Economy and Environment: The Relative Efficiency between NATO Nations

Tommy Jun, Cynthia Zheng, Francesca Beller, Shaoran Sun

I. Introduction and Methods

This paper addresses the impact of modernization on the environment in the past few decades. Conclusions are drawn from the analysis performed in Python. In particular, the questions which are addressed are on the concerns of how the evolving economy affects greenhouse gas emissions and trends which are of significance. Specifically, the three criteria which are analyzed are: (a) the productivity of nations as measured by the annual dollar Gross Domestic Product per capita (GDP/capita in $/year), (b) the carbon dioxide production, in terms of tons per person (tons CO_2 /capita), and (c) the efficiency as measured by the ratio of the previous two. With these criteria, two specific analyses are performed. First, on the efficiency of the United States to other NATO countries through analyzing their respective confidence intervals. In general, this will gauge where individual countries stand in terms of the three criteria. The second analysis constructs multiple time series model for the United States and the other, combined, NATO countries. Both the productivity and emissions data are analyzed to illuminate the general trends of these two groups.

Regarding the nature of the data, the raw data is drawn from the Gapminder.org database, an international organization that aggregates statistics on global economies. Although it is well maintained and cross-validated by other international organizations, such as the World Health Organization, there are many unrecorded values. Therefore, the data cannot be used in its raw form and must be transformed by some means. Normality is tested by quantile-quantile plots and appropriate transformations are taken. Through linear regression the coefficients for the line-of-best-fit are determined and missing values are extrapolated. With the data complete, the analysis follows. The results of the first analysis shows positive relationship between productivity of nations and production of CO_2. Also, the United States has higher production of CO_2 per unit productivity, thus it needs to invest in more advanced technology to decreases the efficiency of productivity. The results of the second analysis show that the improvements in technology have increased the productivity of nations without significant costs to environment impact, in terms of emissions per unit productivity. Future predictions suggest that the emissions for the United States follows an ARMA(2,2) process and that the mean slowly converges to a finite value. Of course, the accuracy of this prediction decreases over time and only immediate prediction values are of significance. Recommendations on what to take away from this analysis are that further analysis should be performed for individual countries as was done for the United States in this paper. Then models between each country can be compared in order to determine if there are any underlying trends common to all countries. Furthermore, it is important to note that these analyses make several assumptions, namely that technology, emissions, and productivity are correlated. Multivariate analysis involving several potential influential variables would only improve the analysis.

II. Data Cleaning

The two data sets used in the project are the annual GDP/capita and CO_2 emissions (tons)/capita for NATO countries. Data is drawn from the Gapminder Foundation which aggregates global data on the Wealth and Health of Nations. The data is provided as an Excel sheet which is then exported as a csv. The data is trimmed prior to analysis and contains 28 columns (countries) and 56 rows (years). The GDP data is well documented and complete while the CO_2 emissions data has many missing entries due to when certain countries have entered into NATO[1]. In all, there should be a total of 3136 entries; however, 283 are missing. For this reason, these values have to be extrapolated. To do so, a line of best fit is found through performing linear regression on the data. But first, the normality assumption must be

[1]The only country which is missing data completely is the Czech Republic. This reduces the number of countries analyzed from 28 to 27.

met. Below is a sample raw time series of the United States:

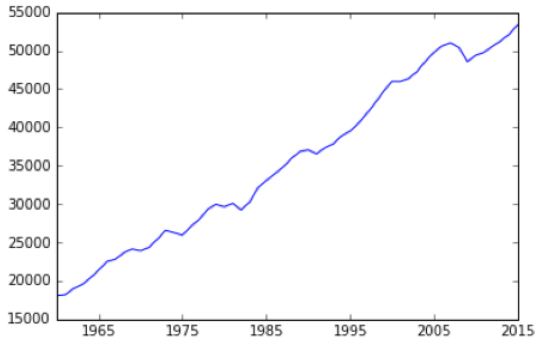

Fig. 1. Time Series of GDP/Capita for the United States

Although the data appears to follow a linear trend, the quantile-quantile plots reveal that the data may instead follow a logarithmic trend:

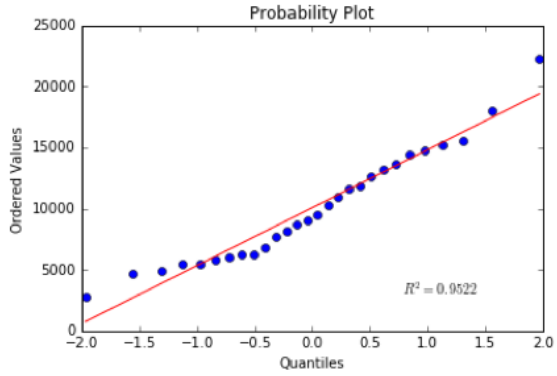

Fig. 2. Q-Q Plot of GDP/Capita for the United States Raw

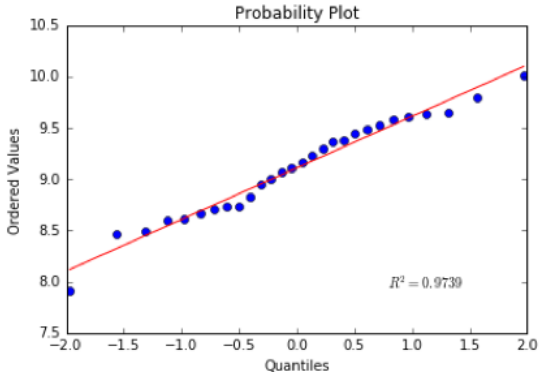

Fig. 3. Q-Q Plot of GDP/Capita for the United States Log Transformed

The r-squared value is slightly higher in the log transformed case for GDP; however, the difference is far more substantial for the log transformed CO_2 data:

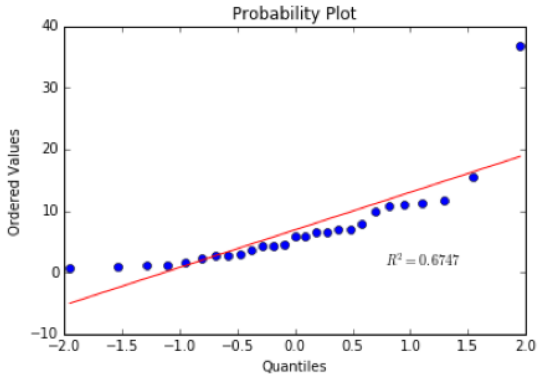

Fig. 4. Q-Q Plots of CO_2/capita for the United States Raw

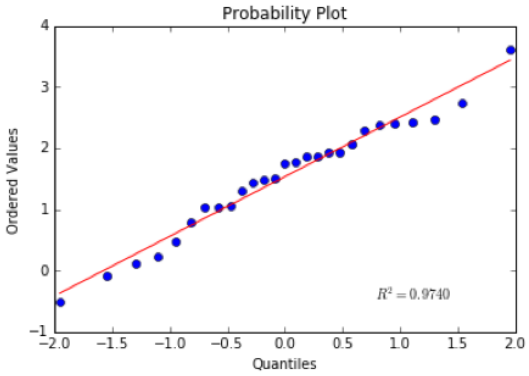

Fig. 5. Q-Q Plots of CO_2/capita for the United States Log Transformed

The r-squared values for the log transformed case (0.9740) is far greater than the untransformed case (0.6747). For this reason the team assumes that the CO_2 data is logarithmically distributed, and so logarithmically transform the CO_2 data before continuing with linear regression. Now that the normality assumption is satisfied, below is a sample of the result of the log transformation:

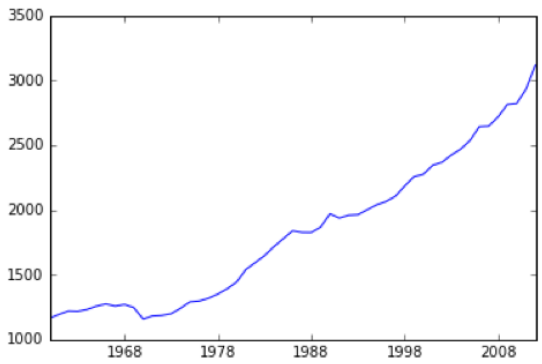

Fig. 6. Time Series of the Logarithmically Transformed GDP/CO_2 for the United States

It is evident that the time series is sufficiently linear, and this is also the case for the other countries. Assuming that linearity of this relationship[2] is common between countries, missing values from, such as from Slovenia, are estimated. By iterating linear regression on the data, the coefficients of the line-of-best-fit are found. These values are substituted to estimate the values for the missing entries of the CO_2 values. Iterating this method through each country and filling in the tables completes the data. Below is a sample of the methods described for Slovenia and Spain.
(Figure 7 and 8)

After visually inspecting the extrapolated data for each country, a linear trend is sufficient for the analysis which follows. Below are all time series for emissions superimposed on one another before and after the extrapolation.
(Figure 9 and 10)

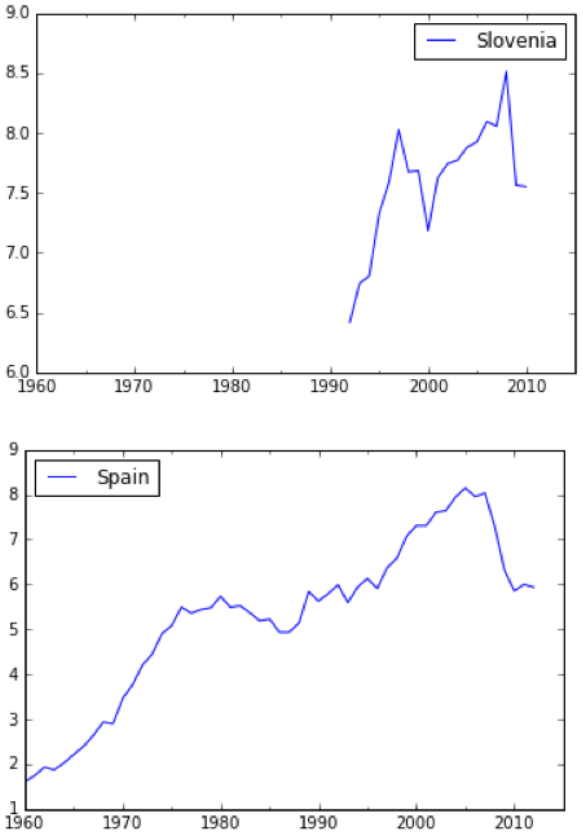

Fig. 7. Time Series of CO_2/capita for Slovenia and Spain Raw

[2]This is a large assumption; however, justified as generally most countries do follow a linear, rather than quadratic or higher order trend.

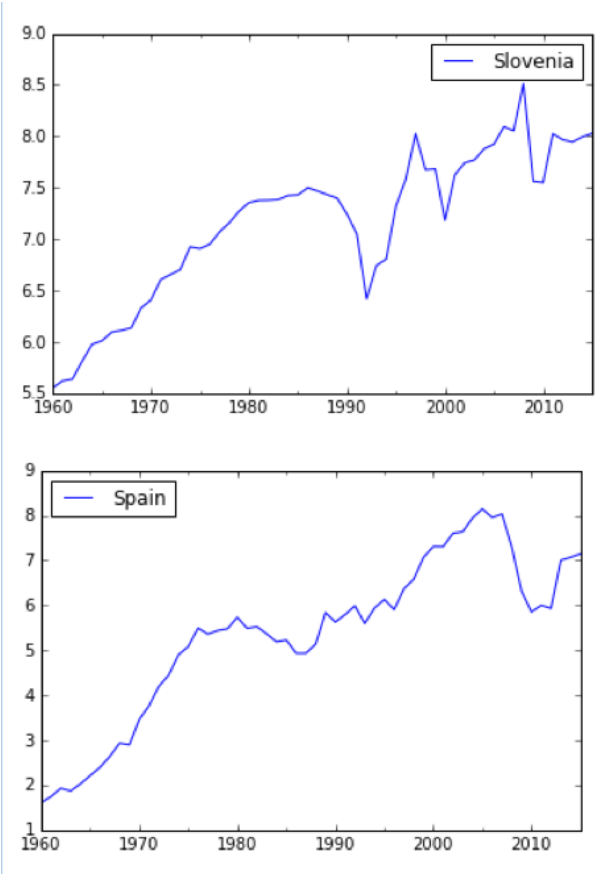

Fig. 8. Time Series of CO_2/capita for Slovenia and Spain Extrapolated Data

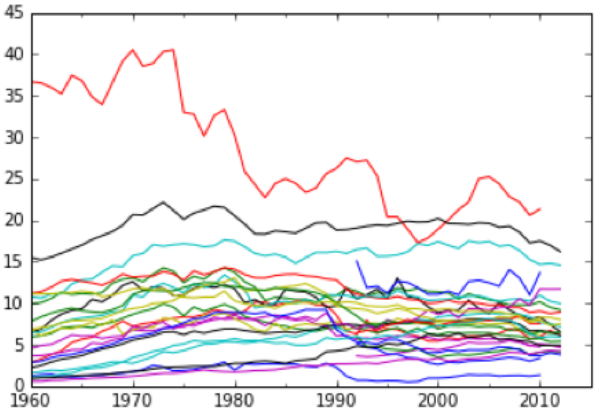

Fig. 9. Time Series of CO_2/capita for Slovenia and Spain Raw

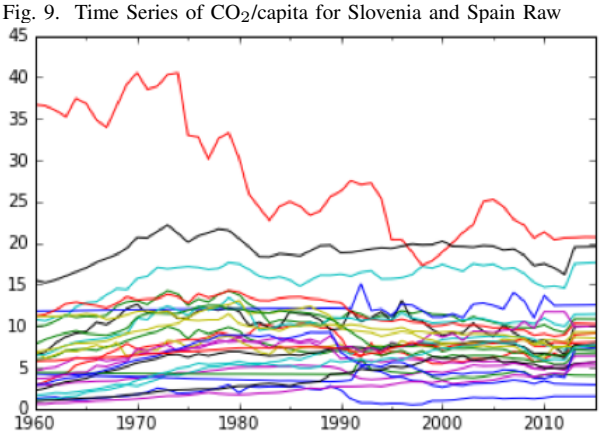

Fig. 10. Time Series of CO_2/capita for Slovenia and Spain Extrapolated Data

III. Confidence Intervals

NATO	95% Confidence Interval
Emissions/Capita	[8.1614, 8.5304]
GDP/Capita	[9.3969, 9.5752]
Efficiency of Productivity	[0.8685, 0.8909]

TABLE I
95% Confidence Interval for NATO

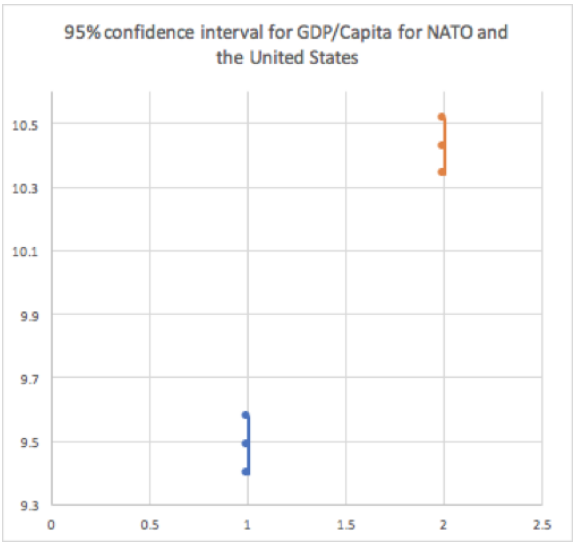

Fig. 11. 95% Confidence Intervals for Emissions and GDP for the United States

US	95% Confidence Interval
Emissions/Capita	[18.6421, 19.4921]
GDP/Capita	[10.3420, 10.5155]
Efficiency of Productivity	[1.8026, 1.8536]

TABLE II
95% Confidence Interval for the United States

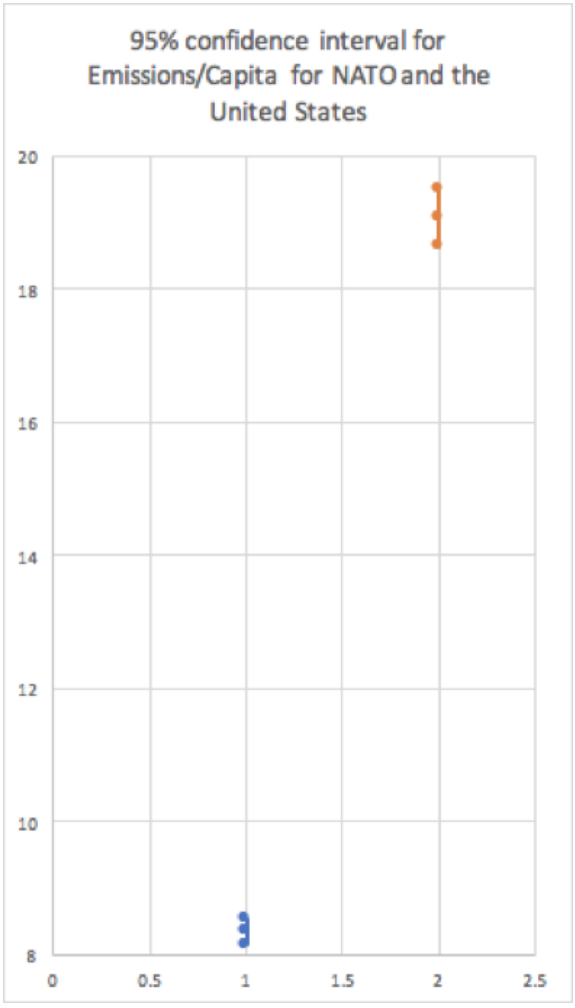

Fig. 12. 95% Confidence Intervals for Emissions and GDP for NATO

Over repeated samples, 95% of the confidence intervals will contain the true population mean. By comparing the 95% confidence intervals for Emissions/Capita for NATO countries and the United States, there seems to be significant difference between them since the confidence intervals do not overlap. Similarly, by comparing the 95% confidence intervals for GDP/Capita for NATO countries and the United States, there seems to be significant difference between them.

Furthermore, the United States has higher lower and upper bounds for both Emissions/Capita and GDP/Capita than the NATO countries. Therefore, the confidence intervals indicate positive relationship between the annual dollar Gross Domestic Product per capita (GDP/capita in $/year), and the carbon dioxide production, in terms of tons per person (tons CO_2 /capita).

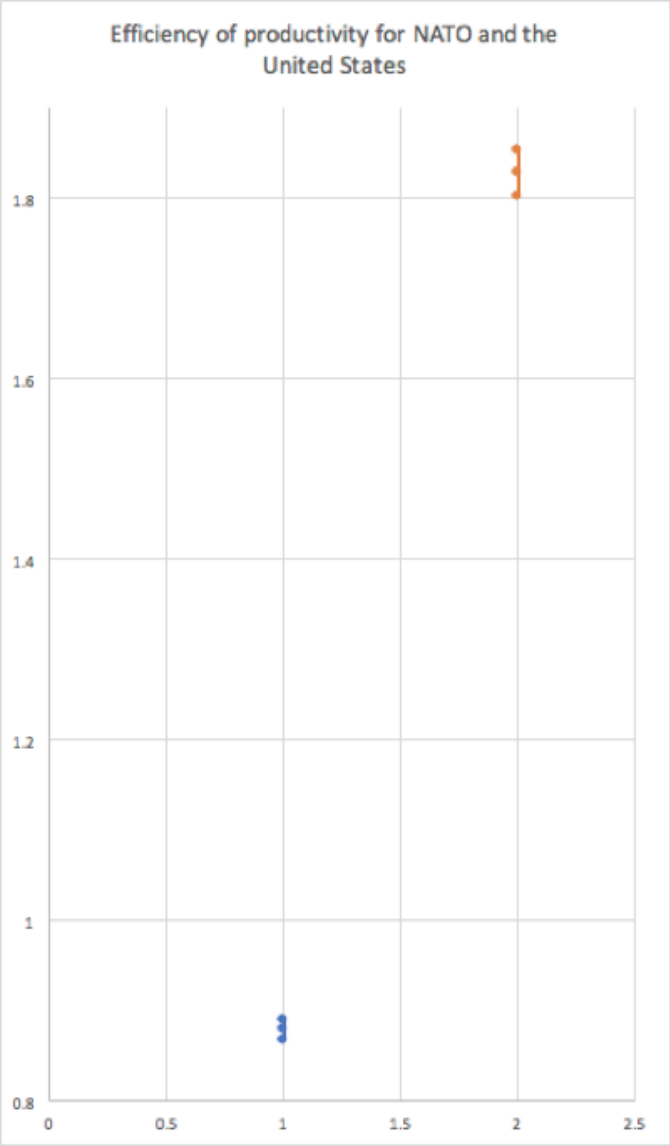

Fig. 13. 95% Confidence Intervals for Efficiency for NATO and the United States

After comparing the two confidence intervals for efficiency of productivity for NATO countries and the United States, it is concluded that the US produces more CO_2 per GDP. Therefore, the United States needs to use more advanced technology to reduce the production of CO_2.

IV. Time Series

Perform a time series analyses of the data and compare and contrast the results of the United States alone to all of the NATO countries combined. To do this, the mean of all countries CO_2/capita and GDP/capita was taken for each date in the dataset. The first analysis to be done was on the CO_2/capita. Below are the time series plots for the United States and all NATO countries:

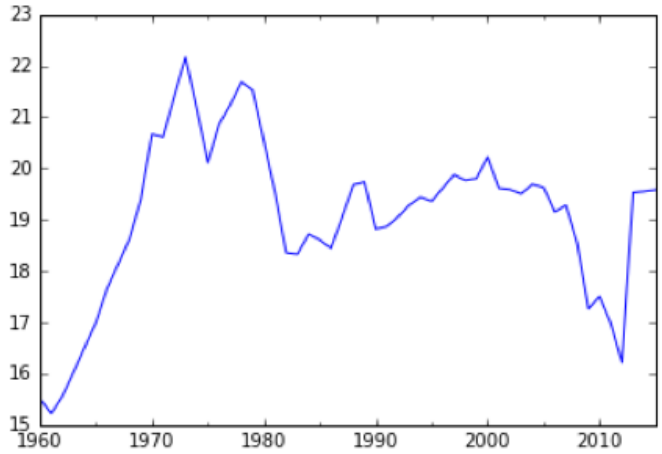

Fig. 14. Time Series for CO_2 of United States (logarithmic data)

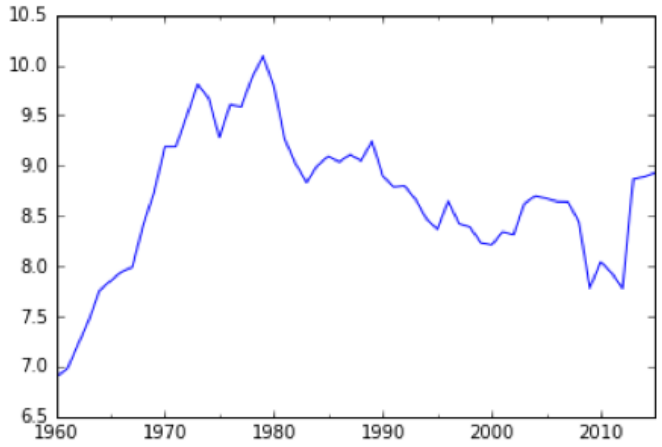

Fig. 15. Time Series for CO_2 of NATO countries (logarithmic data)

Looking at the plots, the most noticeable points are those at the time points of the economic recessions in the early 1970s, 2008, and 2012. This is more than likely in large part due to the collective effort of citizens to save money where possible, such as by reducing the use of cars to save on gasoline, etc. The CO_2 emissions drop during the 2012 recession in the United States is much more extreme. Also, while there is a general upward trend in CO_2 emissions in the United States from 1980 to the mid 2000s, the trend is downward for the combined NATO nations. From this, the question arises as to whether or not other NATO nations are making a more consistent, concerted effort to reduce CO_2 emissions than the United States. In general, the mean emissions over the time period appears to be stagnant. Next, time series plots were created for GDP/capita:

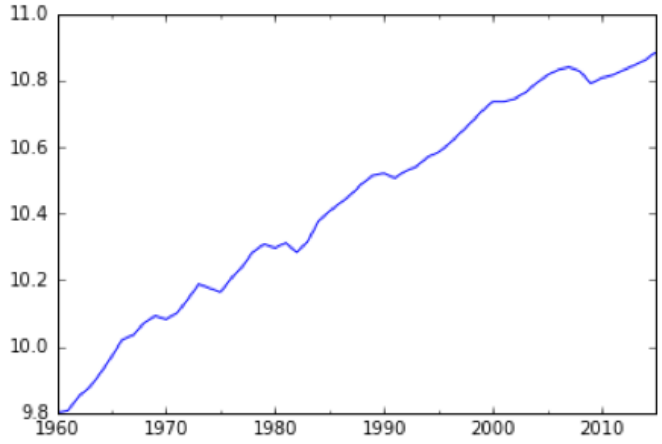

Fig. 16. Time Series plots for GDP/capita of United States

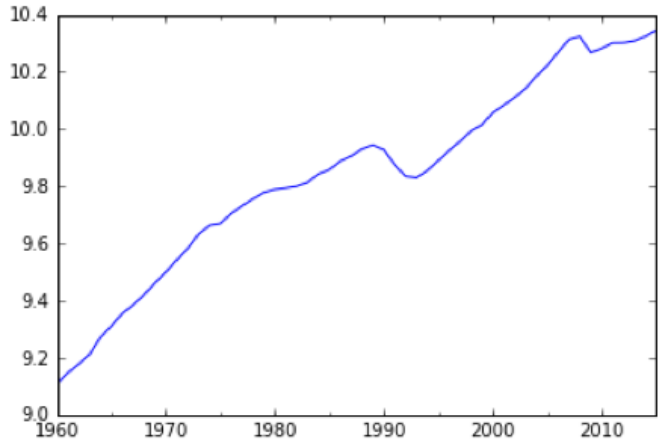

Fig. 17. Time Series plots for GDP/capita of NATO countries

From these plots, it is clear to see that economic output has a positive trend for the countries in the dataset. While the United States was seemingly unaffected by the global recession of the early 1990s, it is easily seen in the second time series plot. Again, as in the previous time series plot, the 2008 recession is visible, although not nearly as pronounced as in the CO_2 emissions data. In general, the GDP/capita is higher for the United States, with the start time for the dataset being at 9.8 and increasing from there, while the NATO countries begin at 9.1.

The upward linear trend for GDP and stagnant emissions data suggests that, overall for all NATO nations, advancements in technology have improved GDP without the cost of significant environmental impact in terms of emissions per productivity.

Next, a model is fitted for the emissions data of the United States. First, the ACF and PACF is generated. From this, an ARMA(1,3) seems to be the most plausible model; however, computing the AIC and BIC error terms reveal that an ARMA(2,2) is more appropriate (AIC: 138.909 vs. 137.674). Then, after choosing the ARMA(2,2) model, predict the future values of the emissions. The model reveals a sinusoidal trend whose amplitude slowly converges.

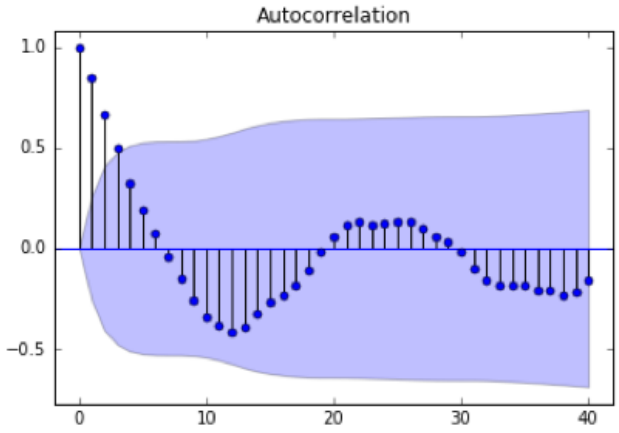

Fig. 18. ACF of United States Emission Data

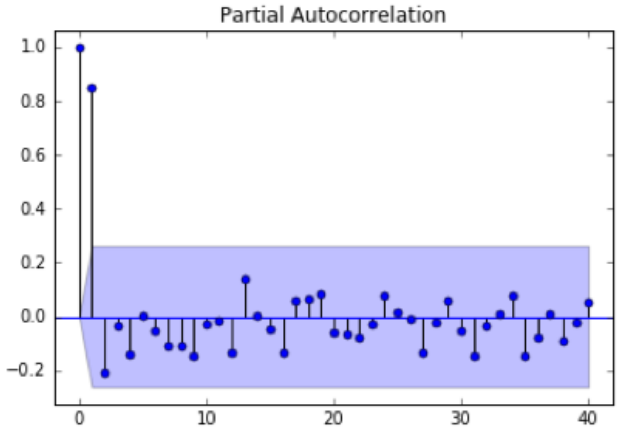

Fig. 19. PACF of United States Emission Data

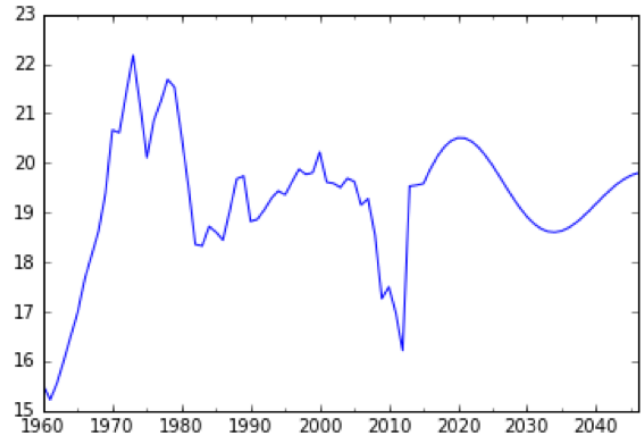

Fig. 20. Predicted Values of United States Emission Using an ARMA(2,2) Model

V. Code Samples

```
### Q-Q PLOTS ###
series = gdplogdf.iloc[0].values
series = series[~np.isnan(series)]
somedata = stats.probplot(series, dist="norm", plot=pylab)

### LINEAR REGRESSION ###
import pandas as pd
from sklearn import linear_model
# Linear Regression
for country in countries[countries!="Czech Republic"]:
somedf=pd.concat([gdplogdf[[country]],emissionslogdf[[country]]],axis=1)
    otherdf=somedf.dropna()
    regression = linear_model.LinearRegression()
    regression.fit(otherdf[[0]].values,otherdf[[1]].values)
    missing=somedf[np.isnan(somedf[[1]].values)]
    missing[[1]]=regression.intercept_+regression.coef_*missing[[0]].values
    emissionslogdf[[country]]=emissionslogdf[[country]].fillna(missing[[1]])

### CONFIDENCE INTERVALS ###
newemissionsdf # CO2/Capita
gdplogdf # GDP/Capita
newemissionsdf.__delitem__('Czech Republic') # remove Czech Republic
gdplogdf.__delitem__('Czech Republic') # remove Czech Republic

# Emission/Capita for NATO countries
newemissionsdf['Sum']=newemissionsdf.sum(axis=1,skipna=True)
# get the sum of countries for each year
newemissionsdf['Avg']=newemissionsdf.Sum/28
# divide the sum by 28 since there are 28 countries to get the average value
# for each year
xbar = np.mean(newemissionsdf.Avg)
# mean emission/capita of the mean values for each year
s = np.std(newemissionsdf.Avg,ddof=1)
# standard deviation emission/capita of the mean values for each year
n = len(newemissionsdf.Avg) # number of years
tstar = t.ppf(.975, n-1) # tstar
lcl = xbar - tstar*s/np.sqrt(n) #lower bound
ucl = xbar + tstar*s/np.sqrt(n) #upper bound
print("95% confidence interval for Emissions/Capita for NATO: ")
print([lcl,ucl])

# GDP/Capita for NATO countries
gdplogdf['Sum']=gdplogdf.sum(axis=1,skipna=True)
# get the sum of countries for each year
gdplogdf['Avg']=gdplogdf.Sum/28
# divide the sum by 28 since there are 28 countries to get the average
 value for each year
```

```
xbar = np.mean(gdplogdf.Avg)
# mean GDP/capita of the mean values for each year
s = np.std(gdplogdf.Avg,ddof=1)
# standard deviation GDP/capita of the mean values for each year
n = len(gdplogdf.Avg) # number of years
tstar = t.ppf(.975, n-1) # tstar
newlcl = xbar - tstar*s/np.sqrt(n) #lower bound
newucl = xbar + tstar*s/np.sqrt(n) #upper bound

print("95% confidence interval for GDP/Capita for NATO: ")
print([newlcl,newucl])

# Efficiency of productivity for NATO countries
[lcl/newlcl, ucl/newucl]

# Emission/Capita for the United States
...

# GDP/Capita for the United States
...

### TIME SERIES ###
newemissionsdf # CO2/Capita
gdplogdf # GDP/Capita
cogdpdf = newemissionsdf/gdplogdf # CO2/GDP

emmeans = newemissionsdf.mean(axis=1) # Emissions means of all NATO nations
gdpmeans = gdplogdf.mean(axis=1) # GDP means of all NATO nations
cogdpmeans = cogdpdf.mean(axis=1) # Emissions/GDP of all NATO nations

usem = newemissionsdf['United States'] # US emissions/capita
usgdp = gdplogdf['United States'] # US GDP/capita
uscogdp = cogdpdf['United States'] # US emissions/GDP

# TS plot of US emissions/capita
usem.plot()
...

# ACF and PACF
dta = usem
acf = sm.graphics.tsa.plot_acf(dta.values, lags=40)
acf.show()
pacf = sm.graphics.tsa.plot_pacf(dta.values, lags=40)
pacf.show()
```

```
# Let's try some models.
for i in range(0,4):
    for j in range(0,4):
        try:
            print("The model is ARMA with coefficients ",i,"and",j)
            somemodel = sm.tsa.ARMA(dta,(i,j)).fit()
            print("The AIC estimate is.")
            print(somemodel.aic)
        except:
            print("None")
            pass

# We can see that the best estimate is an ARMA(1,3) or ARMA(2,2).
# Try bic now.
...

# ARMA(2,2) is the best model by a small margin. Now make the next few
# prediction values using this model.
model = sm.tsa.ARMA(dta, (2,2)).fit()
pred = model.predict("2016","2046",dynamic=False)
predict = pd.concat([dta,pred]).plot()
```