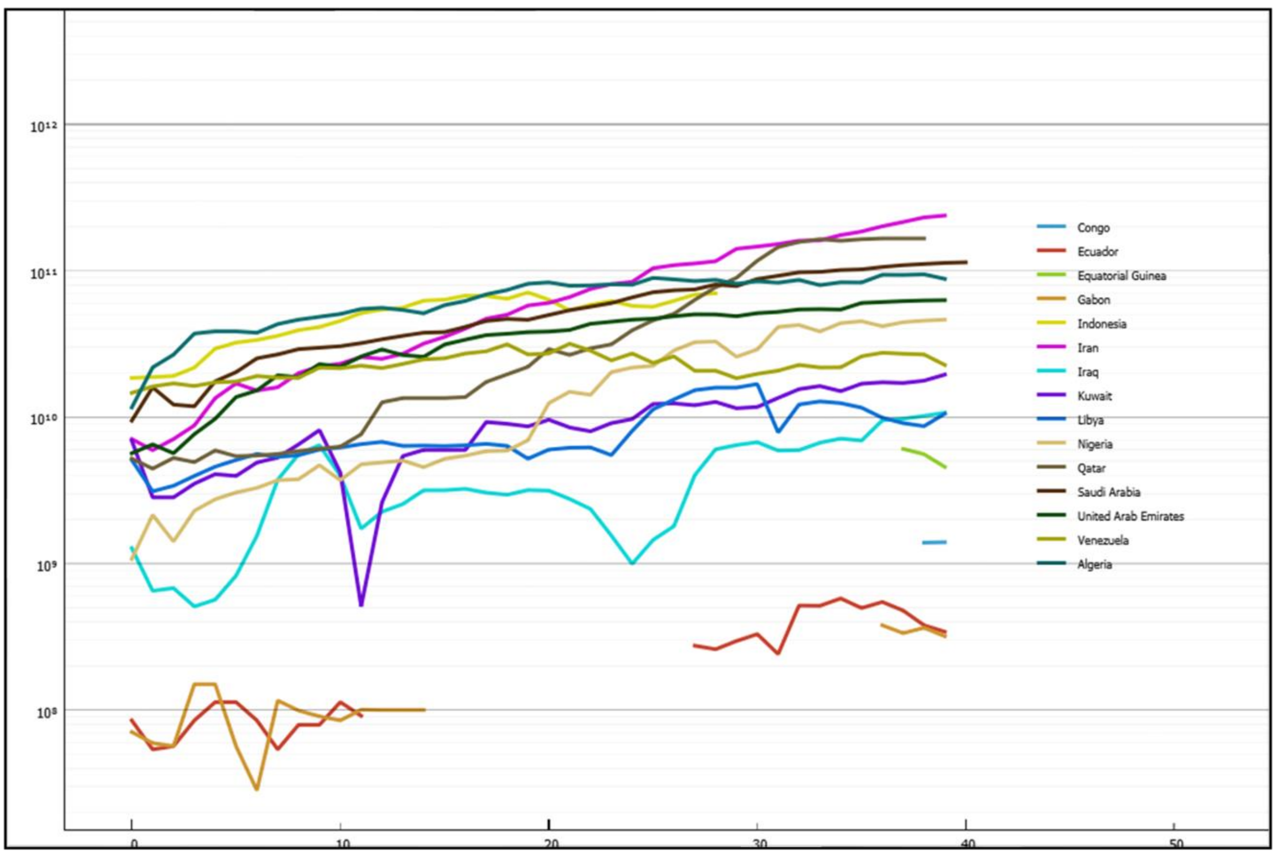


Figure 29. OPEC Gas Production Relative Graph

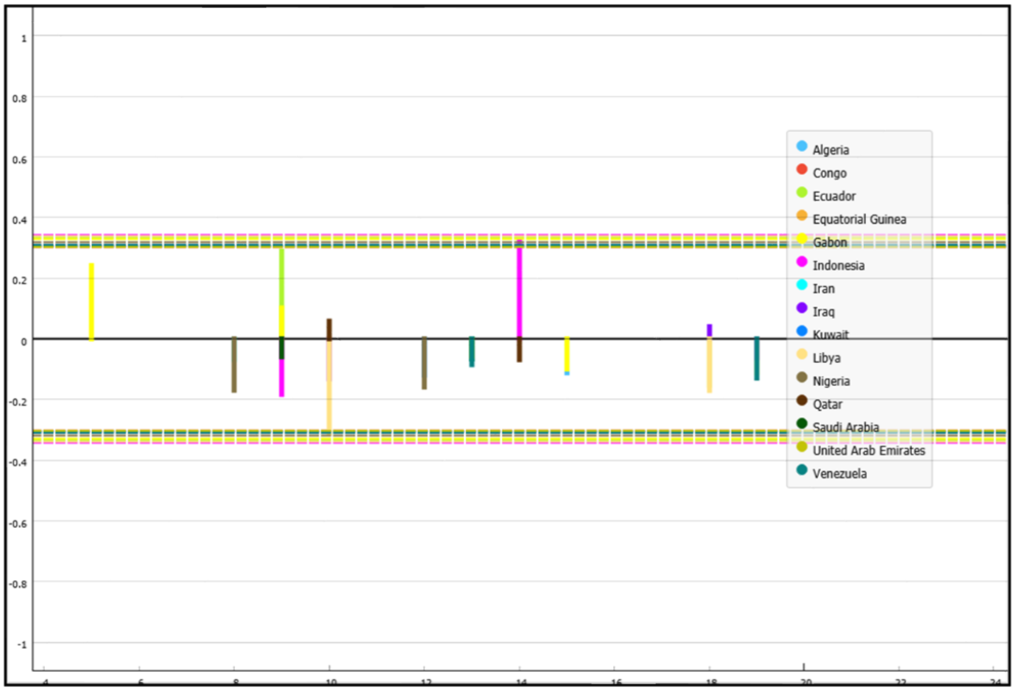


Figure 30. OPEC Gas Production Correlogram

As for the correlogram for this data series, there is a very statistically significant non-zero correlation for Indonesia, Libya, and Ecuador. These correlations suggest some significantly nonrandom data. The heat map reaffirms the previously observed trends. The United States produces the most gas of any country, followed by Russia. Overall, the non-OPEC countries are producing most of the gas, and OPEC nations do not have a history of producing massive amounts of gas, though they are increasing their production numbers over time.

As for the non-OPEC gas consumption, the United States are far above the other countries, with Russia in a firm second place. Note, data are missing from Russia in the earlier years, with data appearing in the early 1990's (corresponding with the fall of the Soviet Union). China has overtaken Canada and the United Kingdom as the third highest gas consumer.

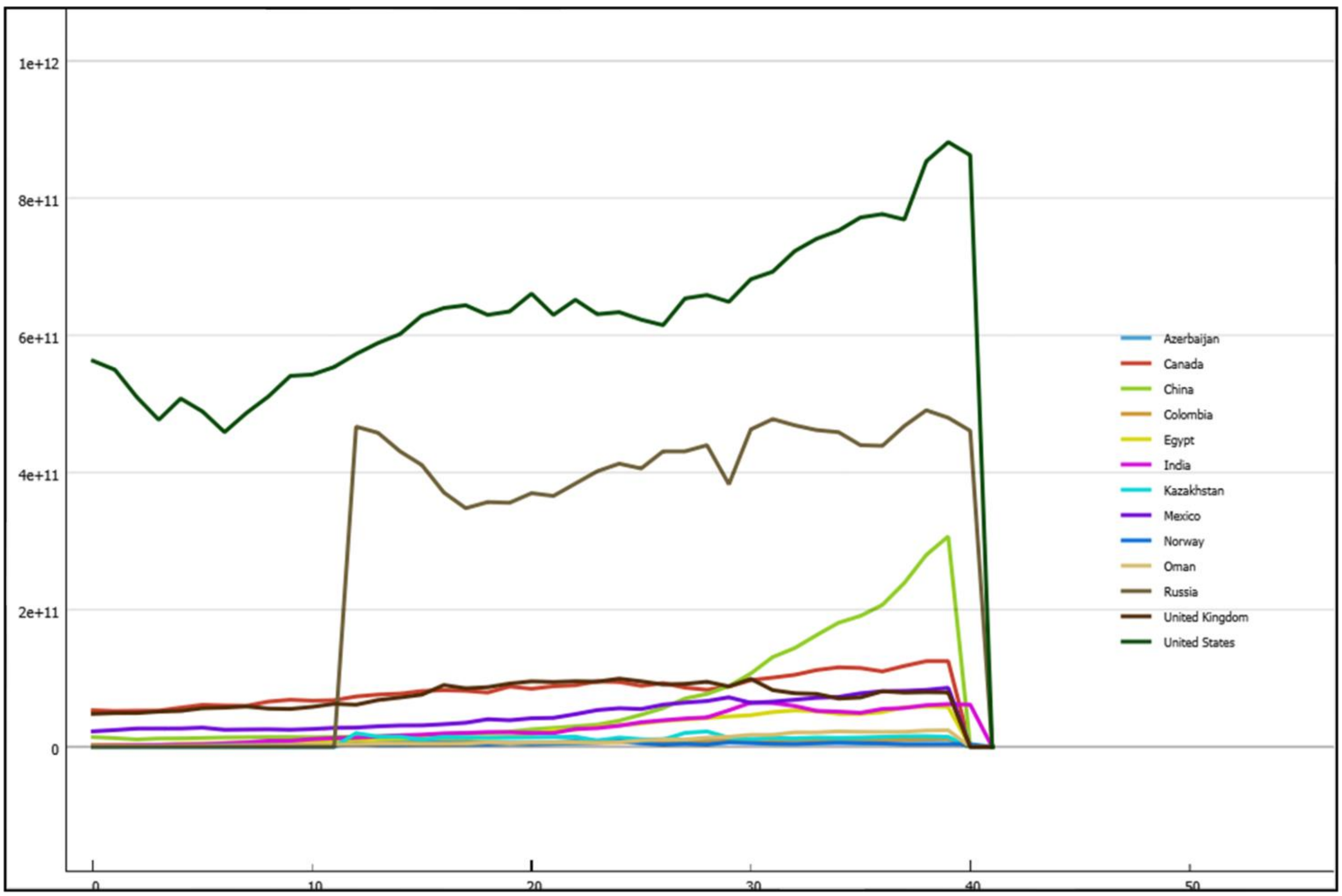


Figure 31. non-OPEC Gas Consumption Graph

The relative chart displays the gas consumption trends and does a better job of illustrating how China overtook Canada and the United Kingdom as well as showing the other non-OPEC countries as they change over time.

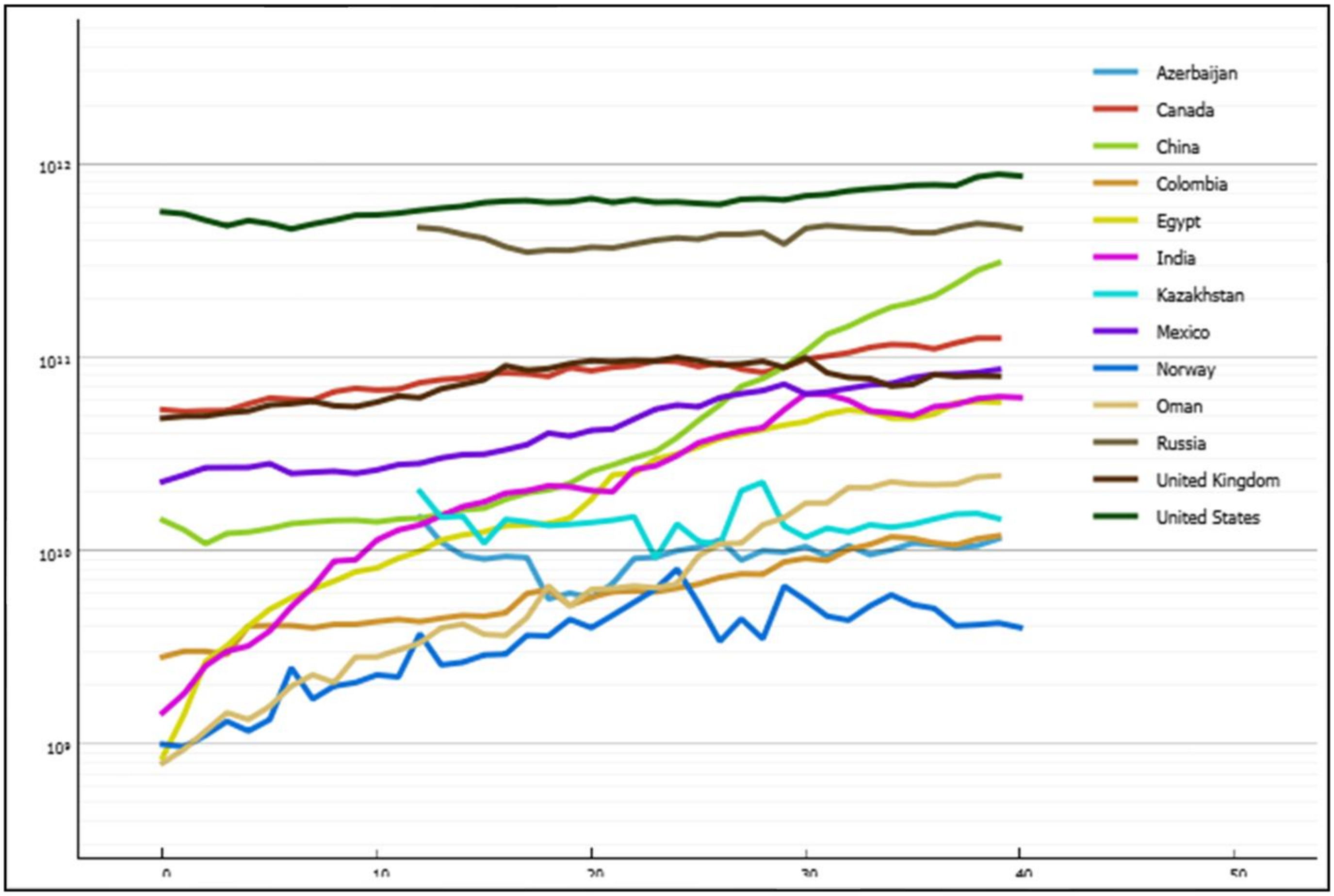


Figure 32. non-OPEC Gas Consumption Relative Graph

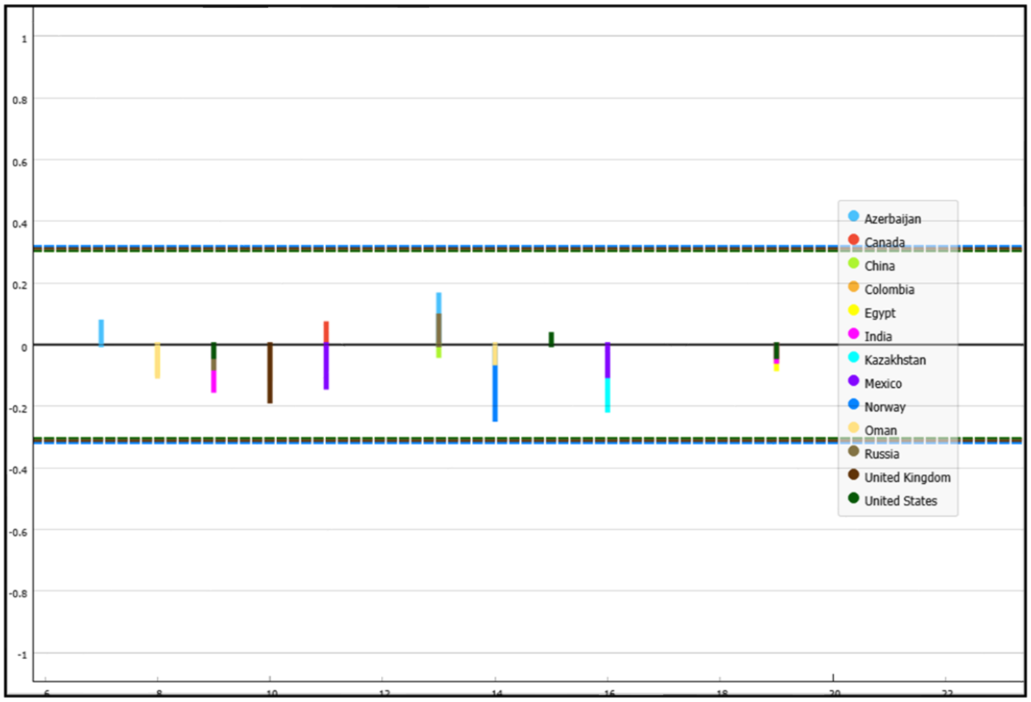


Figure 33. non-OPEC Gas Consumption Correlogram

The correlogram for Non-OPEC Gas consumption does not contain any statistically relevant correlations or non-random data. However, there seem to be slight correlations between Kazakhstan and Mexico, Mexico and Canada, Russia and the UK, and India and the US as well as a few others. This likely can be explained by trade and present another avenue for further research.

The gas consumption for OPEC countries show that Iran is the clear leader, Saudi Arabia is the second, followed by the United Arab Emirates.

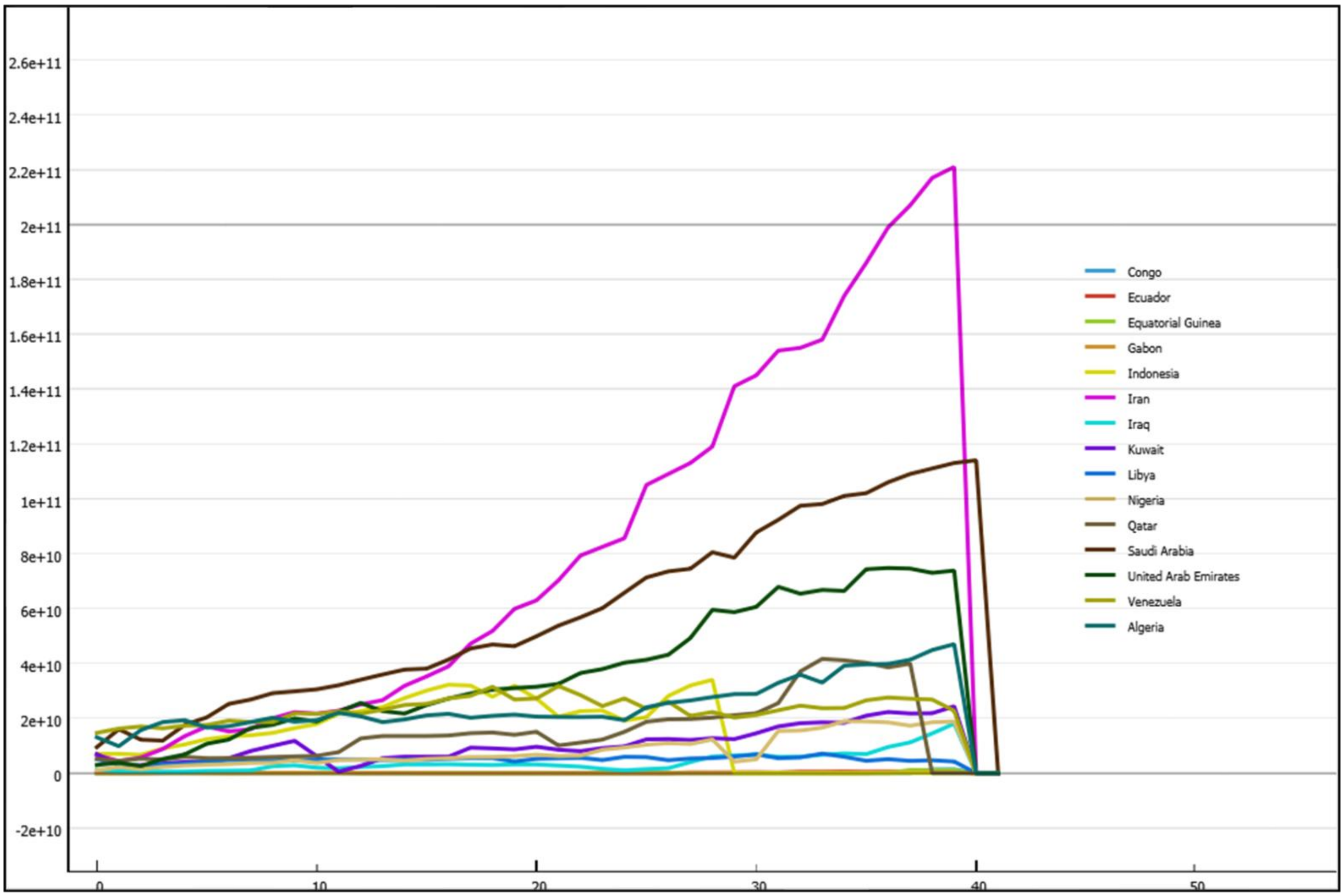


Figure 34. OPEC Gas Consumption Graph

In order to view the trends for the lesser consumers in this dataset, the relative chart was consulted. This chart also showcases holes in the data which represent the years where OPEC members did not belong to OPEC. However, it highlights the trends present in countries like Iran and Kuwait, who's consumption fluctuated during the 1990's and early 2000's.

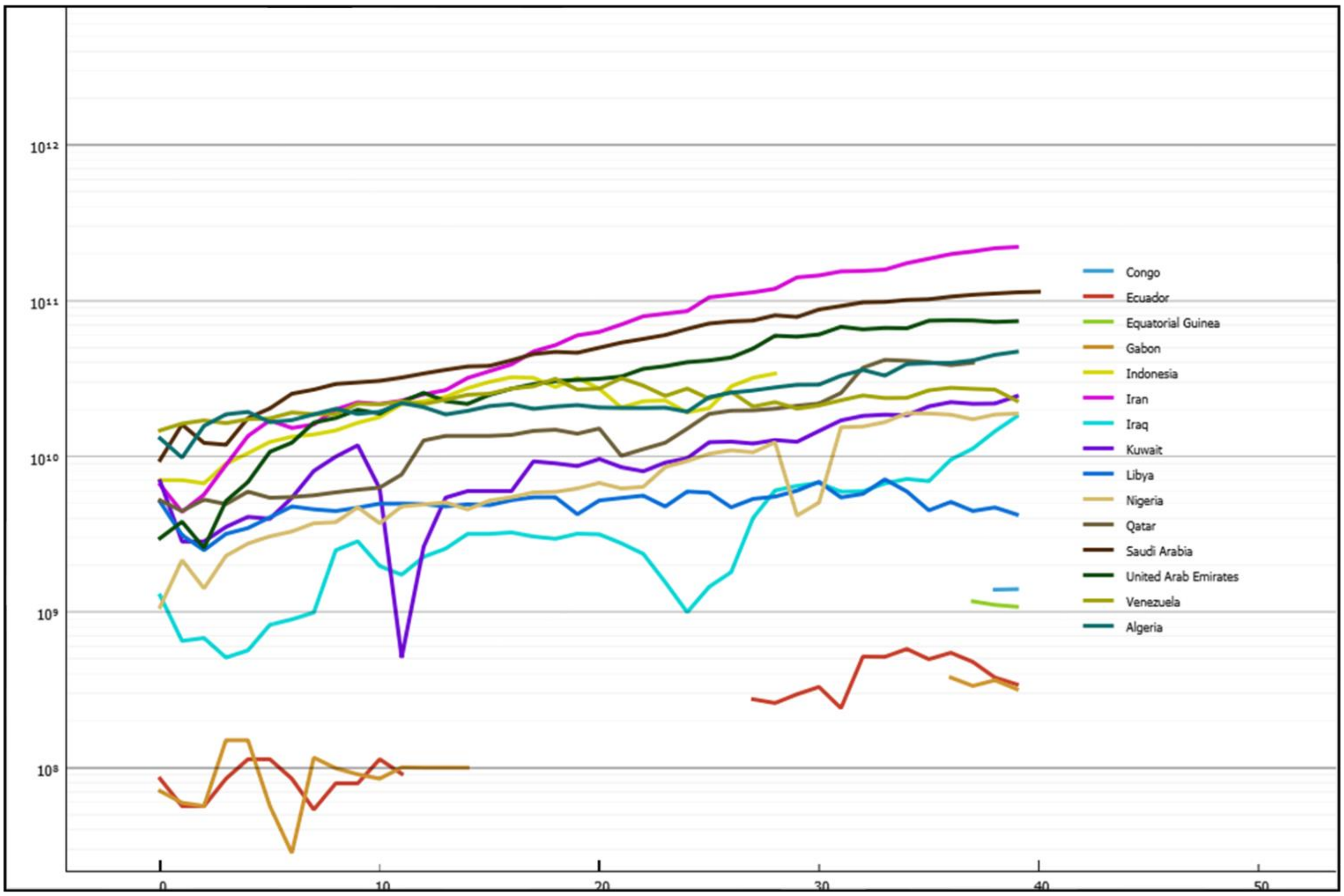


Figure 35. OPEC Gas Consumption Relative Graph

The gas consumption heat map showcases some interesting trends. China's consumption is close to Iran in terms of magnitude, and Russia has been consuming natural gas at an almost consistent rate since the fall of the Soviet Union. The United States seems to be transitioning over to the use of natural gas, and this is potentially a response to climate change, as the United States seems to be pivoting away from coal.

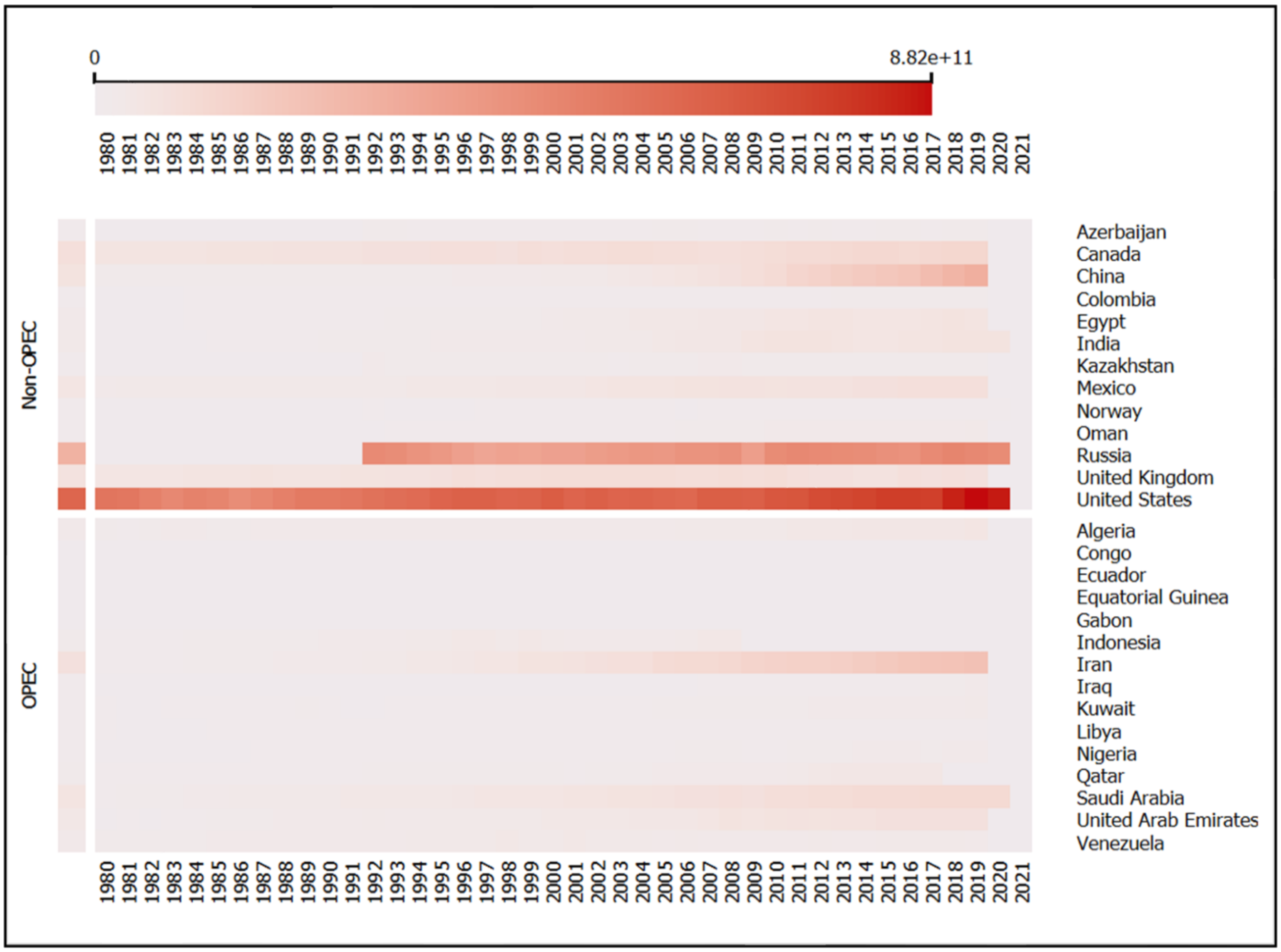


Figure 36. OPEC Gas Consumption Heat Map

Oil Production & Consumption Trends

Oil is the final fossil fuel being analyzed by this case study, and the trends are displayed in the following charts. For non-OPEC countries, the United Kingdom has produced more than the United States for most of the years included in this study, with the United States overtaking the United Kingdom for the highest producer in recent years. After checking the relative chart, it can be noted that Norway has decreased its production in recent years and Azerbaijan and Kazakhstan have increased their production since the 1990's.

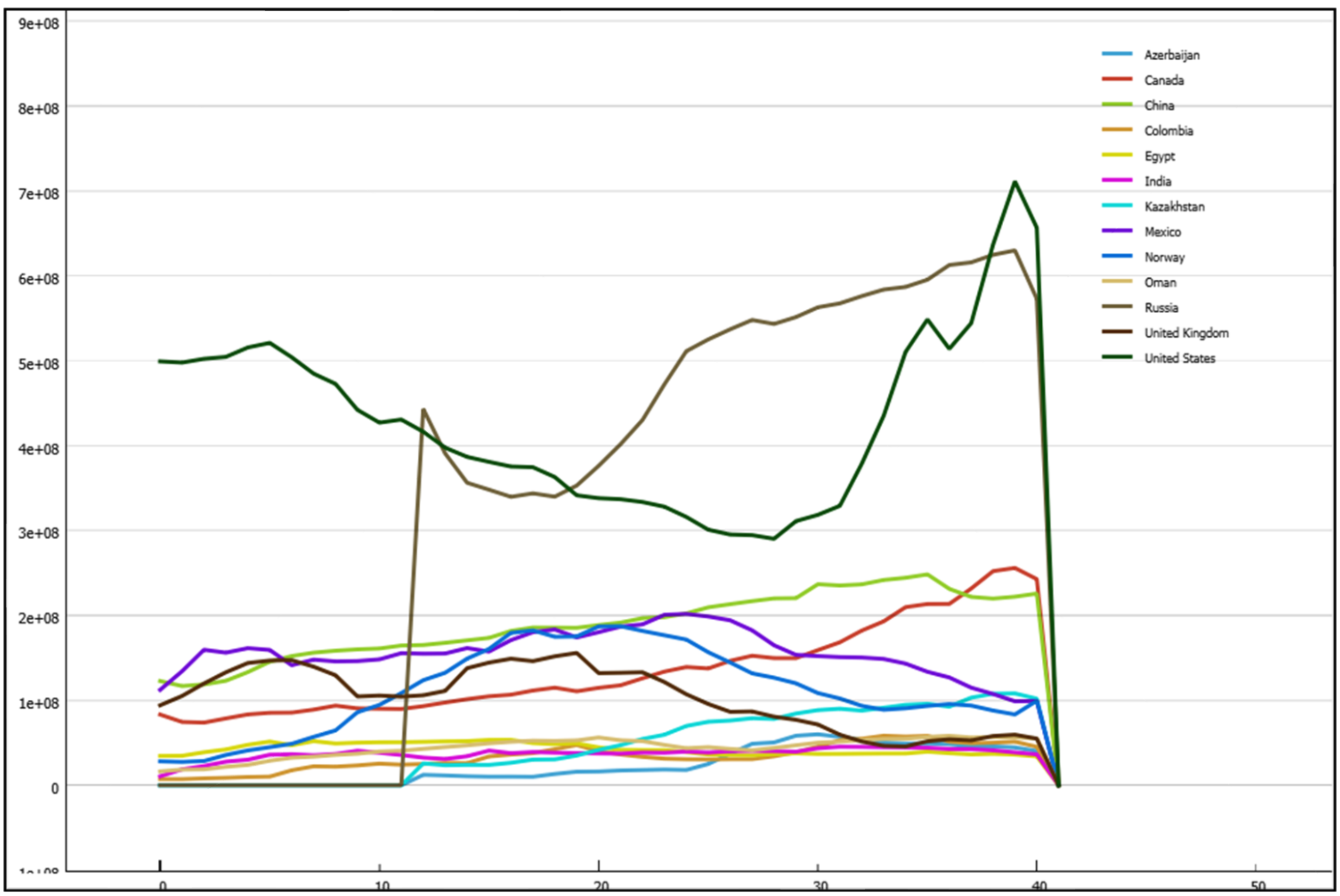


Figure 37. non-OPEC Oil Production Graph

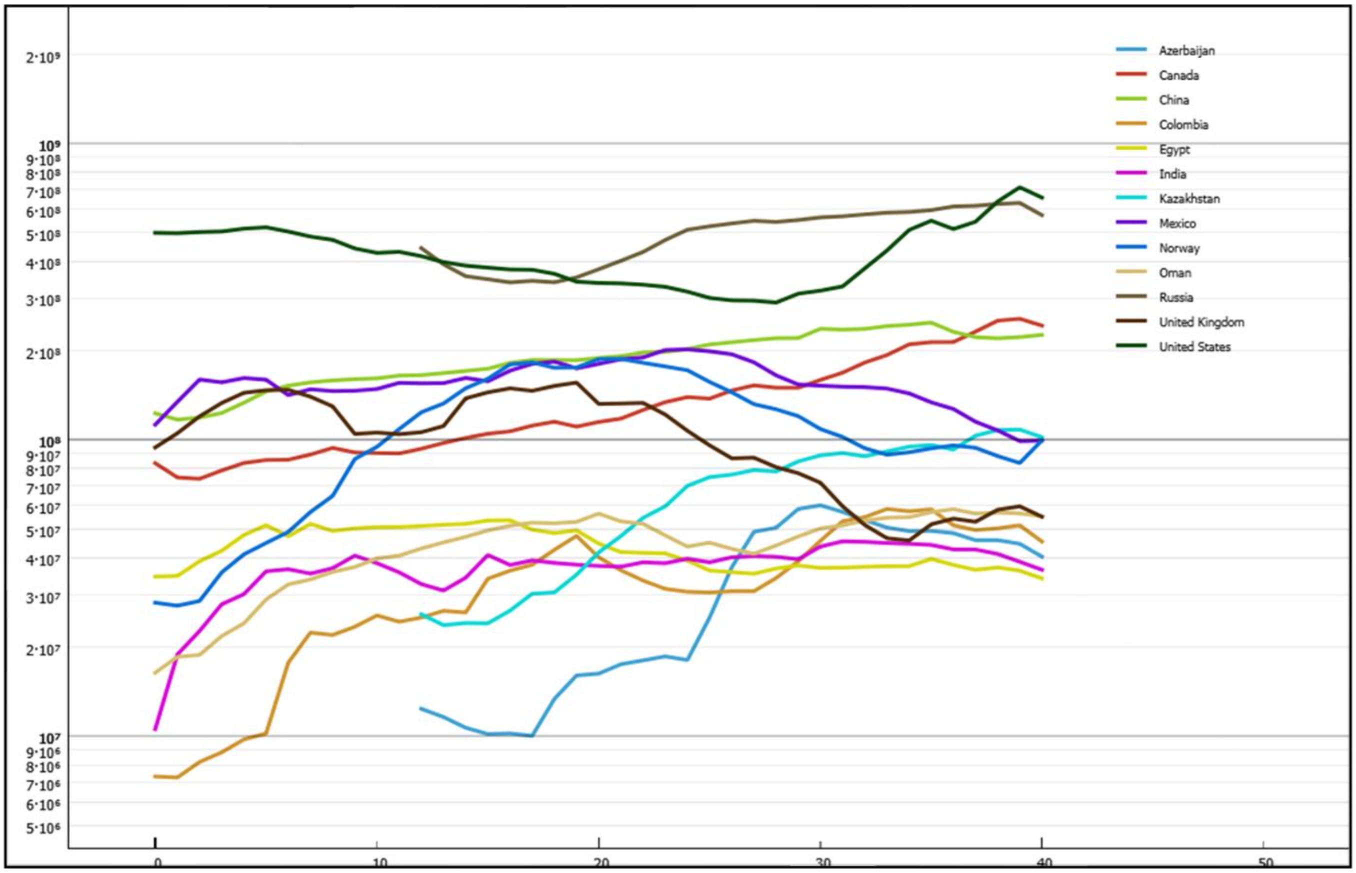


Figure 38. non-OPEC Oil Production Relative Graph

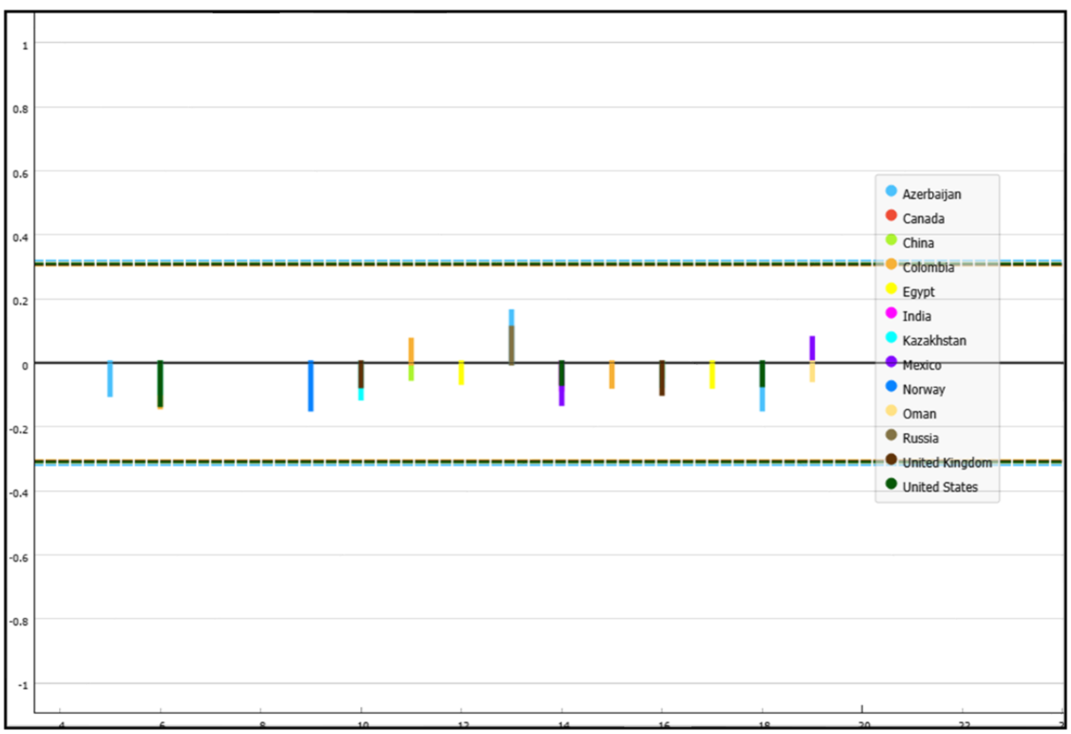


Figure 39. non-OPEC Oil Production Correlogram

Most countries have shown slight dips in production in 2020, likely due to the proliferation of the coronavirus disrupting international production. The non-OPEC correlogram does not contain any statistically significant correlations or nonrandom data.

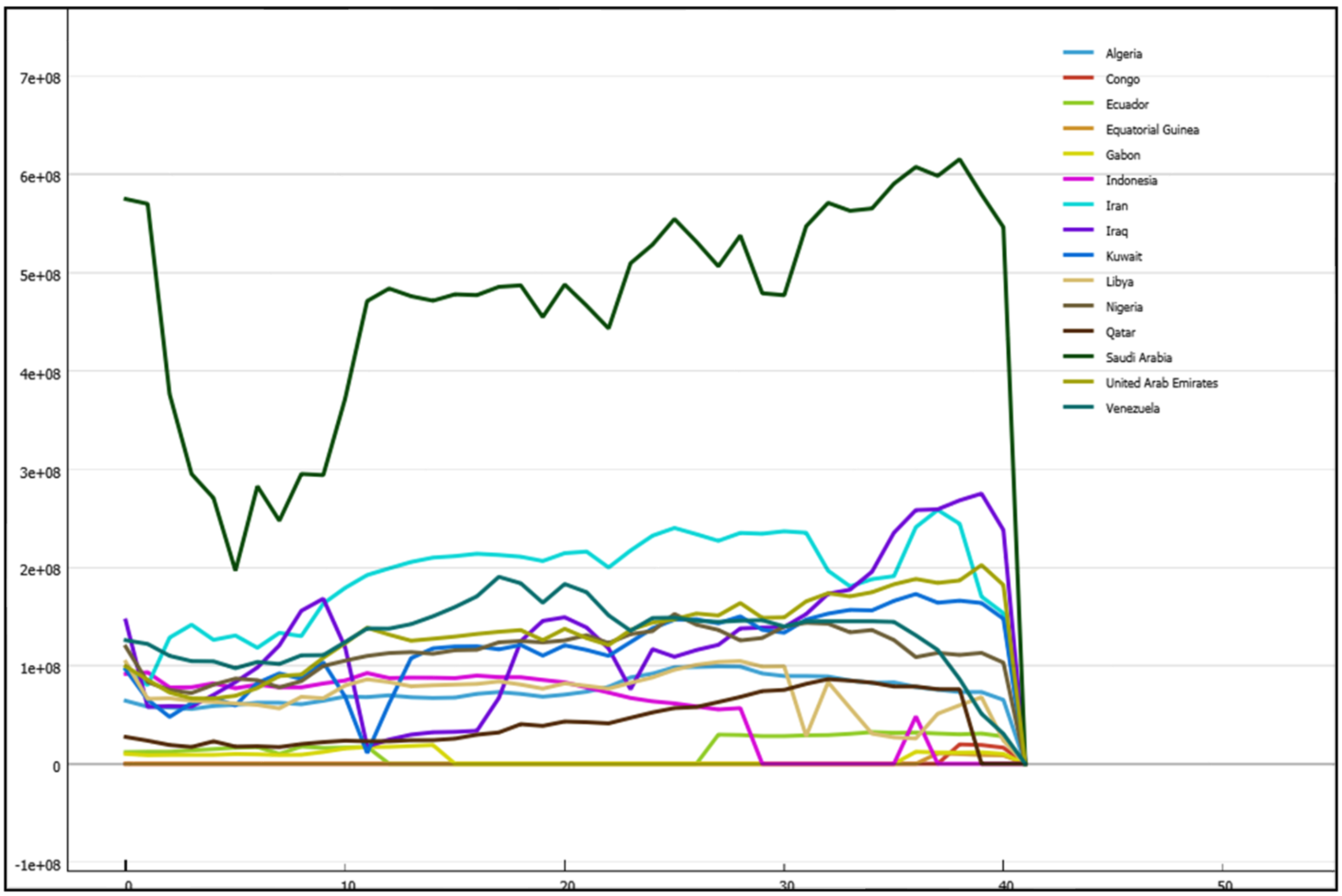


Figure 40. OPEC Oil Production Graph

For the OPEC oil production chart, it was noted that Saudi Arabia is the clear leader in terms of oil production and followed by Iran and Venezuela. The majority of OPEC members seem to produce similar amounts to one another. This may be due to OPEC being a coalition of petroleum producing countries, who may produce a minimum amount of fuel. Work can be done to analyze the requirements for OPEC entry and identify other countries that might meet the requirements for entry.

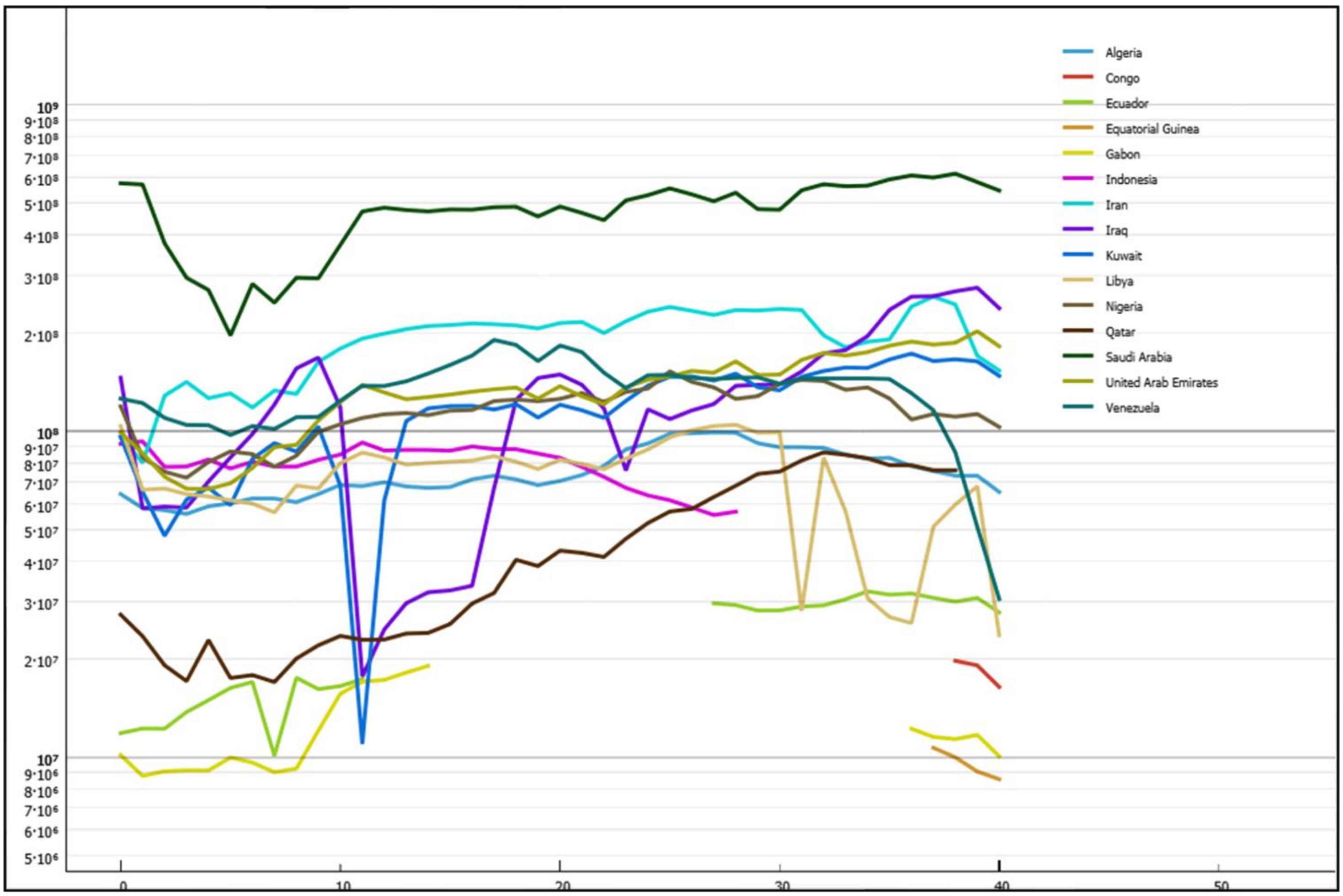


Figure 41. OPEC Oil Production Relative Graph

The OPEC producer correlogram shows that there are some non-random data, as well as some statistically significant correlations. Iran and Equatorial Guinea seem to be related based on their overlap in the chart, as well as Indonesia and Equatorial Guinea. Ecuador and Gabon also seem to share some correlations.

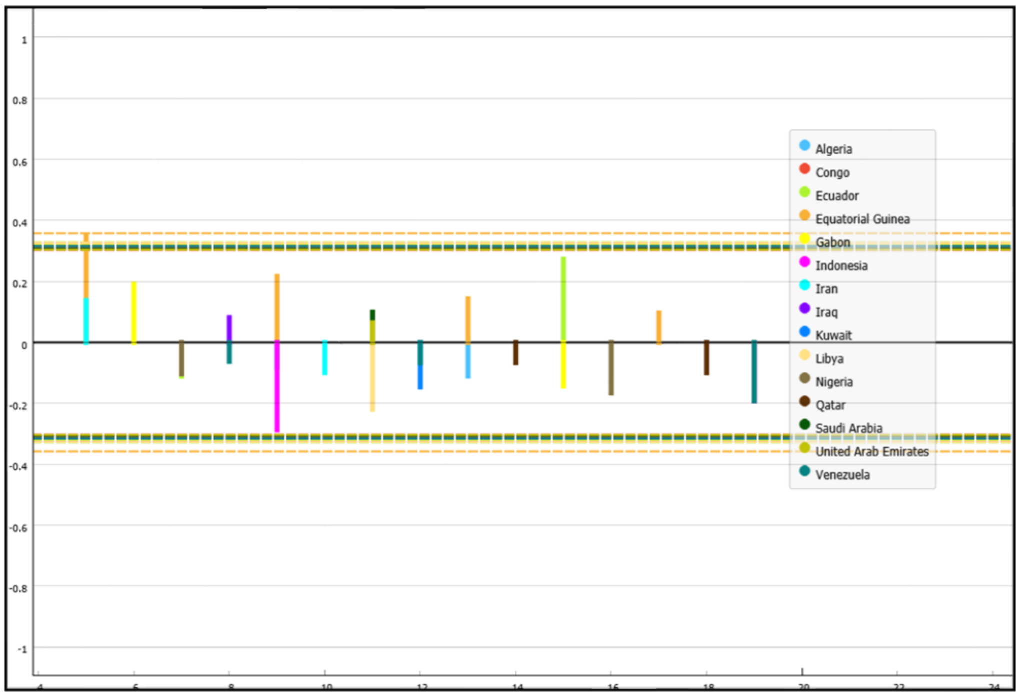


Figure 42. OPEC Oil Production Correlogram

The heat map produced from the oil production data reaffirmed the observation that OPEC countries have a greater focus on oil production than any other fossil fuels. When comparing the two groups, it seems that the United States, Saudi Arabia, and Russia are the three most prominent oil producers in the world, with a greater proportion of OPEC nations following suit and producing copious amounts of oil. China and Canada also produce substantial amounts of oil; however, India seems to not produce much compared to their other fossil fuel production habits.

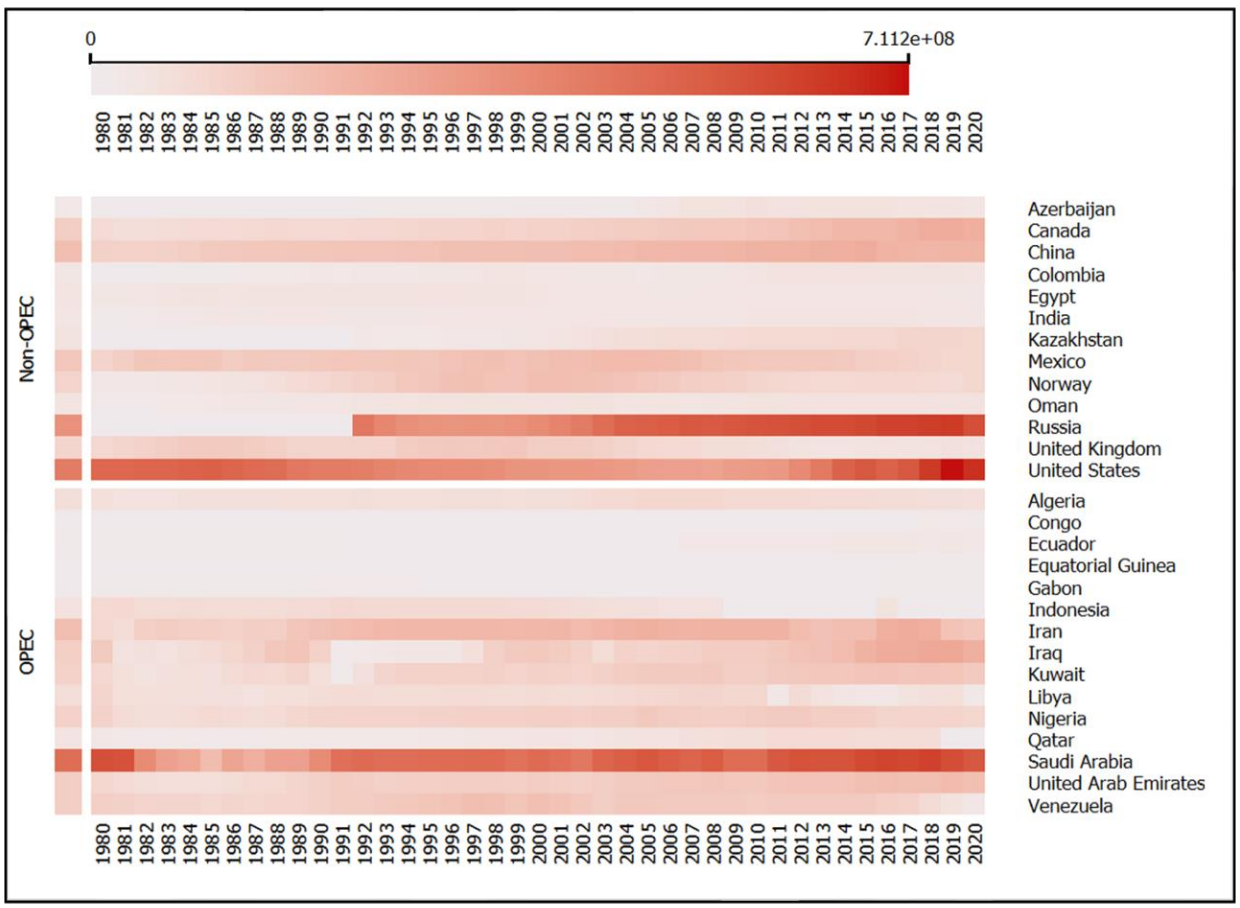


Figure 43. Oil Production Heat Map

For oil consumption, there seems to be a clear dominant force: the United States. By a wide margin, this is the country using the most oil followed by China who has been rising over time and is now the second highest oil consumer. The relative chart shows that Azerbaijan is steadily lowering their usage of oil, and Oman is ramping up their usage. The correlogram for non-OPEC users displays some potentially statistically significant correlations between Canada and the United States as well as Russia and Mexico. No nonrandom data seems to exist in this dataset.

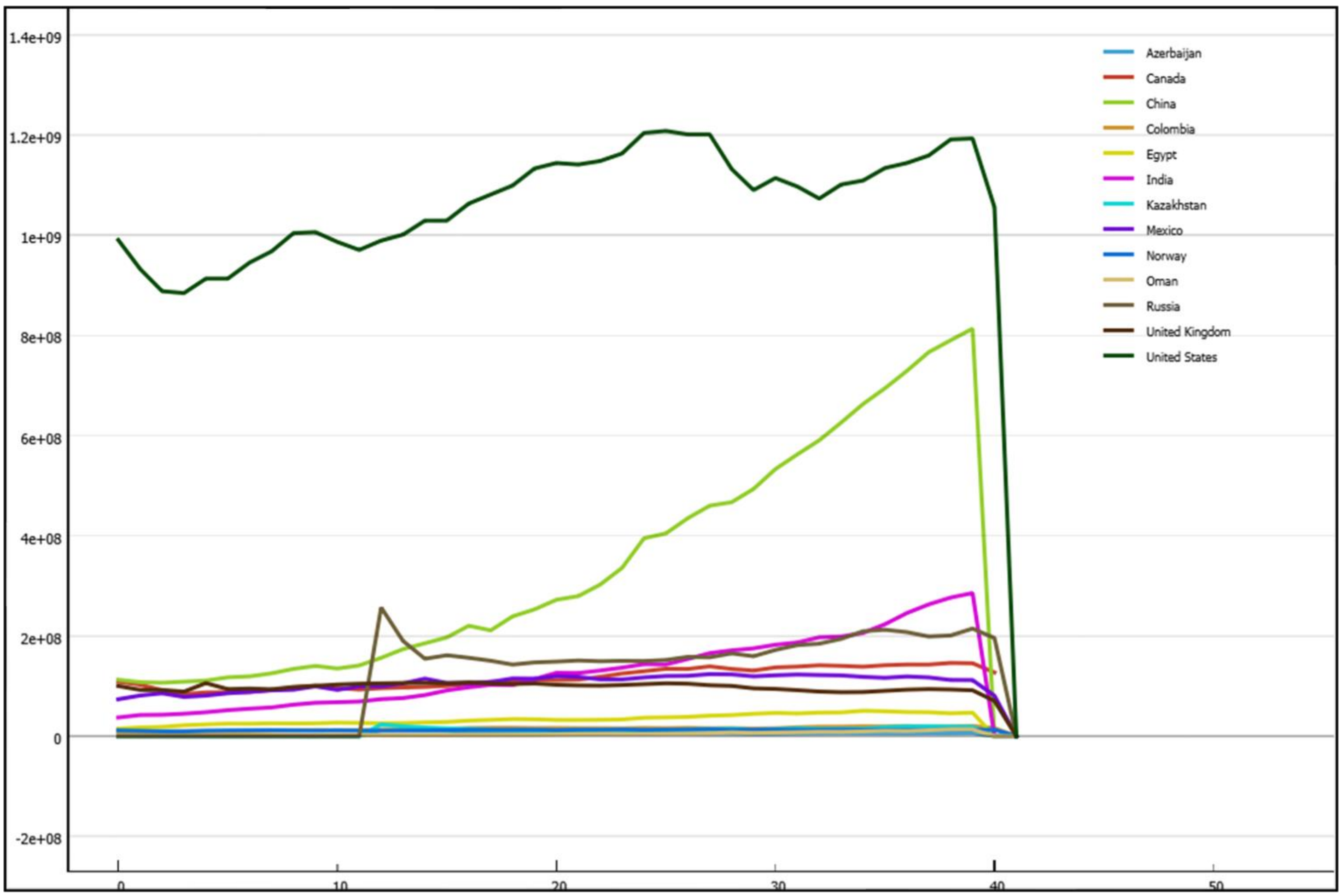


Figure 44. non-OPEC Oil Consumption Graph

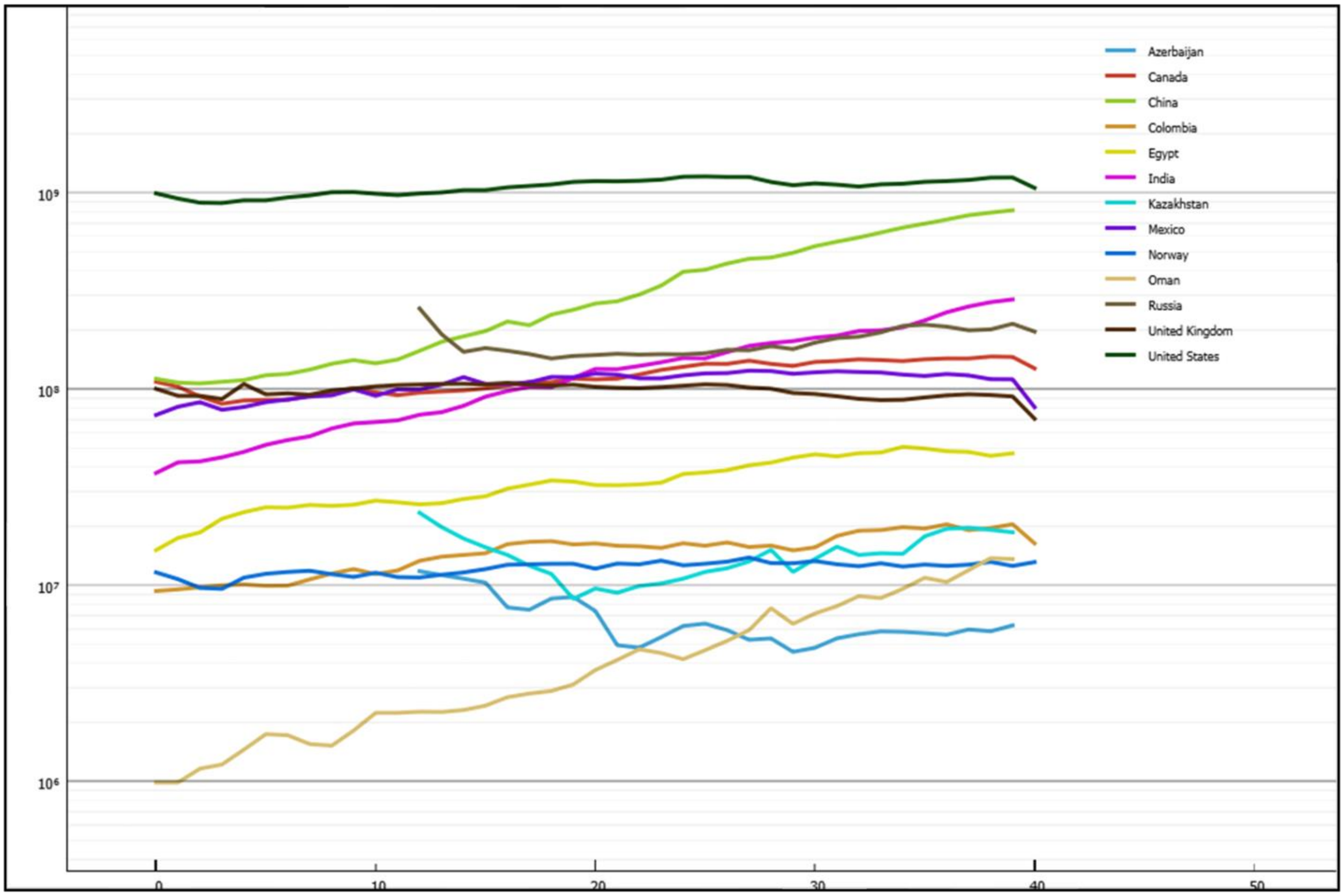


Figure 45. non-OPEC Oil Consumption relative Graph

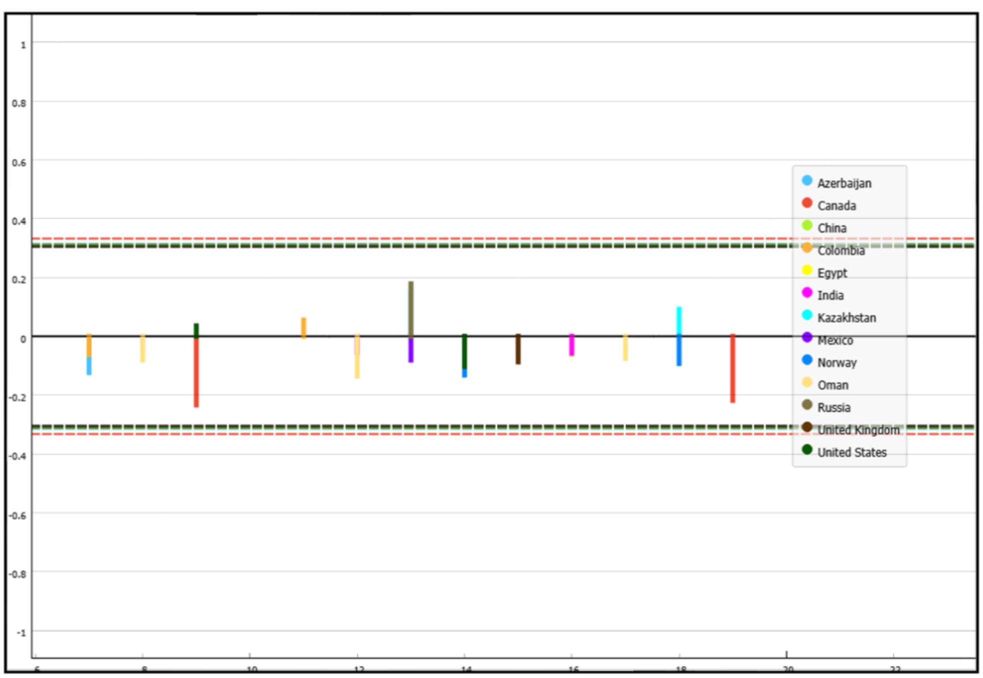


Figure 46. non-OPEC Oil Consumption Correlogram

Oil consumption by OPEC countries seems to be, in a similar fashion, dominated by a single country: Saudi Arabia. Iran is the next largest consumer followed by Indonesia. Note that there are major holes in the data for Indonesia, this is another artifact in the charts due to the removal of the data associated with them when the country left OPEC, returned for a brief period of time in 2016, and left the same year. Consulting the relative chart provides a more robust view of the data, with Qatar ramping up their consumption, and Ecuador having a large hole in their data due to leaving OPEC. The final correlogram highlights some statistically significant correlations between Ecuador, Iraq, and Gabon. Libya shows some nonrandom trends, as well as Indonesia, Gabon and Nigeria having potentially significant nonrandom data.

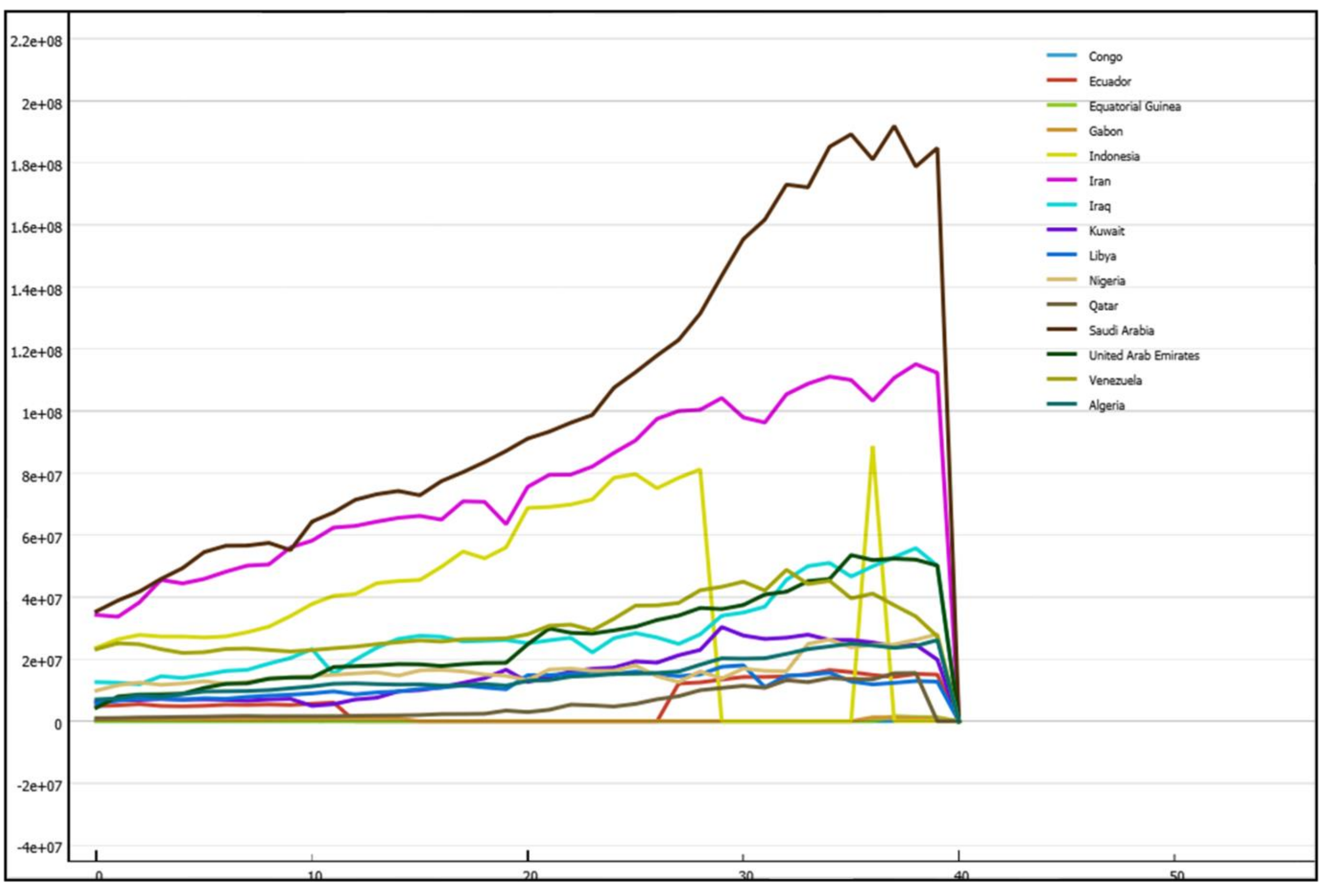


Figure 47. OPEC Oil Consumption Graph

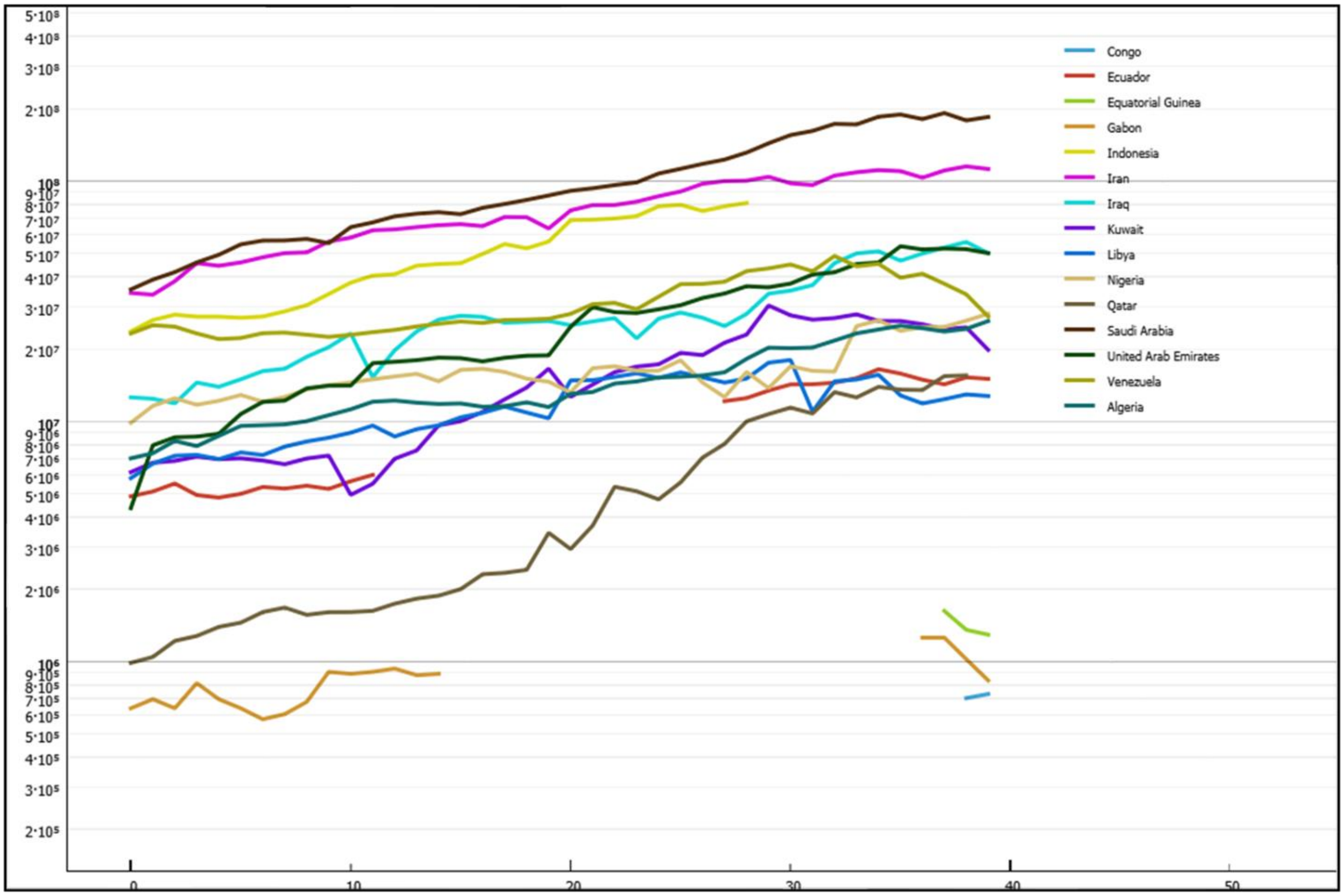


Figure 48. OPEC Oil Consumption relative Graph

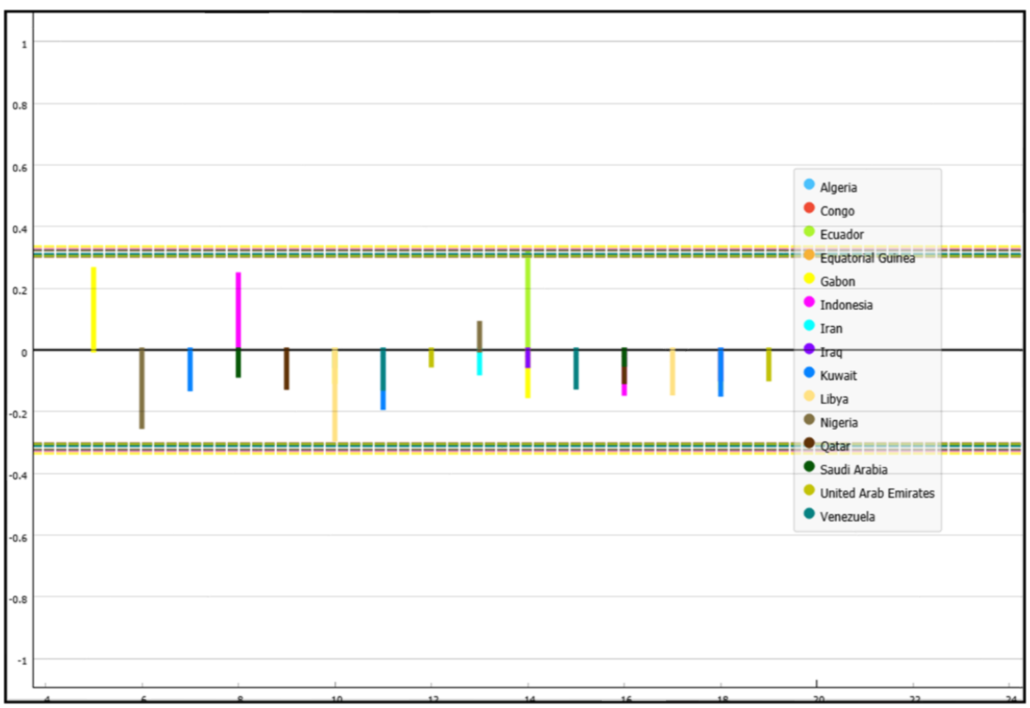


Figure 49. OPEC Oil Consumption Correlogram

The heat map displayed just how dominant the oil consumption of the United States was. China is the next largest consumer worldwide, however some of the lesser consumers completely eclipse the entirety of the OPEC countries. These observations can also be used to research the worldwide trade trends and to make predictive observations such as China seems to be expanding and ramping up its' energy usage to compete (and possibly overtake) other large superpowers.

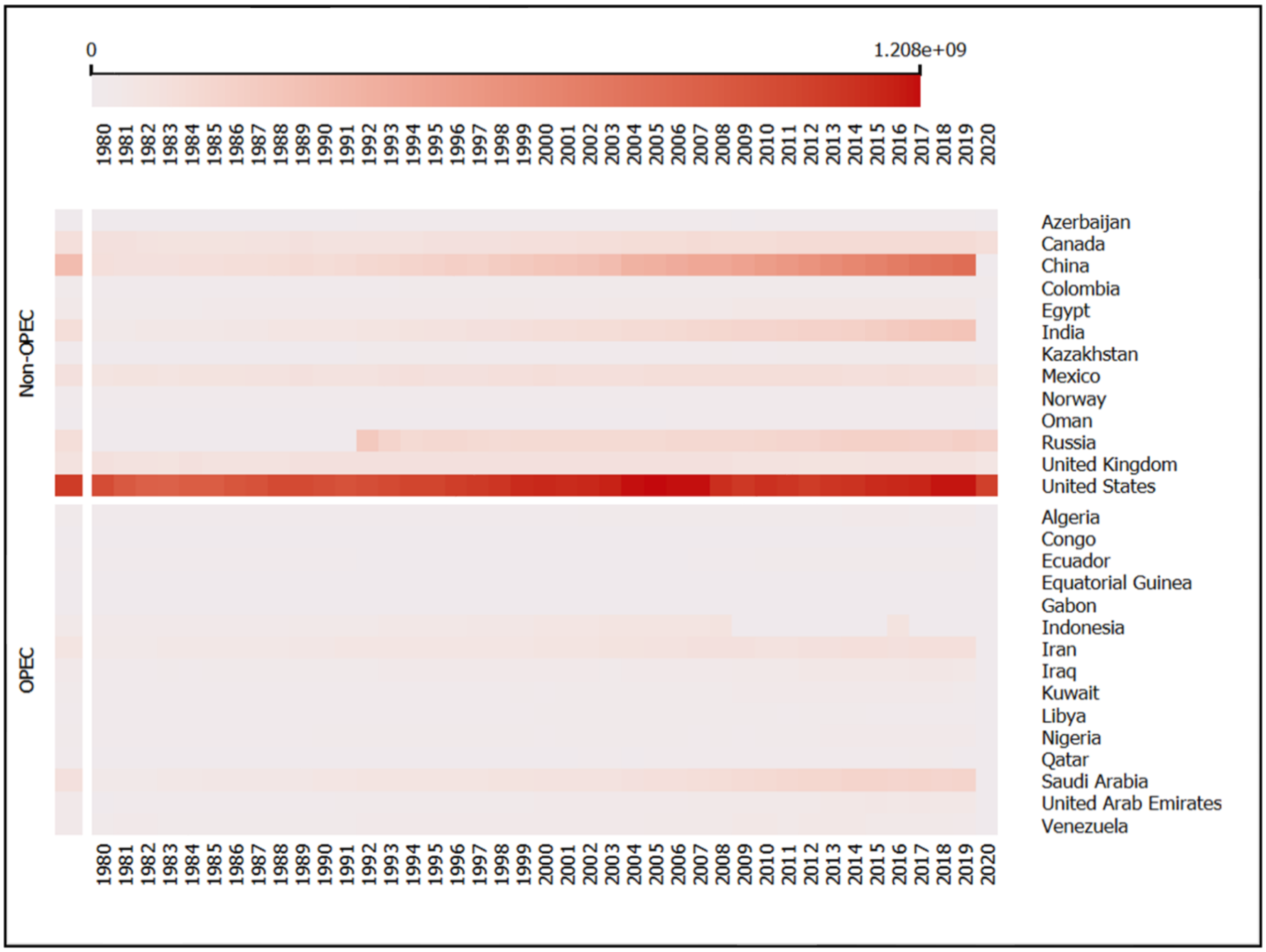


Figure 50. Oil Consumption Heat Map

Integration into GIS & spatial trends

After the data mining software processed the data, a final export of mined data was organized for use in a GIS. In this use case, the GIS acted as an external widget, and leveraged data for display in a similar manner as the graphs produced earlier in Orange. These data were linked to a shapefile sourced from Esri's ArcGIS Online open-sourced data to provide a spatial context. This map showcases the geographic extent of the countries associated with the data included in this study. The idea here is to create a map for each decade (1980's, 1990's, 2000's & 2010's) for each fossil fuel consumption and production dataset, as well as population and worldwide GDP. Each map leverages the visualization power of GIS to showcase the mean statistic and standard deviation for each country by decade. This initial map highlights the countries included in this study, and was intended to provide a visual reference, as some of the bar charts do get cluttered, and require referencing.

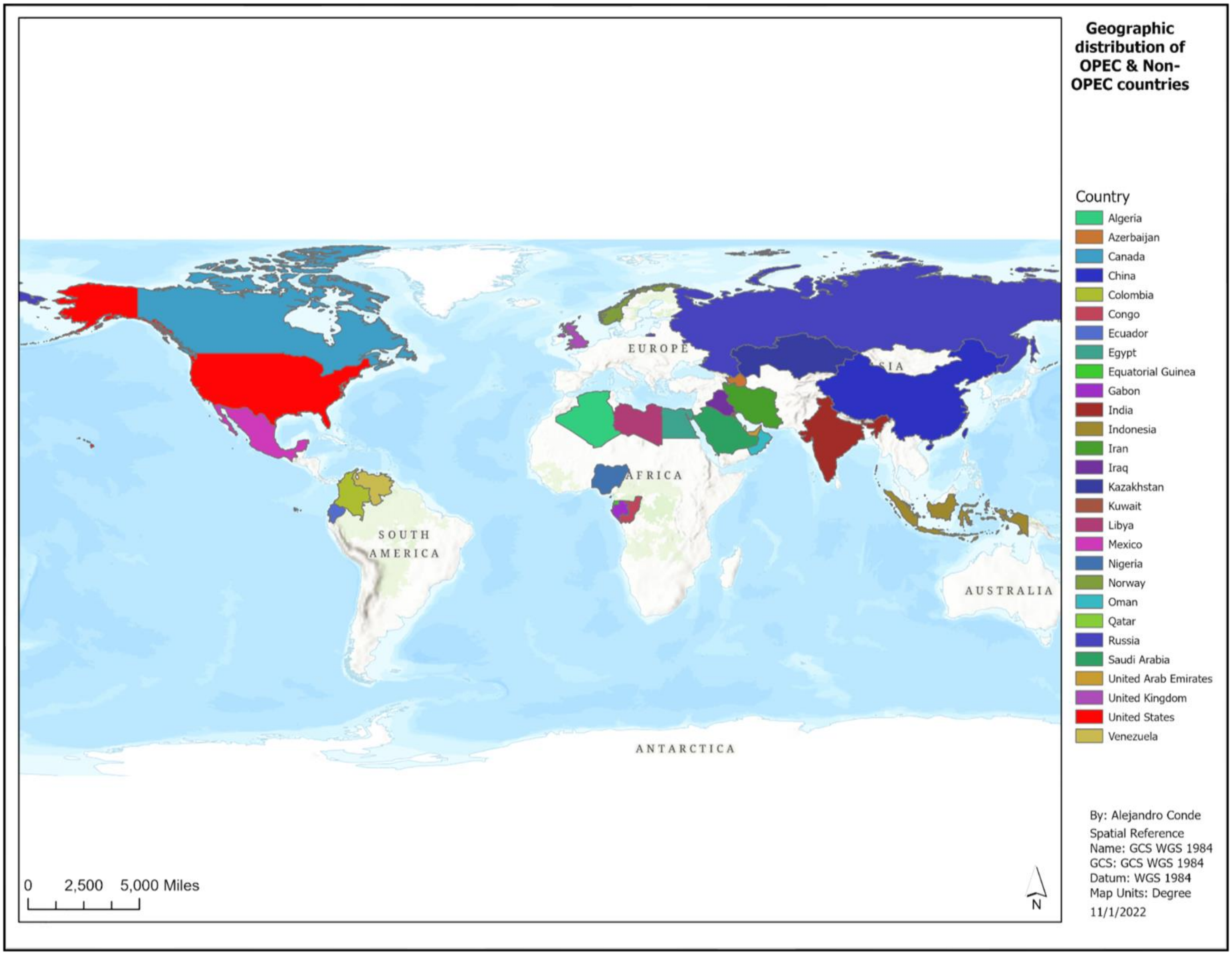


Figure 51. Geographic distribution of Countries included in study

Coal Production & Consumption

GIS is a technology which permits for a sophisticated geography-based data analysis. In this series of maps, OPEC & non-OPEC countries were categorized to provide a high-level OPEC related context to the data. The mean values as well as the standard deviation for each subcategory (coal consumption, coal production, etc.) are displayed, and used to showcase how these data transform temporally. This graph allows a user to visually estimate the average values for the data being represented and the magnitude of change seen throughout the decade, while drawing high level conclusions. For example, Columbia lowered the mean coal production from the 2000’s to the 2010’s, with a larger standard deviation in the 2010’s. This suggests that they’ve lowered their coal production in the last decade, and the shift was both large and relatively quick, as a slow change would be spread out over a few decades.

For coal consumption, it seems many countries have reduced their usage of the fossil fuel, likely due to it being considered a “dirty” fuel with a much higher concentration of carbon produced. Interestingly, the United States seems to have lowered their usage in the last decade and Russia seems to have raised their usage since dissolving the Soviet Union in the early 1990’s. China and India remain some of the highest users, with smaller standard deviations suggesting that their usage is consistent. Norway, on the other hand, seems to have lowered their usage considerably, and they have done so in the last decade. Egypt, Iran, Iraq are others are following suit. However, Ecuador has raised their coal usage considerably in recent years after a large reduction in the 2000’s.

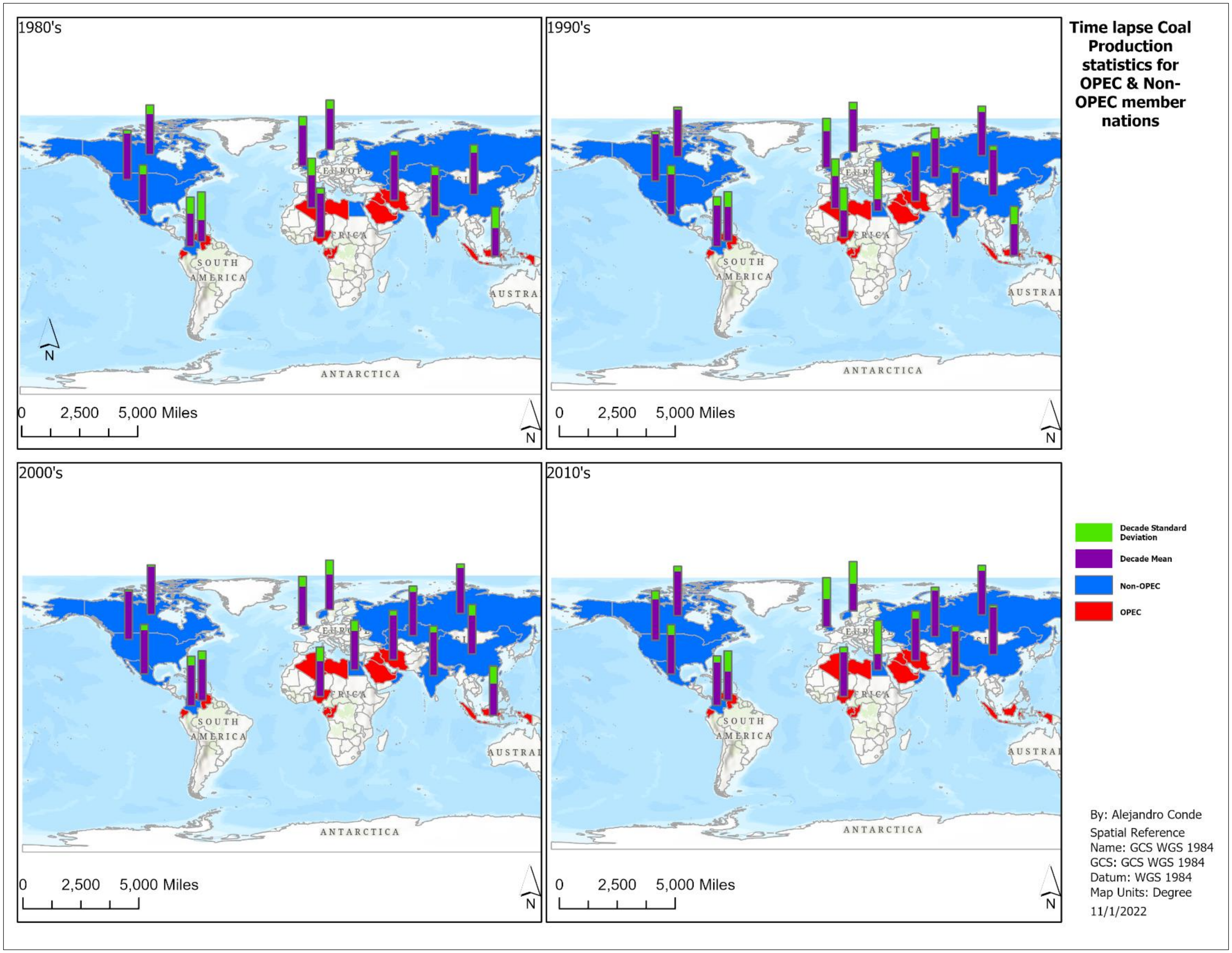


Figure 52. Time lapse Coal Production Map

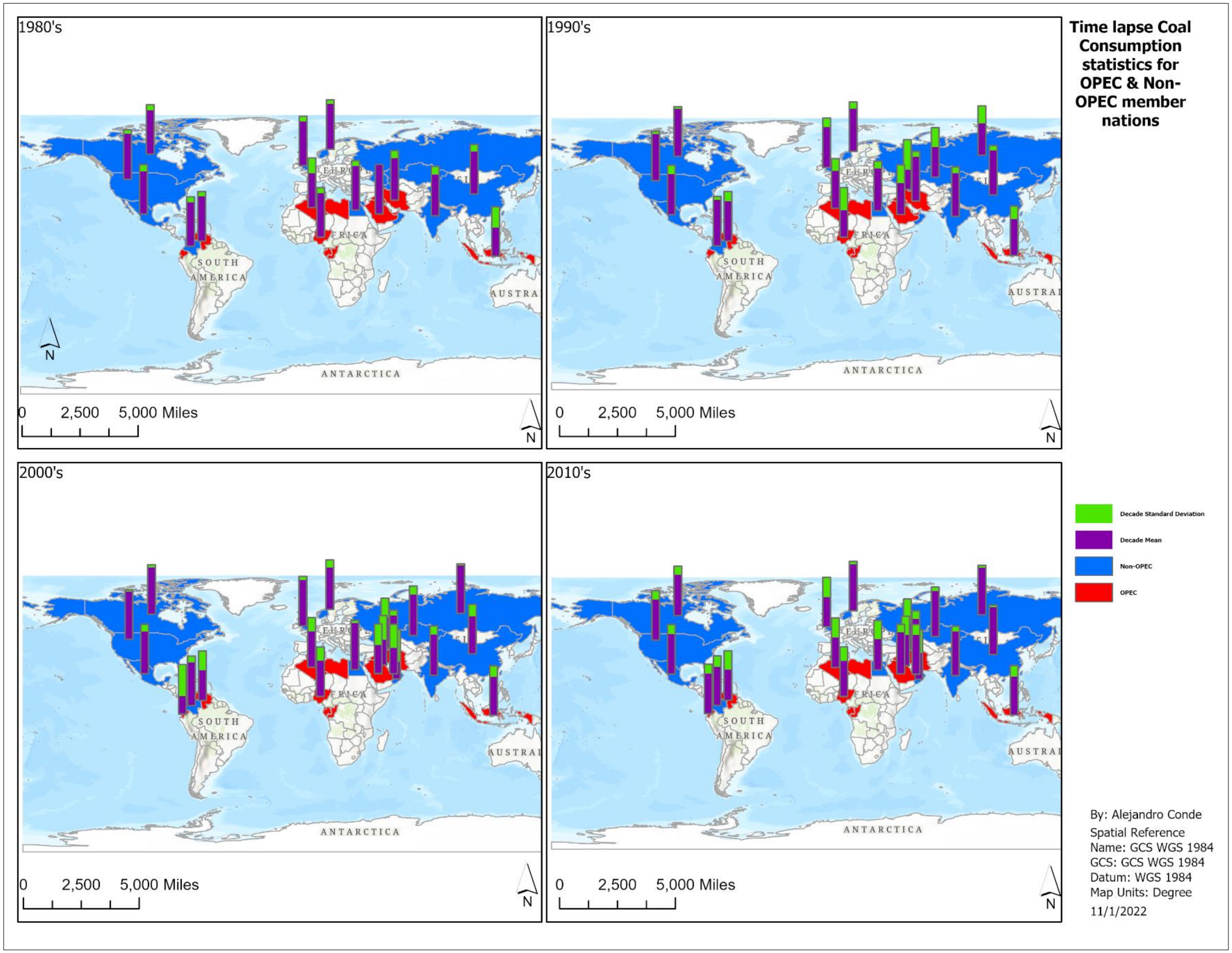


Figure 53. Time lapse Coal Consumption Map

Gas Production & Consumption

Gas production shows that Russia has raised their production since a steep drop in the 1990's. The data for Gabon are missing during the 2000's, as this coincides with a period on non-OPEC membership. The same can be said for Indonesia, where the data for the 2010's is missing because this country left OPEC as well. The trends seem to fluctuate the most in either South America, Africa, or the Middle East. This may be because of political turmoil or war (such as the United States' presence in the Middle East). India has raised their production since the 1980's and has remained a steady natural gas producer since. On the other hand, gas consumption seems to have increased and has remained steadily high since the 2010's (with some exceptions of course). The United States has remained a high consumer throughout this forty-year period. Russia seems to have remained a high consumer of gas, once again with data missing before the dissolution of the Soviet Union. Ecuador has seriously ramped up their gas consumption from the 2000's to the 2010's, and this is a possible avenue for further research.

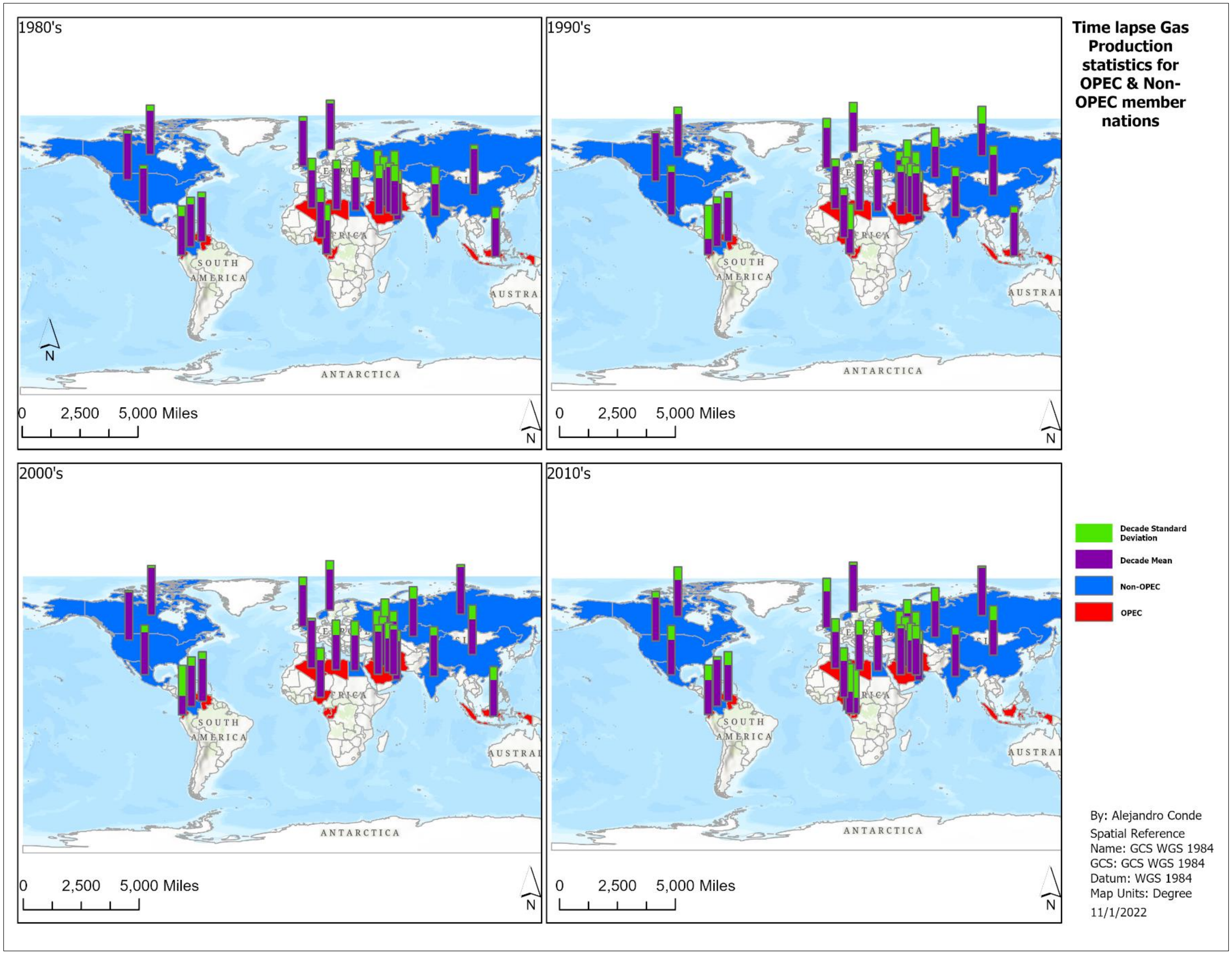


Figure 54. Time lapse Gas Production Map

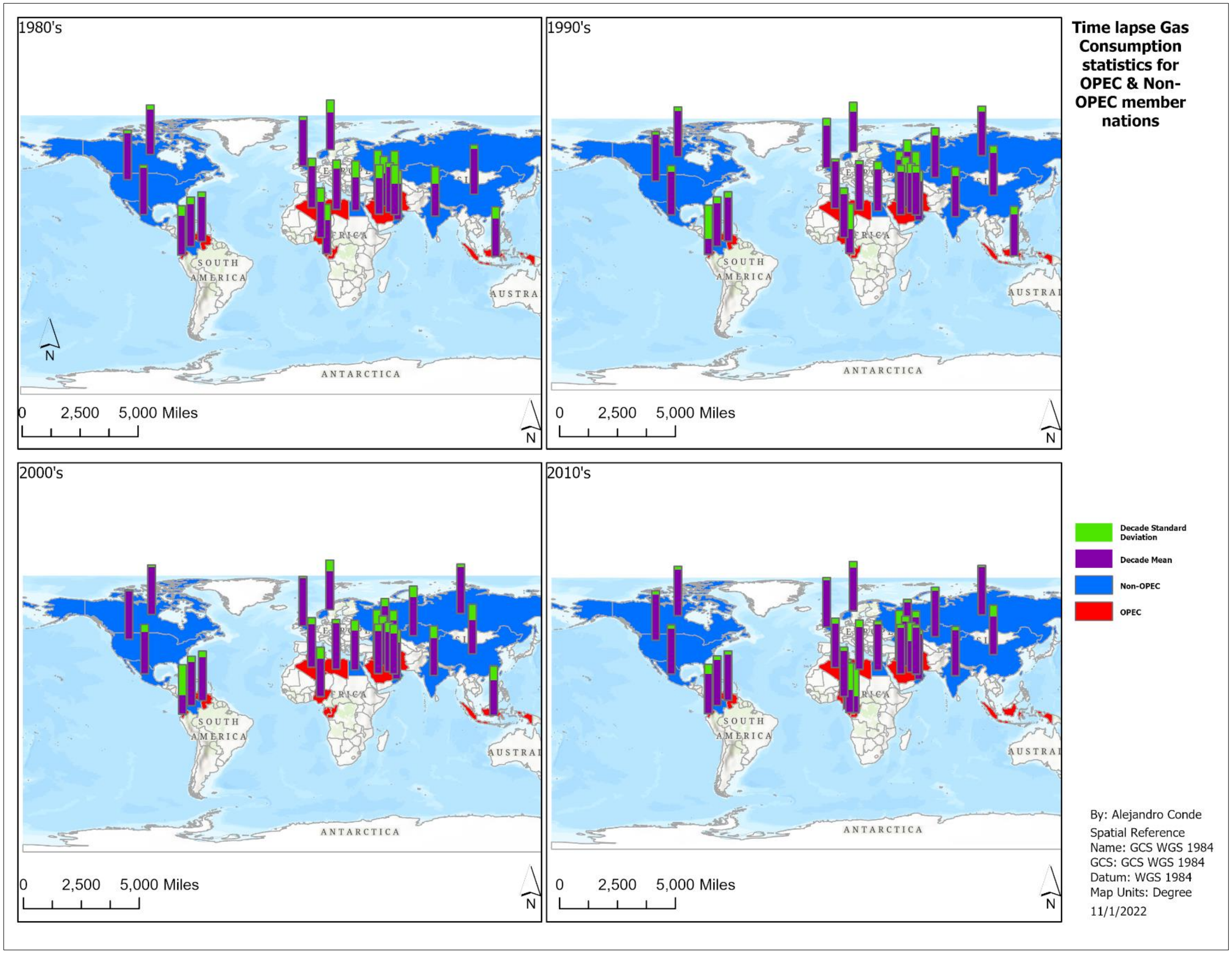


Figure 55. Time lapse Gas Consumption Map

Oil Production & Consumption

A high-level trend regarding oil production is that most countries seem to be increasing or maintaining production. This may be because of its' value as both a fuel and its use in other petroleum products. Ecuador lowered its production during the 1990's and 2000's, but has significantly ramped up production during the 2010's. Most of the Middle East and Africa seem to have raised their production since the 1980's and have remained consistently high producers. However, countries like the Congo, and Indonesia have greatly decreased their output in recent years (2010's onward).

Oil consumption displays a similar trend with respect to Ecuador with two decades of chaotic shifts and generally lower consumption followed by rapid expansion of steadier use. North American countries are consistently high in their use, with most of the Middle East and Africa also being large users. Note Indonesia's drop in usage was due to their exit from OPEC, not from significantly lower usage. Gabon has seemed to lower their usage significantly in the last decade, and the Congo only had recent data on their consumption. This number is overall low but has fluctuated slightly, causing a massive shift in its standard deviation due to how small the sample size is.

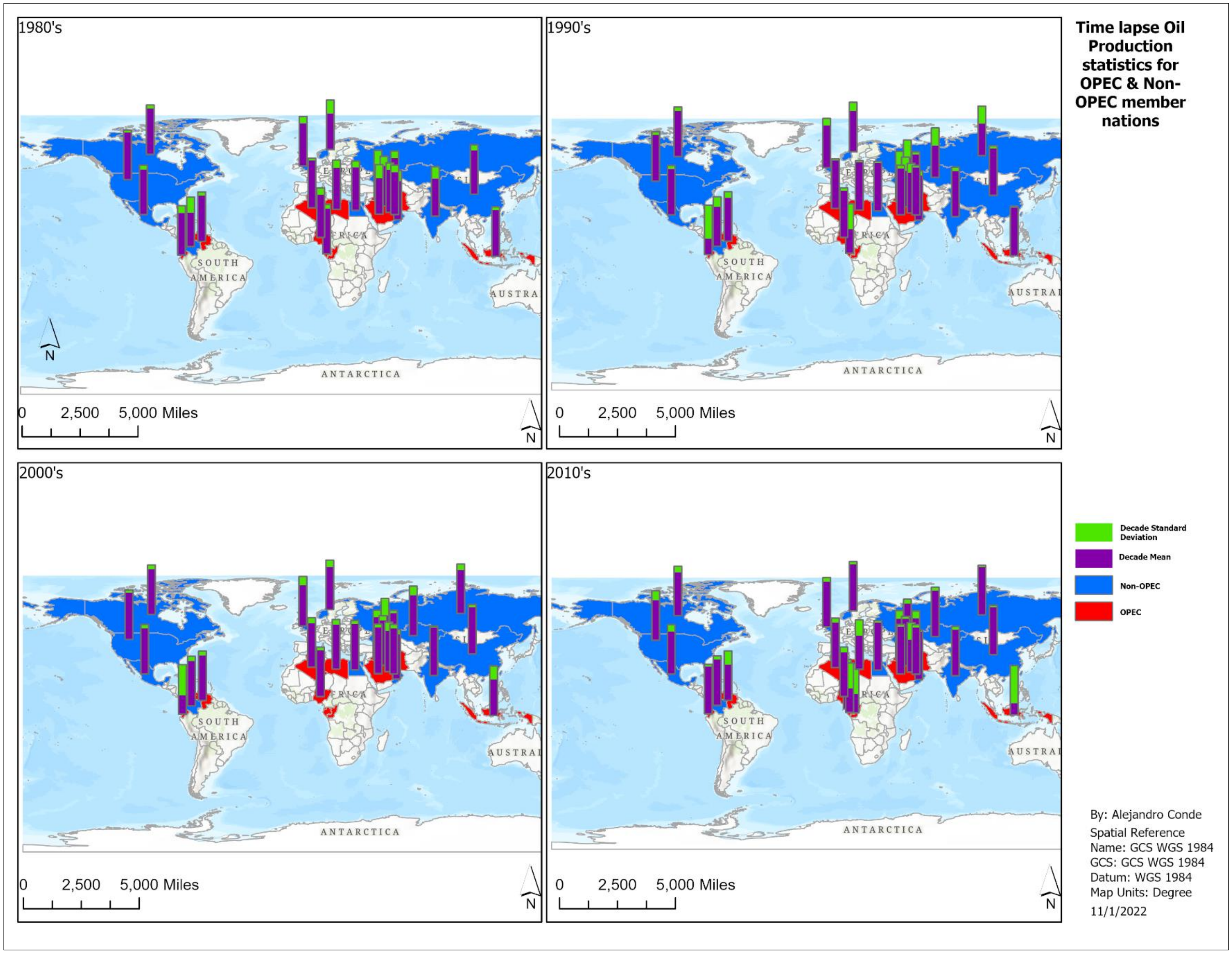


Figure 56. Time lapse Oil Production Map

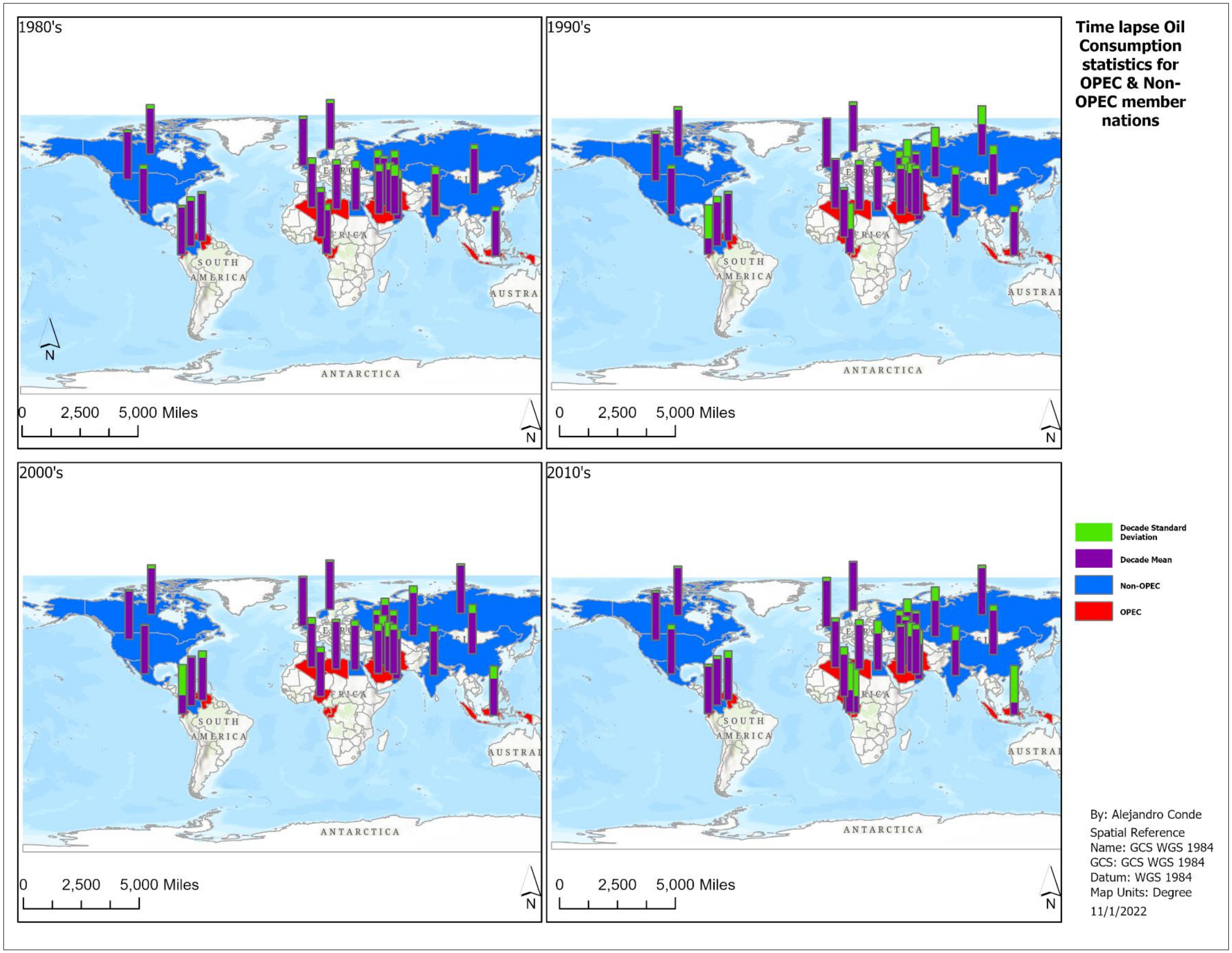


Figure 57. Time lapse Oil Consumption Map

Gross Domestic Product (GDP) & Population

Gross domestic product is highlighted in this map, displaying the fluctuations in the average total of the monetary value of all final goods and services produced in each country. On a high level, almost all countries have seen an increase in mean GDP plus an increase in stability (from the decrease in the magnitude of standard deviations). Russia data during the 1980's, and these figures suggest a high level of instability and very small GDP. This was confirmed in the bar graphs; however, this may be due to deception or isolation with respect to the Soviet Union. This information showcases how the flow of money has fluctuated in each country but does not indicate how dependent each countries monetary resources are on fossil fuels, and additional research could be applied to integrate the percentage of GDP related to fossil fuels to provide more context as to the impact of fossil fuels on these countries trade.

Population data are vital for this study, as it provides a baseline context into why the production values are so high. Population is related to production because larger populations require more fuel for transportation, heating, petroleum products, etc. It is likely not as straightforward as a country producing only what it needs; many countries with an excess trade fossil fuel or stockpile their surplus for future use. This question of tracking how population and need relate to a countries production, as well as comparing available trade information will provide further context into GDP and population as it relates to fossil fuel trends and will provide another avenue for further research.

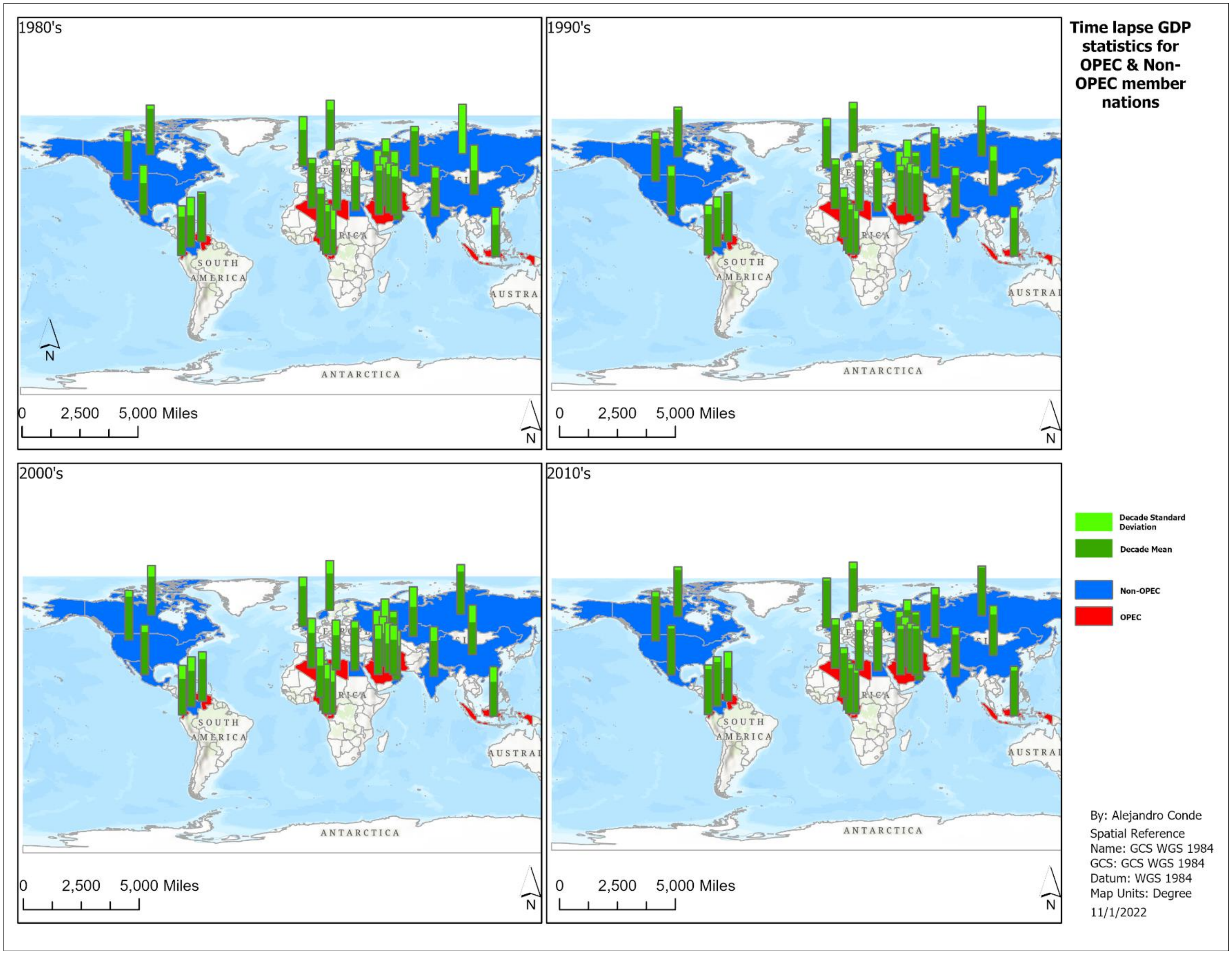


Figure 58. Time lapse Gross Domestic Product Map

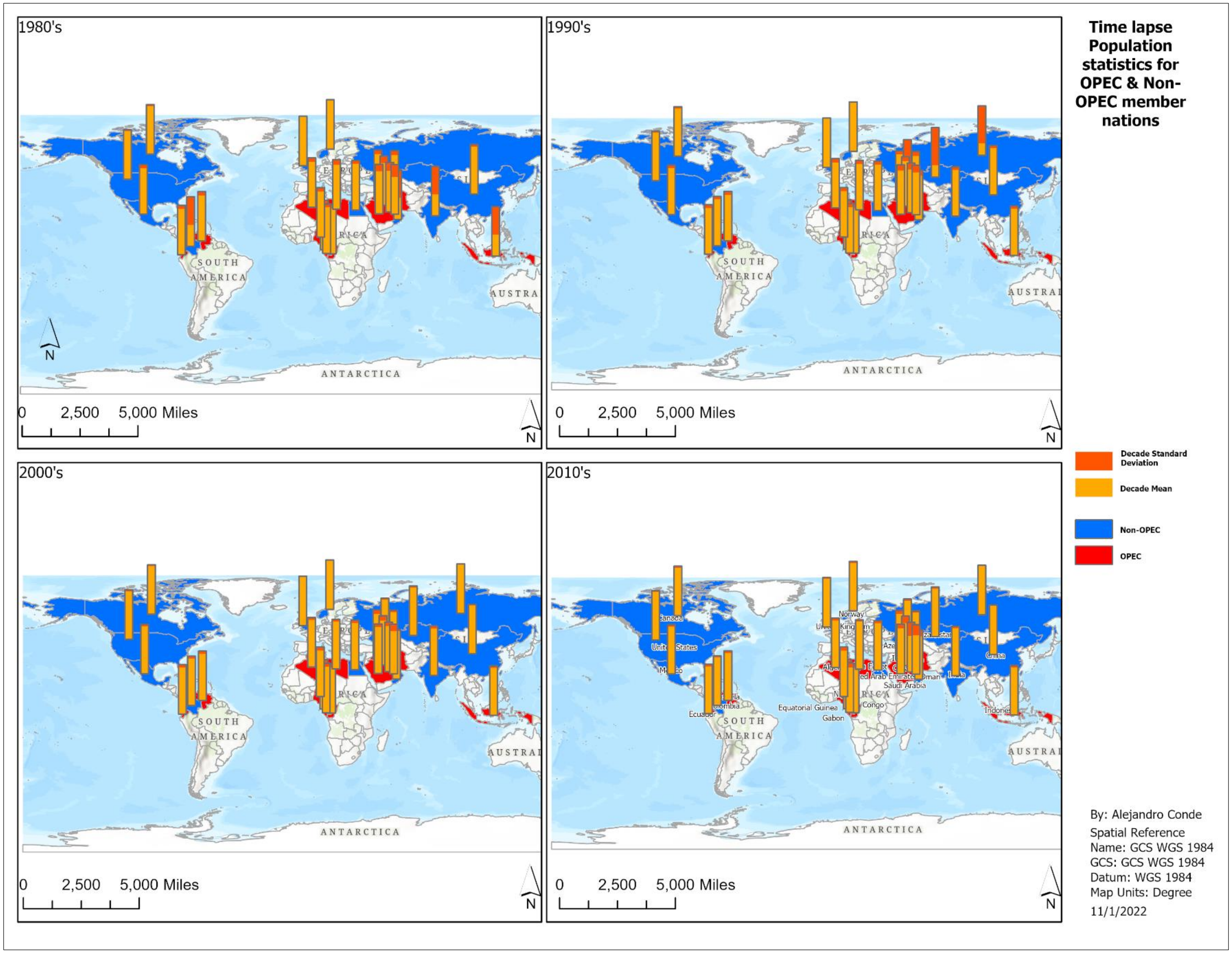


Figure 59. Time lapse Total Population Map

Discussion

Throughout the exported data, trends have appeared that highlight just how interconnected the data are; this is sensible, considering that global trade is a major driver behind production and consumption of fossil fuels. As for correlating OPEC and non-OPEC producers, it is not so straightforward to correlate these countries with overall consumption at such a high level. Each country prefers their production type, such as China and India with respect to coal and Saudi Arabia and the United States with Oil. So, on this level, it is relatively easy to identify the trends and connect them to a country.

This case study demonstrates the high-level analytical capabilities of applying data mining to large datasets. One of the key issues encountered here is an almost cascading effect: as we refine data, create connections, and draw conclusions, more questions inevitably arise alongside more opportunities for analysis. In this study, certain kinds of analysis were omitted (mainly machine learning based approaches or AI applications supported by Orange). These two approaches are powerful, however at the level of analysis leveraged for this study, there were few opportunities to refine the data further with the express purpose of applying these deep learning capabilities and search each possible avenue in the data. This deep learning is the next step in this analysis but was not considered appropriate considering the high-level analysis being performed.

On a lower-level, subtle trends can be identified by means of further refining data (by isolating data based on individual countries and not by dividing them into two large groups) and additionally processing this refined and grouped data into such a manner that the machine learning capabilities of Orange can be utilized to identify trends in the data that would not be obvious at the level utilized in this study. In order to robustly apply data mining to produce undetected conclusions (rather than the general observations present in this study) from a series of massive datasets, lower-level observations will need to be made and will influence the direction in the research. This will allow future data analysts to draw more insightful conclusions that might have been lost in the mélange of such massive series of datasets.

Areas of Further Research

In this study, more data was processed than could be handled within the scope of this study, and many avenues of additional research were mentioned throughout. As this data is refined and structured, more questions arise regarding how these data interact with one another. As follows the avenues identified within this study are as follows:

- How has each country responded to climate change in recent years, and how is this impacting the production and consumption of fossil fuels?
- How does trade affect the production and consumption of each country? For each country, which trade companion affects their production and consumption trends the most or least?
- Which years showed the largest and smallest shift in the use of fossil fuels?
- What portion of GDP is related to fossil fuels?
- Why do the observed trends in this study exist? Do the trade interactions account for more variance in the data, and if not, what does?
- Regression analysis can be run on each country, the years, production, consumption, GDP, etc. Each variable can be compared to one another to numerically estimate which variables contribute statistically to the propagation of each other.

While these subjects do not encapsulate all possible interactions between these multidimensional datasets, they're an excellent start. It should be noted that the OPEC and non-OPEC datasets are a subset of a much larger data set. Fossil fuel data exists for most countries around the world (including a worldwide average). With the removal of this OPEC related constraint, the possibilities increase exponentially.

Conclusions

This case study is the foundation from which further GIS related analysis can be built upon. While the capabilities of GIS are undeniable, more sophisticated means of processing, cleaning, and preparing data for display are needed to draw insightful conclusions and enable GIS to convey more sophisticated information. GIS is an already robust means of communicating space related information and is greatly enhanced by the addition of other software suites that specialize in the mining and processing of data. While geospatial data exists, a GIS will always be present in some form to leverage this data and convey information. However, as time progresses, the field of data mining will grow more conventional, and support GIS as a more sophisticated means of preparing data for visualization and other purposes. Due to how harmoniously these two systems coexist together, data mining and GIS will continue to enrich our understanding of the world around us through analysis on all scales.

References

Akman, Engin, and Ibrahim Bozkurt. 2016. "On the Mixed Indirect Effects of Oil Prices on International Trade." *OPEC Energy Review* 40 (4): 374–96. doi:10.1111/opec.12087.

Atluri, Gowtham, Anuj Karpatne, and Vipin Kumar. 2019. "Spatio-Temporal Data Mining." *ACM Computing Surveys* 51 (4): 1–41. doi:10.1145/3161602.

Bhattacharyya, Siddhartha, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. 2011. "Data Mining for Credit Card Fraud: A Comparative Study." *Decision Support Systems* 50 (3): 602–13. doi:10.1016/j.dss.2010.08.008.

Casey, Gregory, and Oded Galor. 2016. "Population Growth and Carbon Emissions." doi:10.3386/w22885.

"GDP (Current US$)." 2022. *Data*. Accessed September 4. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2021&start=1960&view=chart.

Gregersen., Erik. 2022. "Data Mining." *Encyclopedia Britannica*. Encyclopedia Britannica, inc. Accessed October 8. https://www.britannica.com/technology/data-mining.

Guerreiro, Thais de, Janice Kirner Providelo, Cira Souza Pitombo, Rui Antonio Rodrigues Ramos, and Antonio Nelson Rodrigues da Silva. 2017. "Data-Mining, GIS and Multicriteria Analysis in a Comprehensive Method for Bicycle Network Planning and Design." *International Journal of Sustainable Transportation* 12 (3): 179–91. doi:10.1080/15568318.2017.1342156.

Malik, Rajat, and Rainu Nandal. 2018. "Comprehensive Study of Applications of Spatio Temporal Data Mining in GIS." *International Journal of Advanced Research in Computer Science* 9 (2): 208–10. doi:10.26483/ijarcs.v9i2.5686.

"Member Countries." 2022. *OPEC*. Accessed September 4. https://www.opec.org/opec_web/en/about_us/25.htm#:~:text=The%20Organization%20of%20the%20Petroleum,Founder%20Members%20of%20the%20Organization.

Moussa, Tamer, and Hassan Dehghanpour. 2022. "Evaluating Geothermal Energy Production from Suspended Oil and Gas Wells by Using Data Mining." *Renewable Energy* 196: 1294–1307. doi:10.1016/j.renene.2022.06.090.

Tong Wang, Xian-wen Gao, and Kun Li. 2012. "Application of Data Mining to Production Operation and Control System in Oil Field." *2012 24th Chinese Control and Decision Conference (CCDC)*. doi:10.1109/ccdc.2012.6242979.

"U.S. Energy Information Administration - EIA - Independent Statistics and Analysis." 2022. *International - U.S. Energy Information Administration (EIA)*. Accessed September 14. https://www.eia.gov/international/data/world.

Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2017. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Zheng, Yu. 2015. "Trajectory Data Mining." *ACM Transactions on Intelligent Systems and Technology* 6 (3): 1–41. doi:10.1145/2743025.