HIGH SCHOOL EDITION
Journal of Student Research

Volume 10 Issue 4 (2021)

Application of OLS Regression and VAR Models to Analyse the Economies of Varying Political Regimes

Amrita Ganeriwalla[1] and Swapneel Mehta[#]

[1]Bombay International School, Bombay, India
[#]Advisor

ABSTRACT

The GDP per capita is a popular method of measuring the economic success of a country. This paper uses regression analysis to predict the GDP per capita (GDPPC) of countries using different independent variables. We applied Ordinary Linear Squares Regression and Vector Autoregression to check for a correlation between the chosen independent variables (Corruption Perception Index, Political Rights score, Civil Liberties score, Gender Inequality Index, Consumer Price Index, Population Density, and the percentage of people using the Internet) and the GDPPC. Using empirical evidence, we determine which model might be more accurate to attain this goal. Four countries of varying political regimes are studied - USA and Canada are categorised as democratic countries and China and Russia are non-democratic countries. Our results show trends in the correlations between the independent and dependent variables, and we can draw a distinction between the political regimes. We found that Corruption Perception Index and Population Density negatively correlates with the GDPPC of all 4 countries. We also noticed that the percentage of people using the internet and Gender Inequality Index correlates negatively with the GDPPC for non-democratic countries and in democratic countries the Consumer Price Index negatively influences the economy.

Introduction

Is there a relationship between the type of political regime in a country and its economic prosperity? Can we use machine learning to quantify it? Firstly, it is important to look at why a distinction in political regimes may matter. Intuitively speaking, a government in a democracy i.e., the ruling party, strives to achieve economic prosperity to increase their chances for a re-election. However, following this reasoning, why is a government in a non-democratic country motivated to maximize economic output? Przeworski et.al (1993. p.51-69) details 3 possible reasons: "(1) the state has a role to play to make the economy function efficiently, (2) the state must be insulated from private pressures if it is to perform this role well; and (3) the state apparatus wants to perform this role well."

The World Economic Forum's Global Competitive Index (2017), which categorises countries into 3 types of economies based on 12 index measures. The three types of economies are factor-driven economies, efficiency-driven economies, and innovation-driven economies. The Gem Consortium (2021), a reputed project's website that asses the levels of entrepreneurial activity prevalent in different countries globally, explains the distinctions between the economies: factor-driven economies are the least developed and depend on subsistence agriculture and rely heavily on unskilled labour and natural resources. Efficiency-driven economies are characterised by more efficient production processes and better product quality. Innovation-driven economies are comparatively the most developed. The United States has an innovation-driven economy while China and Russia have efficiency-driven economies, according to Zhang, Kinser and Shi (2014).

Machine learning applications in political science and econometrics is a growing space. Machine learning is an application of Artificial Intelligence that specialises in working with self-improving algorithms. It has numerous applications in data analysis because it uses experiences and past data to give more accurate answers. In this paper, we use linear models, namely the OLS Regression and VAR frameworks for analysis. OLS Regression estimates the relationship between one dependent variable and multiple independent variables. This statistical method of analysis has been previously used in the fields of economics and political science to estimate the GDP. Khan et.al (2016) have used the Human Development Index as an indicator of economic development in authoritarian regimes and have used OLS Regression to predict the HDI. VAR is often used in economic forecasting and several papers have studied nowcasting and forecasting of economic growth in countries. Yoon (2020) uses machine learning to forecast real GDP growth. VAR is a multivariate time-series algorithm. Essentially, it is used to predict the relationship between multiple variables over a period of time. VAR forecasting assumes that the forecasted variable has some dependency on other variables. One key difference between these two models is time-dependence: OLS Regression assumes each datapoint is independent of the others; whereas VAR does not rely on the assumption that data are independent and identically distributed. To summarise, machine learning helps us figure out the relation between sets of random variables and use that knowledge to make informed predictions about their future values.

Predictions of economic indicators like GDP or GDP per capita (GDPPC) help countries make important decisions related to the economy like planning fiscal policies and budgeting. Machine learning is used widely for this because it useful in making near accurate predictions. It uses past data to calculate the correlations between variables and use that to make predictions and accounts for any trends that have emerged. GDPPC is a metric used to calculate a country's economic output per person. Cambridge dictionary defines it as "the total value of all the goods and services produced in a country in a particular year, divided by the number of people living there" or simply, the GDP divided by the population of the country. GDP can primarily be calculated using three primary methods, all of which should give the same value. Matthews (2012) details the 3 methods: The Expenditure Approach, The Production (Output) Approach, and The Income Approach. The GDP of the United States of America and the Russian Federation is calculated using the Expenditure Approach.

Through my research, I empirically analyse if the independent variables affect the GDPPC and if they can be used to predict it. A noteworthy paper is Magee and Doces (2015, p. 223-237) who state that authoritarian regimes tend to overstate their true economic growth rates by about 0.5-1.5 percentage points in the data that they report to the World Bank. Hollyer, Rosendorff and Vreeland (2011) conclude in their paper that democracies are more transparent than other political regimes. Interestingly, this paper also uses an OLS regression to assess the relationship between their dependent and independent variables.

This paper seeks to use machine learning to analyse if certain factors affect the economy of a country and if and how those factors affect the economy of democratic or non-democratic countries. The next section, section 2, outlines and visually represents the datasets. In section 3, applications of Ordinary Linear Squares (OLS) Regression and Vector Autoregression (VAR) are used to show correlation between the factors and the GDPPC, which quantifies the economic prosperity. The democratic countries used in this paper are Canada and the United States of America (USA); the non-democratic countries are People's Republic of China (referred to as China) and the Russian Federation (referred to as Russia). They are classified as Democratic and Non-Democratic for this study based on an index, Freedom House's Global Freedom Score. As of 2021, China (scored 9 points) and Russia (scored 20 points) are stated as "Not Free", and Canada (scored 98 points) and the USA (scored 83 points) are stated as "Free". Section 4 documents the results of the machine learning frameworks. The analysis and implications of the results are discussed in section 5. Finally, the conclusions of the research are summarised in section 6

Datasets

Dependent Variables

Data for the dependent variable used in this model, the GDPPC or the Gross Domestic Product Per Capita of a country, is a part of the Maddison Project Database 2020. In its documentation, the project is described as "ongoing research project aimed at standardising and updating the academic work in the field of historical national accounting in the tradition of syntheses of long-term economic growth produced by Angus Maddison in 1990s and early 2000s." It consists of data of over 160 countries from the Roman times to the present. In this study, only the data from 1988 to 2018 is used. The datasets for this variable are visualised in Figure 1.

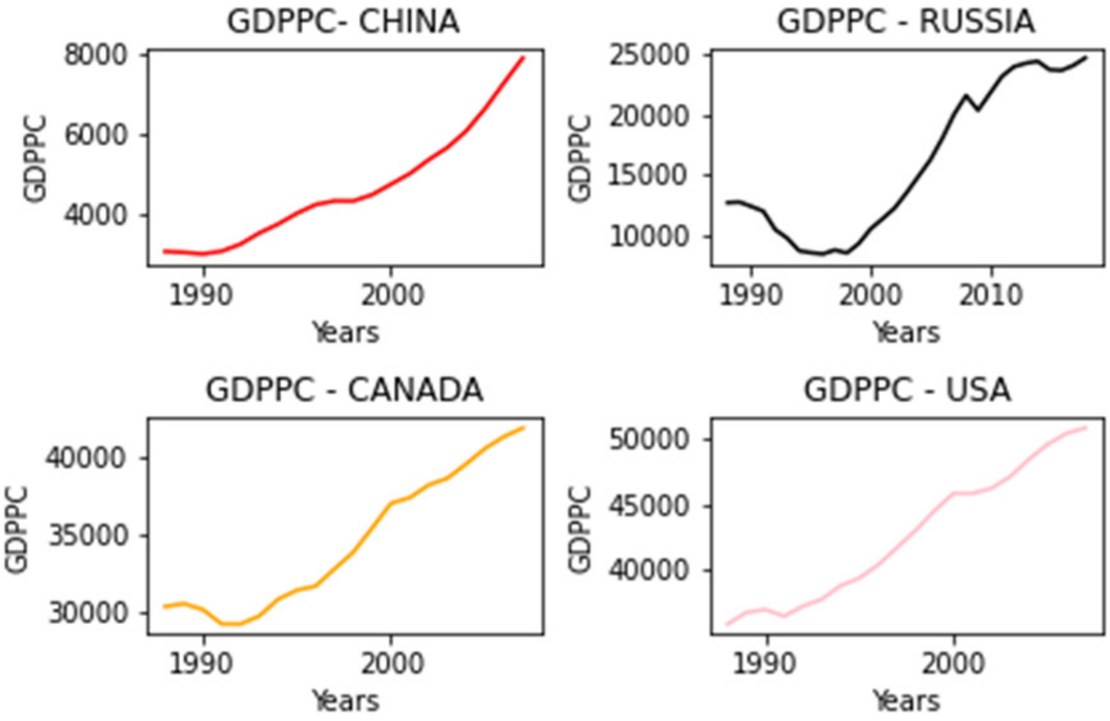

Figure 1. Visual representation of each country's GDPPC by year

Independent Variables

The first independent variable that is used is the Corruption Perception Index. It is referred to as 'CPI'. This index is released annually by Transparency International. It ranks 180 countries and territories by their levels of public sector corruption as perceived by experts and businesspeople on a scale of 0 to 100, where 0 is highly corrupt and 100 is highly clean. 13 different data sources, independent from Transparency International, are used to calculate the index. Each of these 13 datasets are in the form of indices of different scales. In order to compare them, Transparency International standardises the indices. After this, they convert these scores to a scale of 0 to 100. The converted scores are then averaged for each country, producing the Corruption Perception Index. The datasets for this variable are visualised in Figure 2.

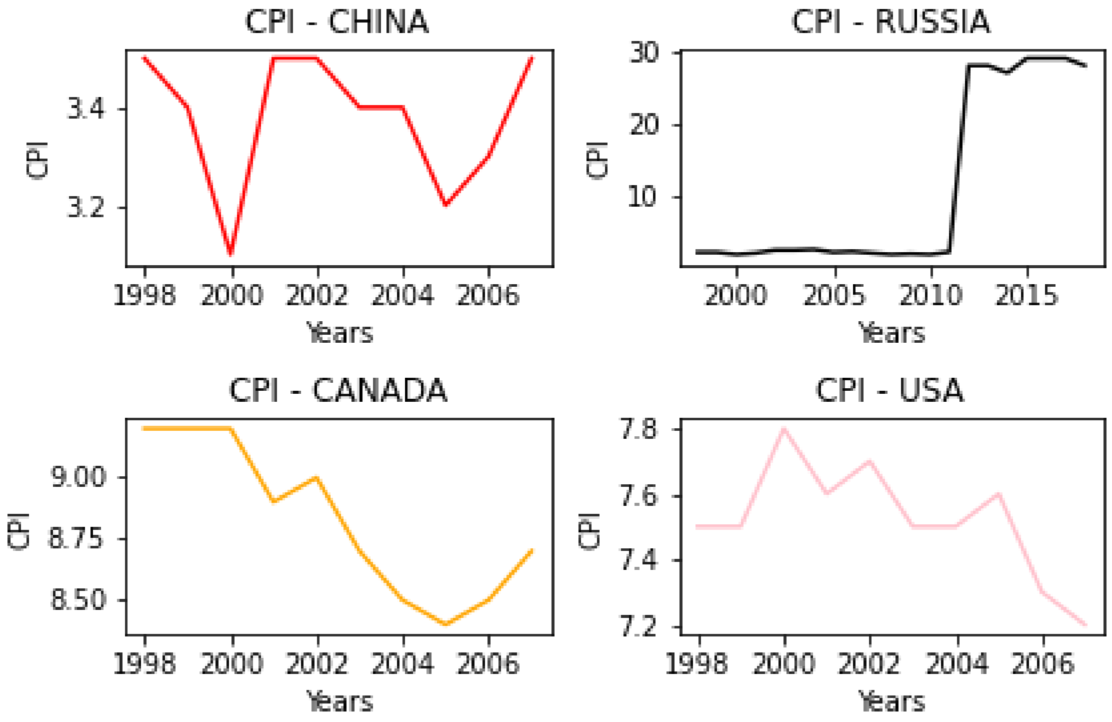

Figure 2. Visual representation of each country's CPI by year

The next two independent variables are the Political Rights and Civil Liberties scores. The Political Rights are referred to as 'pol_rights' and the Civil Liberties as 'civil_lib'. They are part of a dataset released by Freedom House. The Political Rights' variable is calculated with a focus on 3 categories: the electoral process, political pluralism, and participation, and functioning of the government. Each category has some indicators. Electoral process includes executive elections, legislative elections, and the electoral framework. Political pluralism and participation include party systems, political opposition and competition, political choices dominated by powerful groups, and minority voting rights. Functioning of governments includes corruption, transparency, and ability of elected officials to govern in practice. The datasets for this variable are visualised in Figure 3.The Civil Liberties score is calculated with a focus on 4 subcategories: freedom of expression and belief, associational and organizational rights, rule of the law, and personal autonomy and individual rights. Similar to the Political Rights, the Civil Liberties also has indicators. Freedom of expression and belief includes media, religious, and academic freedoms, and free private discussions. Associational and organizational rights consist of free assembly, civic groups, and labour union rights. Rule of law involves independent judges and prosecutors, due process, crime and disorder, and legal equality for minority and other groups Lastly, personal autonomy and individual rights includes freedom of movement, business and property rights, women's and family rights, and freedom from economic exploitation. Each country is awarded 0-4 points for each of the indicators, where 0 represents the lowest degree of freedom and 4 represents the highest degree of freedom. Each country is then assigned the two scores based on the total scores for the 10 Political Rights and 15 Civil Liberties indicators. The data is gathered by their in-house and external expert analysts and advisers. While they are guided by the 25 questions, they also cover supplemental questions. The datasets for this variable are visualised in Figure 4.

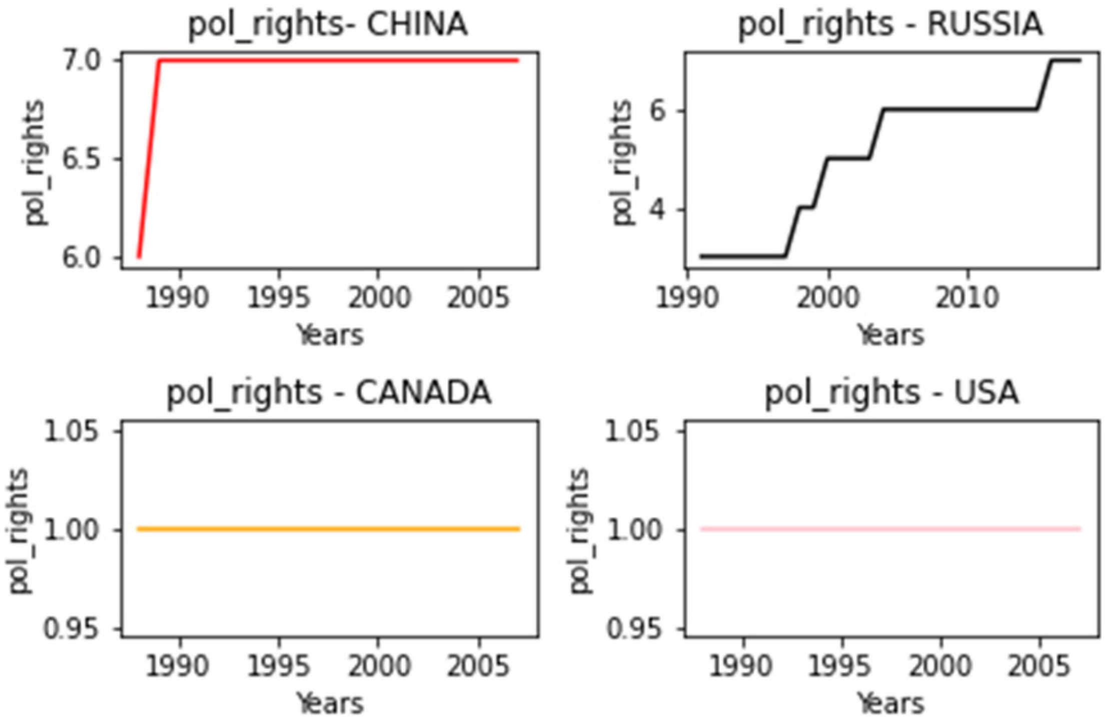

Figure 3. Visual representation of each country's pol_rights by year

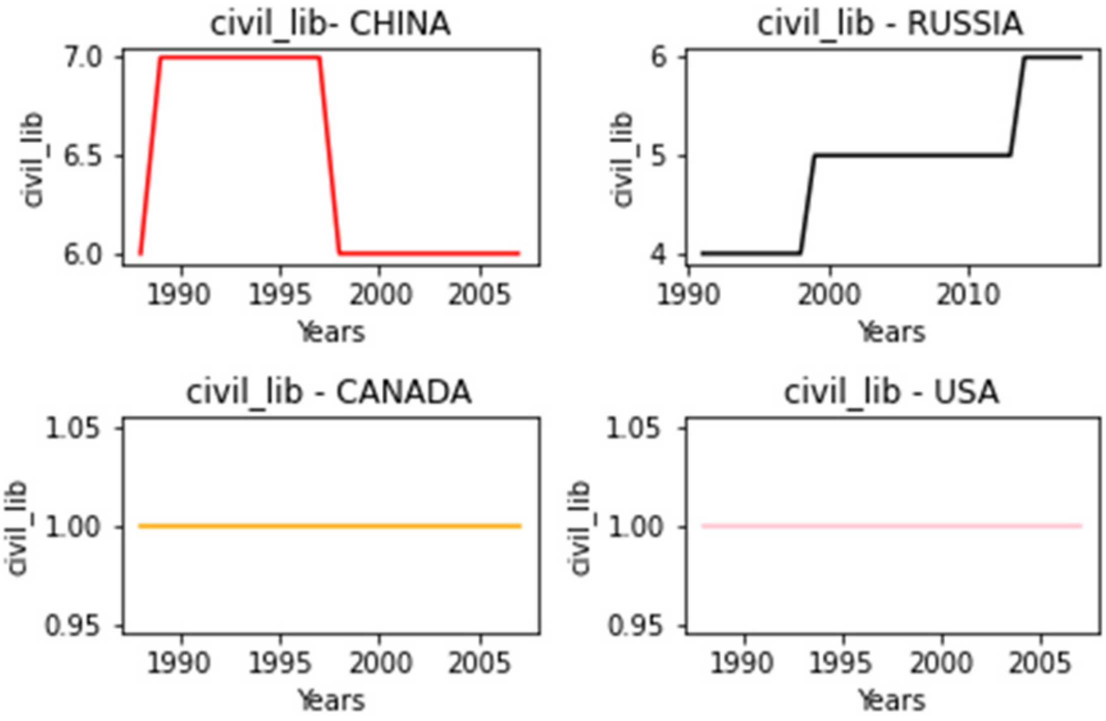

Figure 4. Visual representation of each country's civil_lib by year

The fourth independent variable is the percentage of people using the internet. It is referred to as '%usingint'. This is released by the International Telecommunication Union. It is easily comparable across different countries as it is in percentage form. The ITU defines it as the proportion of individuals who used the Internet from specified locations in the last three months. The Internet, a worldwide public network, provides access to numerous communication services like the World Wide Web and carries e-mail, news, entertainment, and data files. It can be accessed can be via a fixed or mobile network, including wireless access at a Wi-Fi 'hotspot'. Locations of Internet use are their homes, workplaces, schools, community internet access facility, commercial internet access facility, and mobile connectivity. The datasets for this variable are visualised in Figure 5.

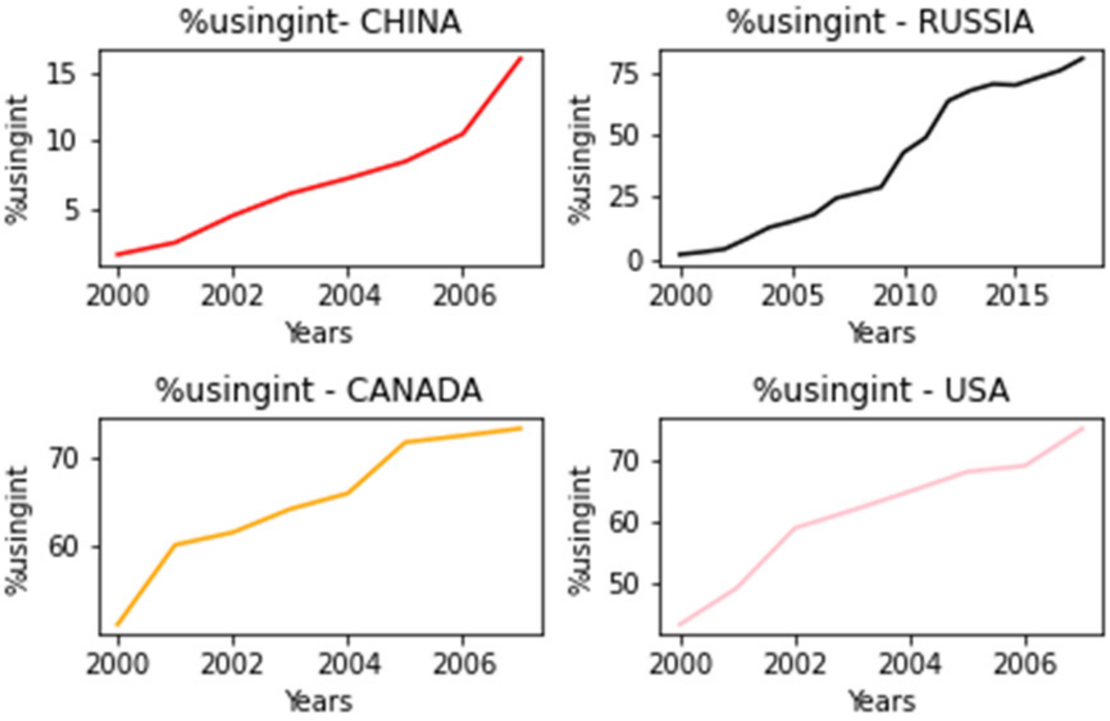

Figure 5. Visual representation of each country's %usingint by year

The fifth independent variable is the Consumer Price Index which is calculated by the International Monetary Fund. This variable is referred to as 'cons_pri_ind'. Its measure changes over time in the general level of prices of goods and services that households acquire (use or pay for) for the purpose of consumption. It is considered a macroeconomic indicator of inflation and is used by governments and central banks. It is especially useful while framing monetary policy, monitoring price stability, and as deflators in the national accounts. The datasets for this variable are visualised in Figure 6.

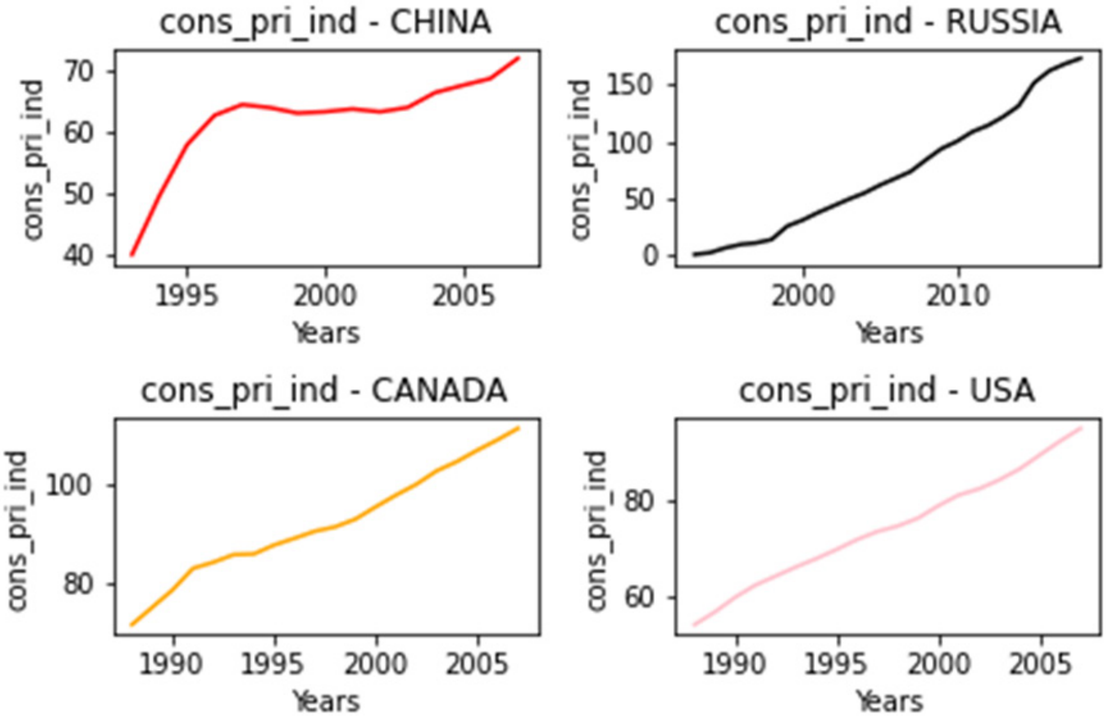

Figure 6. Visual representation of each country's cons_pri_ind by year

The sixth independent variable is the Gender Inequality Index, which is also given by the International Monetary Fund. It is referred to as 'Gend_ineq_index'. The Human Development Reports describe this as an index between 0 and 1. The higher the number, the more inequality there is. The dataset comprises data from 162 countries and takes these 3 important aspects of human development into consideration: reproductive health, empowerment, and economic status. The reproductive health is quantified by maternal mortality ratio and adolescent birth rates. The empowerment is the proportion of parliamentary seats occupied by females and the proportion of adult females and males aged 25 years and older with secondary education. The economic status is quantified as labour market participation and measured by labour force participation rate of female and male population 15 or older in age. The datasets for this variable are visualised in Figure 7.

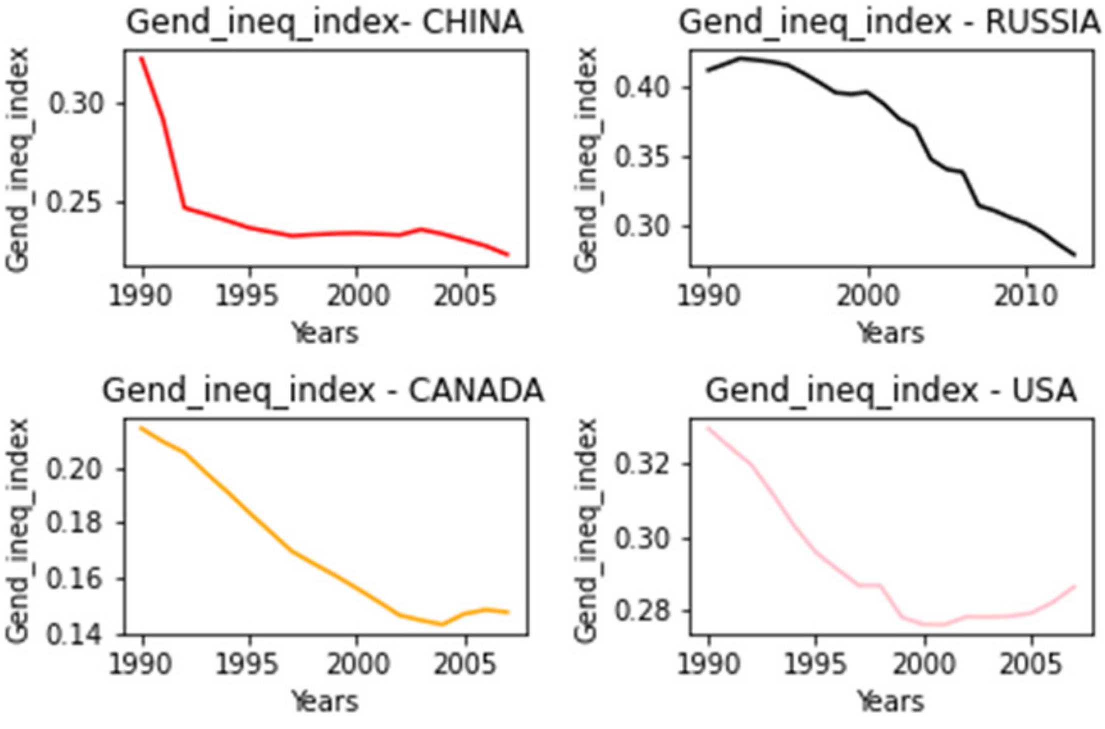

Figure 7. Visual representation of each country's Gend_ineq_index by year

The seventh independent variable is the Population Density, which is calculated by the World Bank. This is referred to as 'POPDEN'. The World Bank says that the population density is calculated by dividing midyear population by land area in a country. Estimates of the population are usually based on national population censuses. Estimates for the years before and after the census are interpolations or extrapolations based on demographic models. Population estimates are taken from demographic models and are susceptible to biases and errors from shortcomings in the model and data. Population includes all residents regardless of legal status or citizenship - except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin. Land area refers to a country's total area, excluding area under inland water bodies, national claims to continental shelf, and exclusive economic zones. The datasets for this variable are visualised in Figure 8.

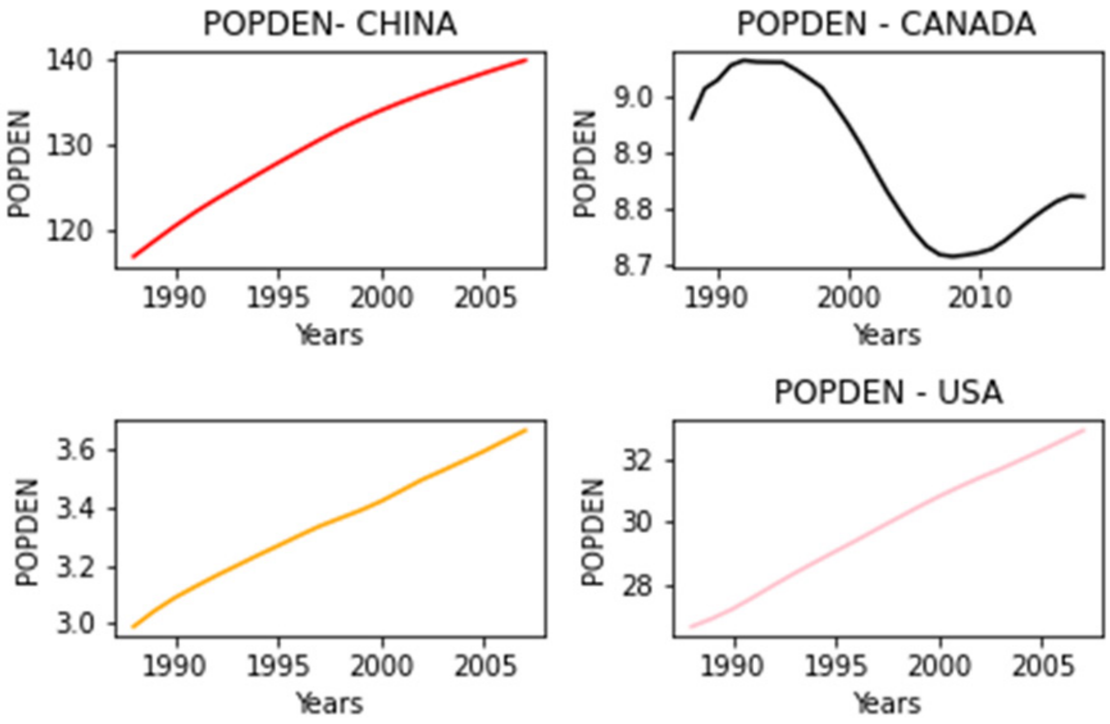

Figure 8. Visual representation of each country's POPDEN by year

Methods

The two machine learning models of OLS Regression and Vector Autoregression are used to predict the GDPPC of a country based on the 7 independent variables defined in section 2. The models serve to check for correlation relationships between the independent variables and dependent variables. The code for the two models was written in Python using Google Collaboratory. The last 7 rows of the Corruption Perception Index were not considered because it seemed like the method for scaling the data had been changed. The missing values were imputed in both the models.

Time Series Model

The first model that was used was the Vector Autoregression model, which is a time series model. VAR is a form of multivariate time series. VAR forecasting is used when there are two or more variables that may affect each other.

The libraries imported and used to build this model include pandas, matplotlib, NumPy, statsmodels, sci-kit learn, datetime, and tabulate.

After importing all the necessary libraries and packages, and uploading the data frame, the data frame is split into a training and test set.

The machine learning framework can be separated into 4 parts.

First we checked for stationarity of the data frame. The missing values were imputed and then scaled. Each column of the data was then visualized. To check for stationarity, the copy of the training data frame was put through the KPSS test and the ADF test.

The KPSS test, also known as Kwiatkowski-Philips-Schmidt-Shin Test is used for checking stationarity of datasets around a deterministic trend. The null hypothesis of the KPSS test is that the data is stationary, and the alternate hypothesis is that the data is not stationary. The ADF Test, or the Augmented Dickey Fuller test, is also a stationarity test. The null hypothesis of the ADF test is that the data is not stationary, and the alternate hypothesis is that the data is stationary. One of the main differences between these two tests are that if a dataset has an increasing or decreasing trend, the KPSS test may still show that it is stationary while the ADF test will not. For this model, the results of the KPSS test will hold more weightage because we realised that the data frame exhibits trends.

After this, the second part of the VAR framework begins. This part works on the actual training set. The missing data was imputed. The column that holds the date is dropped. The variables were then scaled and transformed using the same scaler. At this point, the Political Rights and Civil Liberties datasets were dropped because after scaling, their values become null. The years are then added to this data frame as its index. The reason for doing this, is so that the value of the years does not get scaled. The data columns are then visualised and decomposed seasonally. As a result of this, the conclusion drawn was that the dataset has trends. KPSS test and ADF test was then performed on the dataset and then it is seasonally decomposed again.

The third part of the model involved fitting the VAR model. The training dataset was put into the model as its parameters. The model is fit with the number of lags set as 1 and a constant and linear trend. Next, forecasting is done. 11 forecasts are made, it is the size of the test set. At this point, it is important to note that all the predictions are scaled, so we have to rescale it. At this point, we have the forecasts of the test set in the form of a Pandas data frame with the date as the index.

The fourth and last part focuses on analysing the model. It checks for accuracy of the predictions in the form of MAPE, ME. MPE, MinMax and RMSE. Finally, plots are drawn to visualise the predictions and actual GDPPC. MAPE is Mean Absolute Percentage error, MAE is mean absolute error, ME is mean error and RMSE is Root Mean Squared Error.

Ordinary Linear Squares (OLS) Regression Models

The next model that was applied was the Ordinary Linear Squares Regression (OLS Regression). OLS Regression estimates the relationship between one or more independent variables and a dependent variable. This is done by minimizing the sum of squared residuals (i.e., the difference between the observed and predicted values of the dependent variables which are in the form of a straight line) from the line of best fit.

While writing the OLS Regression model, the packages imported and used were NumPy, pandas, matplotlib, and sci-kit learn. The regression code was written in the form of a function which had the data frame as the parameters. So essentially, by calling the function with a country’s data frame as the parameters, the function would return the values of the predicted GDPPC.

There was one master() function defined. Inside that function, the data frame was first split into a training dataset (with 20 rows of data) and a test dataset (with 10 rows of data). Due to the limited data, there was no validation set. The Political Rights and Civil Liberties columns were dropped as independent variables because they were not used in the Time Series model.

Next, the missing values in the training set were imputed. After imputing the missing values, the independent variables of the training set were scaled and transformed.

Following this, a LinearRegression() model was defined. The sci-kit learn package was used. The scaled independent variables of the training set and the unscaled GDPPC were fit to the model.

After this, the test set was considered. The missing values in the test dataset were imputed using the same function. Then the independent variables were scaled using the same scaler. It is important to note that the predictions and the GDPPC are on the same scale, removing the need to rescale the predictions. The function finally returned the predictions of the training set and test set.

Predictions for the test set were made by inputting the scaled independent variables of the test set. The model generated predictions of the GDPPC. After this, the model intercept was found and printed. The error was calculated using RMSE or root mean squared error. RMSE is the standard deviation of the residuals.

In the second half of this model, each of the country's data frames were uploaded. The master() function, which was described in the paragraph above, was called with the data frames as its parameters. It returned the predictions of its training and test set. After all the 4 countries were run through this process, a graphical representation was made to visualise the results of the model by comparing the actual GDPPC and the predicted GDPPC.

Results

This subsection deals with the results generated by the two models. First the predicted and forecasted values of the GDPPC by the OLS and VAR models respectively are compared to the actual GDPPC of the test set. The correlation matrix of residuals and the correlation coefficients generated from the VAR model are tabulated after this. Then the accuracy of the VAR model is measured using various statistical tests. The OLS Regression model's scaler mean, scaler variance and model intercept for each country is documented. To compare the relative accuracy of both the models, the RMSEs are compared for each country. Lastly, the seasonal decomposition graphs and the prediction graphs are mentioned.

Predictions

Table 1 shows the forecasts of the VAR model, the predictions of the OLS model and the actual GDPPC of the test set tabulated.

Table 1. Comparison of the outputs of both the models with the actual values

COUNTRY	YEAR	VAR	OLS	ACTUAL
CANADA	2008	42302.5	41864.67622336	41896.4256
	2009	42590.3	41168.54122785	40312.6217
	2010	43088.2	41298.90899048	41209.4262
	2011	43765.5	43346.136617	42197.0000
	2012	44694.1	43876.21638391	42445.0000
	2013	45757.4	44577.02437957	42994.0000
	2014	46882.9	43988.72414453	43607.0000
	2015	47966.6	44261.12351262	43619.0000
	2016	48957.4	44458.76779745	43745.0000
	2017	49823.7	44708.93236583	44591.6417
	2018	50571.6	45189.52625616	44868.7435
CHINA	2008	8274.72	8930.33042993	8190.1121
	2009	8567.69	10068.25740225	8734.0406
	2010	8866.89	11246.54359036	9658.4186
	2011	9185.61	11938.46569029	10221.0000
	2012	9555.95	12823.24675523	10680.0000
	2013	9983.99	13577.76490576	11328.0000
	2014	10458.7	14043.8644025	11944.0000
	2015	10962.5	14581.77588662	12244.0000
	2016	11477.1	15221.22017222	12569.0000
	2017	11988.1	15565.28419937	12733.9314
	2018	12488.1	13682.72718819	13101.7064

RUSSIA	2008	21338.2	21354.09869546	21563.4619
	2009	22893	22738.91285776	20335.5577
	2010	24202.9	26065.06412214	21737.3839
	2011	25076.5	27868.75903337	23130.0000
	2012	25509.9	31217.84708806	23931.0000
	2013	25443.1	32693.0514294	24224.0000
	2014	24761.9	34273.61993038	24387.0000
	2015	23351.8	36366.3737743	23691.0000
	2016	21108.4	38032.6804029	23635.0000
	2017	17933.8	39192.93969525	24042.6341
	2018	13736.9	40645.21850729	24668.9079
USA	2008	51398.9	51974.34935606	50275.7463
	2009	52179.3	51834.50730374	48452.9335
	2010	53098.8	52169.49233792	49266.9159
	2011	54125.9	53435.88619763	49675.0000
	2012	55136.3	54674.72047946	50394.0000
	2013	56071	55339.81292184	50863.0000
	2014	56919.5	55673.63901534	51664.0000
	2015	57707.6	55794.509078	52591.0000
	2016	58473.1	56455.11027457	53015.0000
	2017	59247.6	57328.05659038	54007.7698
	2018	60048	58160.20933125	55334.7394

It is interesting that the OLS Regression model gave numerically higher predictions for Canada, China, and Russia. For the USA, the VAR model gave numerically higher results.

Correlation matrix for residuals for VAR:

A correlation matrix for residuals for VAR shows a correlation coefficient for 2 variables in Table 2. This table summarises the information that the independent variables do not describe. It shows where the unexplained variance is located. It must be noted that the Political Rights and Civil Liberties were not considered.

Table *2.* Tabulation of correlation matrix of residuals with respect to the GDP per capita

COUNTRY	%usingint	CPI	POPDEN	Gend_ineq_index	Cons_pri_ind
Canada	0.493296	-0.499680	-0.389700	0.516099	-0.218687
China	0.266796	-0.383470	-0.491099	-0.764410	-0.310601
Russia	0.910121	-0.594447	0.499096	-0.248103	-0.037758
USA	-0.070966	-0.388543	0.406447	0.007467	0.123872

Correlation coefficients of GDPPC using VAR:

Table 3 summarises the correlation coefficients of the independent variables with the GDPPC. The results of this table are from the Summary of Regression Results for equation gdppc.

Table *3*. Tabulation of correlation coefficients for equation of gdppc

COUNTRY	%usingint	CPI	POPDEN	Gend_ineq_index	Cons_pri_ind
Canada	0.273660	-0.004167	-5.716612	0.136321	-0.340399
China	-0.342253	-0.004632	-0.394174	-0.016188	-0.089075
Russia	-0.597602	-0.044436	-0.560238	-0.279290	0.083305
USA	-0.089136	-0.086090	-1.419640	-0.125713	-0.956205

Using the correlation coefficients, the influence of different variables on the GDPPC can be quantified. In Canada, the Population Density has the greatest correlation and negatively influences the GDPPC, while the Corruption Perception Index has the least correlation and negatively influences it. In China, the Population Density has the greatest correlation but negatively influences the GDPPC, while the Corruption Perception Index has the least correlation and negatively influences it. In Russia, the percentage of people using the internet has the greatest correlation and negatively influences the GDPPC and the Corruption Perception Index has the lowest correlation and negatively influences the economy.

In the USA, the Population Density seems to have the highest correlation and negatively influences the GDPPC, while the Gender Inequality Index has the lowest correlation and also has a negative effect on the GDPPC.

Forecast Accuracy of the VAR Model

Table 4 summarises the results of the error tests run for the VAR Model. This again helps analyse the accuracy of the model.

Table 4. Summary of errors in the VAR model

COUNTRY	MAPE	ME	MAE	MPE	RMSE	MINMAX
Canada	0.0731	3174.0151	3174.0151	0.0731	3576.8881	0.0671
China	0.078	-872.2645	887.6482	-0.0761	989.799	0.078
Russia	0.1167	-908.1341	2752.2215	-0.034	4088.8401	0.1132
USA	0.0861	4442.3626	4442.3626	0.0861	4597.5869	0.0789

OLS REGRESSION MODEL:

In Table 5, each country's model intercept is tabulated.
The Model Intercept is the expected value of the Y variable when X=0. The intercept of a graph is the point where the line crosses the y-axis.

Table 5. Model intercept of the OLS Regression Model

COUNTRY	MODEL INTERCEPT
Canada	34412.087065
China	4622.524735
Russia	11978.52546
USA	42688.82896

RMSE

The RMSE of each country, namely the OLS Regression and the VAR models, are found in Table 6. This will help quantitatively analyse the best model for each country and regime.

Table *6.* Comparative RMSEs for each country

COUNTRY	OLS REGRESSION	VAR MODEL
Canada	842.1095800459977	3576.8881
China	1970.645443327808	989.799
Russia	10100.879304863656	4088.8401
USA	3469.838210955317	4597.5869

Taken together, the findings indicate that for China and Russia, the VAR model have more accurate forecasts; in the cases of Canada and USA, the OLS Regression model makes more accurate predictions.

Time Series Graphs

The graphs below (Figure 9, 10, 11, and 12) represent the seasonally decomposed data for each country. Trend refers to increasing or decreasing values in the series and seasonality is the any repeating short-term cycle in the series. Seasonal decomposition provides a framework which helps easily analyse the data to inform the forecasting models.

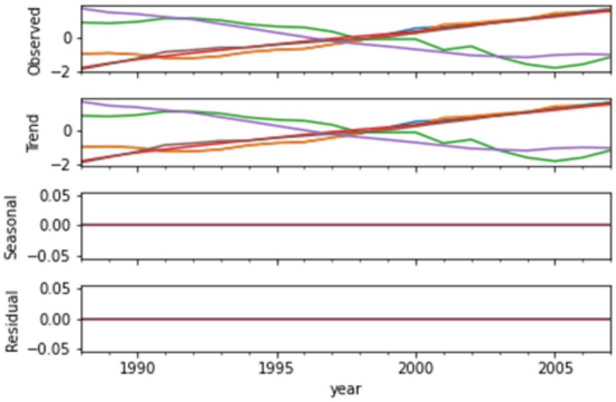

Figure 9. Seasonally decomposed data for Canada

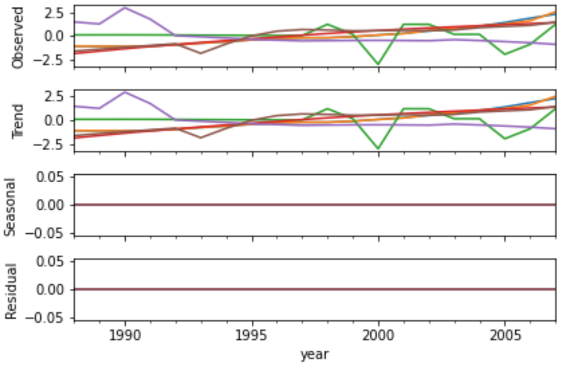

Figure 10. Seasonally decomposed data for China

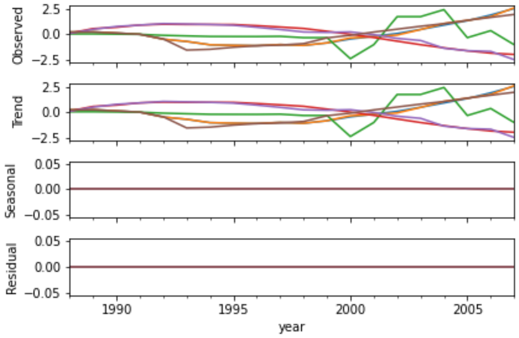

Figure 11. Seasonally decomposed data for Russia

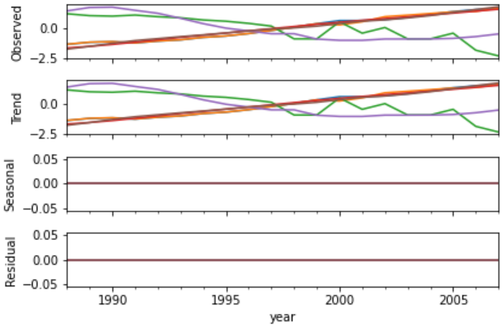

Figure 12. Seasonally decomposed data for USA

Figures 13, 14, 15 and 16 compare the predicted GDPPC for each country (in red) to the actual GDPPC of each country (in black) of the VAR model. The predictions of the training set are not graphed. The red curve represents the GDPPC predictions of the test set, and the black curve represents the actual GDPPC.

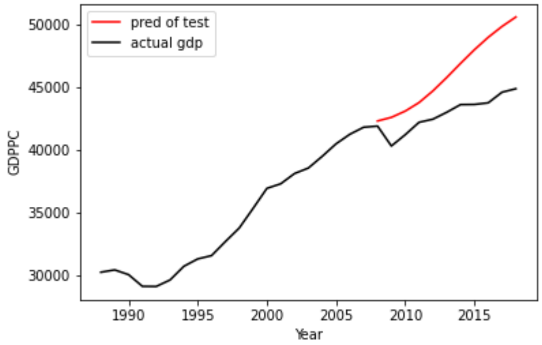

Figure 13. Actual vs Predicted GDPPC for Canada

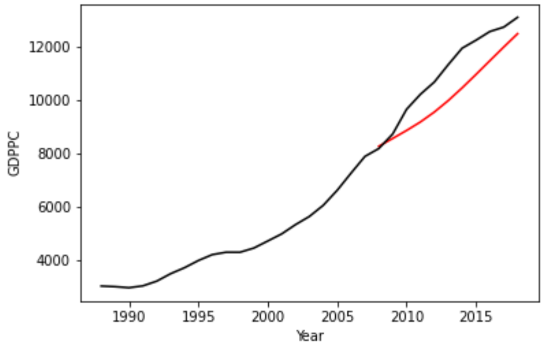

Figure 14. Actual vs Predicted GDPPC for China

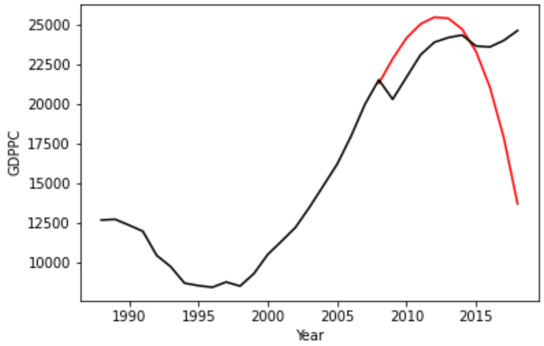

Figure 15. Actual vs Predicted GDPPC for Russia

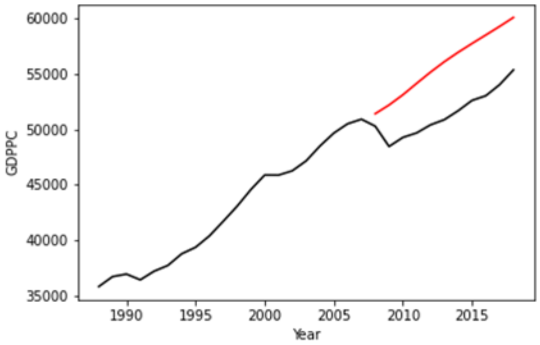

Figure 16. Actual vs Predicted GDPPC for USA

OLS Model

Figure 17 has 4 graphs. Each graph in the figure has 2 lines – the red one represents the predictions of GDPPC for the test set made by the OLS Regression model and the black line represents the actual GDPPC.

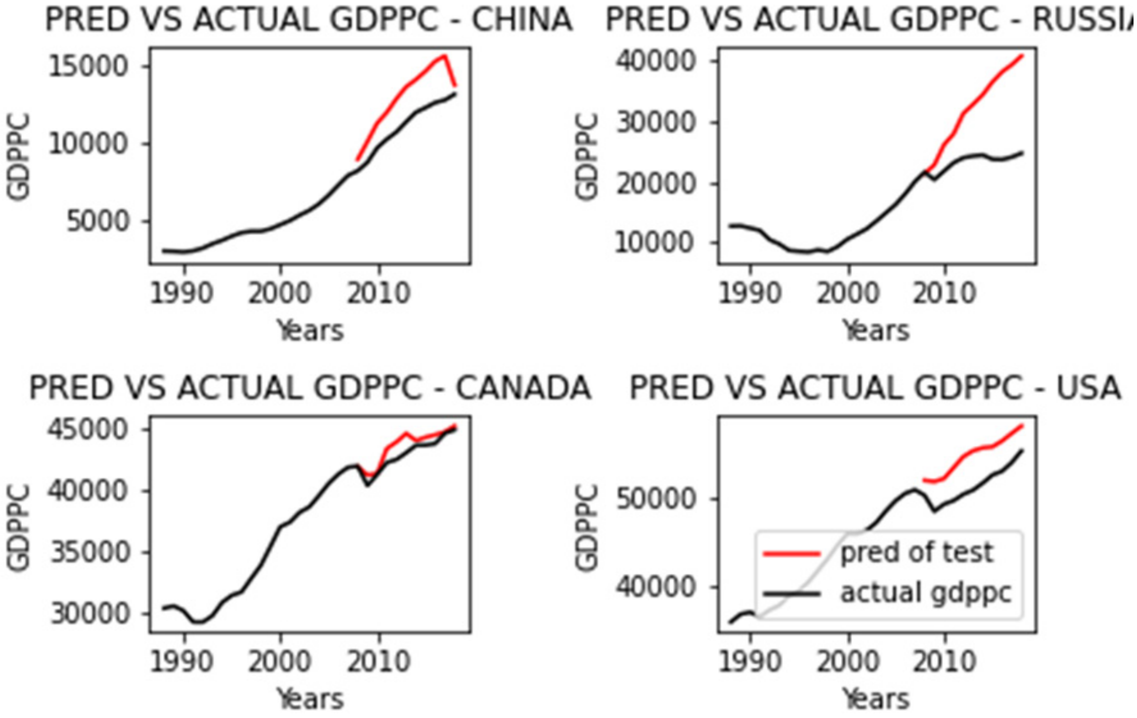

Figure 17. Actual vs Predicted GDPPC for all countries for OLS Regression model

Discussion and Limitations

This paper aimed to quantitively analyse if there is a correlation between the independent variables and the dependent variable using machine learning models. The study includes comparing the models to see if one is better than the other and if the type of political regime in the country plays a role.

Machine learning has been used to predict GDP in the past. Yoon (2020) produced a forecast for Japan's real GDP growth using machine learning models of gradient boosting and random forest model. The results in the paper showed that machine learning models produced a more accurate results than the benchmark forecasts used. Richardson et. al. (2018) also show that machine learning algorithms outperform the traditional statistical models when nowcasting real GDP growth.

In Mohd Zukime Hj Mat Junoh (2004), the paper had a comparative case study between neural network and econometric approaches to predict GDP growth and found that neural network techniques can better predict GDP growth based on certain factors. In a similar vein, this study compared two machine learning models to predict GDPPC. The type of models used differ in both the studies. Previously. Espinoza et al (2011) used VAR models to forecast the GDP of the USA or euro-area using only financial variables. Their use of VAR models helps strengthen our decision to implement this model in the field of economics. While our paper uses a financial variable, it also uses non-financial variables to predict or forecast the GDPPC.

It is also interesting to note that the 2008 economic recession is apparent in some countries. The slight dip in the GDPPC is visible for the USA, Canada, and Russia. It is more apparent in the USA graph and least apparent in China's graph. This is one example of how imputed data doesn't account for any shocks to the other variables. Perhaps if the recession wasn't there, the OLS Regression graph would have made more accurate predictions. This also proves that it is important to look at other confounding variables or events that might affect one or more of the variables being studied.

As mentioned in the results section, the findings indicate that for China and Russia, the VAR model have more accurate forecasts; in the cases of Canada and the USA, the OLS Regression model makes more accurate predictions. This finding could be explained by the idea that USA and Canada have had stable economies i.e., the GDPPC has followed a trend or pattern. So, using past data to forecast future data may be more accurate. For China and Russia, the trend isn't very apparent and slight fluctuations might be seen. In this case, it could be that some/all the independent variables might increase or decrease similarly. So, predicting the dependent variable from the independent variables might have yielded more accurate results. In order to overcome this limitation and confirm this reasoning, future research could address any natural calamity or shocks to the economy in a larger time period using more countries.

Interpretation of Results

While there needs to be a careful analysis of confounding and external factors influencing both GDPPC and the independent variables before considering the independent variables to directly impact the GDPPC, we can offer an interpretation of the statistical correlations between these random variables.

From the correlation matrix, some patterns are apparent. In the non-democratic countries, the Corruption Perception Index, the percentage of population using the internet, Gender Inequality Index, and the Population Density seem to negatively correlate with the GDPPC. In democratic countries, the Corruption Perception Index, Population Density, and the Consumer Price Index correlate negatively with the GDPPC.

The relationship between corruption and economic growth has been a widely debated one. The results of my research show that an increase in corruption, correlates to an increase in GDPPC. While that doesn't sound intuitively correct, previous literature explains this phenomenon. Mallik and Saha (2016) conclude that there is a cubic relationship between growth and corruption. This means that for countries with extremely little corruption, corruption affects the growth negatively but for countries with intermediate corruption, it increases growth and at higher levels, it reduces growth.

We could explain the negative correlation between the Population Density and the GDPPC in the following way: for a country to be successful and entrepreneurial, it needs to generate a large amount of financial capital. Countries with larger populations would be spending a lot more on their inhabitants and so, they might take longer to generate the capital. While the absolute volume of the capital will keep growing because of the large population, the capital-per-capita will fall. So, each worker will have lesser capital to begin with, causing them to be less productive which causes a lower GDPPC.

Our finding of Consumer Price Index negatively impacting the GDPPC of Canada, USA, and Russia is consistent with the findings of Fischer (1993). Fischer concludes that inflation (i.e., what Consumer Price Index measures) reduces economic growth by reducing investment and productivity.

The pattern of results of the Gender Inequality Index's correlations are broadly consistent with the results of an IMF (2020) Working Paper. Their study concludes that gender inequality causes un-tapped female potential to be wasted. Our results indicate that in China, Russia, and the USA – the Gender Inequality Index has a negative correlation with the GDPPC.

Other than the results of Canada, our findings for the percentage of the population using the internet are inconsistent with previous literature. For example: Waqar (2015) finds a positive correlation between ICT measures and GDP per worker.

Even though these countries are in no way representative of the regime as a whole, the results suggest that VAR models forecast the GDPPC of democratic regimes better, while OLS Regression predict the GDPPC of non-democratic regimes more accurately. However, it should also be considered that the democratic countries have a more capitalist economy, while the non-democratic regimes lean more towards being planned economies. But as previous literature states: planned economies are a characteristic of non-democratic regimes. Another point of importance is highlighted in Congressional Research Service (2013) which talks about how the Chinese governments modified its currency policy from 2005 to 2008 and then again resumed it in 2010. It then devalued the yuan. And as discussed in previous literature, authoritarian regimes are less likely to report their actual GDP. This means that our model may lack key independent variables to make effective predictions and therefore our VAR predictions are systematically worse. The OLS Regression models does well because it seems to extrapolate a linear trend that predicts that the GDPPC will keep increasing every year. This means that the VAR model might make better predictions for even the non-democratic countries if given the appropriate variables.

Limitations

Certain limitations of this study could be addressed in future research. For example, the number of rows of data used was limited and even in that, a part of it was imputed. Using more rows of data to train and test the models would yield better outputs. As mentioned above, the countries picked are not representative of their population and using more countries would help generalise the results better. The list of 7 independent variables used in this study is not exhaustive.

GDPPC has certain factors used to calculate it, this paper aimed to look at other variables that may correlate to it. To prove causation, using more variables would be necessary. Using the Granger Causality test on the variables used, would give an idea of the other variables that may influence the GDPPC. We feel that further research examining the effect of different independent variables may help better understand what affects the GDPPC.

At the same time, it must be mentioned that GDPPC is not always the best method to measure a country's economic success. van den Bergh (2010) elaborates on why GDP, and by extension GDPPC, should be taken with a grain of salt. In the past, research has been conducted as to why GDPPC or GDP isn't the best metric. It is, however, one of the more popular ones.

Despite these limitations, the present study has enhanced our understanding of the relationship between some variables and the economy of varying political regimes. We hope that the current research will stimulate further investigation of this important area.

Conclusion

Applications of machine learning in political science and economics is a growing field of study. A lot of research has been done to empirically analyse trends and make predictions or suggestions in this niche. This paper applied the machine learning models of OLS Regression and VAR to predict the GDPPC of 4 countries of varying political regimes. Both these models differ from each other by their method of prediction. OLS Regression uses the independent variables to predict the dependent variable and treats each data point independent of any past or future data points. In VAR, the outputs are forecasted by taking past data into consideration.

Making accurate predictions or forecasts for economic indicators like GDPPC is important because it helps governments plan for the upcoming fiscal years and any related policies. While there are certain factors that are used to calculate the GDPPC already, this paper looked for a correlative relationship between the 1 dependent and 7 independent variables.

The frameworks for these models were developed using Python in Google Collaboratory. The outputs supported the idea that OLS Regression models make more accurate predictions for the GDPPC for non-democratic countries and VAR models make more accurate forecasts for democratic countries. However, as discussed in the previous section, our model may lack key independent variables because of which the VAR predictions are not more accurate in the case of non-democratic countries. The OLS Model did well because it seems to detect the linear trend of increasing GDPPC.

We then proceeded to analyse the results and their implications. Using a correlation matrix, conclusions about factors influencing the economies of democratic and non-democratic economies were drawn. However, it is essential to remember that these conclusions have been limited by the number of variables used. We interpreted the statistical correlations which indicated that Corruption Perception Index and Population Density negatively correlates with the GDPPC of all 4 countries. We also noticed that the percentage of people using the internet and Gender Inequality Index correlates negatively with the GDPPC for non-democratic countries and in democratic countries the Consumer Price Index negatively influences the economy. A discussion about the limitations and avenues for future research was also done.

Much work remains to be done before a full understanding of the extent of the correlation and/or causation of different independent variables with the dependent variables is established. This research can be seen as a step further towards integrating machine learning frameworks in economic and political science fields.

Acknowledgments

I would like to thank my mentor, Swapneel Mehta, for his exceptional support. This project would not have been possible without his guidance and expertise. I have benefited greatly from this experience. Many thanks to Tyler Moulton, for his invaluable advice through this process and endless patience. Thanks to Lumiere Education for giving me this wonderful opportunity. And finally, thanks to my family for their endless love and support.

References

Przeworski, A., & Limongi, F. (1993). Political regimes and economic growth. *Journal of Economic Perspectives*, *7*(3), 51-69. https://doi.org/10.1257/jep.7.3.51

The Global Competitiveness Report 2017-2018. World Economic Forum. (2021). Retrieved 18 August 2021, from https://www.weforum.org/reports/the-global-competitiveness-report-2017-2018.

GEM Global Entrepreneurship Monitor. (2021). Retrieved 18 August 2021, from https://www.gemconsortium.org/wiki/1367.

Mathews, R. (2012, September 19). *GDP and the US Economy: 3 ways to measure economic production*. Mic. https://www.mic.com/articles/14943/gdp-and-the-us-economy-3-ways-to-measure-economic-production

Zhang, L., Kinser, K., & Shi, Y. (2014). World Economies and the Distribution of International Branch Campuses. *International Higher Education*, (77), 8-9. https://doi.org/10.6017/ihe.2014.77.5674

Khan, K., Batool, S., & Shah, A. (2016). Authoritarian Regimes and Economic Development: An Empirical Reflection. *The Pakistan Development Review*, *55*(4I-II), 657-673. https://doi.org/10.30541/v55i4i-iipp.657-673

Yoon, J. (2020). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Computational Economics*, *57*(1), 247-265. https://doi.org/10.1007/s10614-020-10054-w

Magee, C., & Doces, J. (2014). Reconsidering Regime Type and Growth: Lies, Dictatorships, and Statistics. *International Studies Quarterly*, *59*(2), 223-237. https://doi.org/10.1111/isqu.12143

Fernando, J. (2021). *Gross Domestic Product (GDP)*. Investopedia. Retrived from https://www.investopedia.com/terms/g/gdp.asp.

GDP per capita. Dictionary.cambridge.org. (2021). Retrieved 2021, from https://dictionary.cambridge.org/dictionary/english/gdp-per-capita.

Rosendorff, B., Hollyer, J., & Vreeland, J. (2011). Democracy and Transparency. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1750824

Democracy under Siege. Freedom House. (2021). Retrieved 18 August 2021, from https://freedomhouse.org/report/freedom-world/2021/democracy-under-siege

Bolt, J., & Luiten van Zanden, J. (2021). *Maddison-Project Working Paper WP-15*. Rug.nl. Retrieved 18 August 2021, from https://www.rug.nl/ggdc/historicaldevelopment/maddison/publications/wp15.pdf.

Congressional Research Service. (2013, July 22). *China's currency POLICY: An analysis of the economic issues*. EveryCRSReport.com. https://www.everycrsreport.com/reports/RS21625.html#_Toc362345928.

Facebook. Facebook. (2021). Retrieved 18 August 2021, from https://www.facebook.com/TransparencyInternational/videos/586267088480385/.

Corruption Perceptions Index. DataHub. (2021). Retrieved 2021, from https://datahub.io/core/corruption-perceptions-index#resource-corruption-perceptions-index_zip.

Freedomhouse.org. (2021). Retrieved 2021, from https://freedomhouse.org/sites/default/files/2021-02/Country_and_Territory_Ratings_and_Statuses_FIW1973-2021.xlsx.

Freedomhouse.org. (2021). Retrieved 2021, from https://freedomhouse.org/sites/default/files/FIW%20Methodology%20Fact%20Sheet.pdf.

Itu.int. (2021). Retrieved 2021, from https://www.itu.int/en/ITU-D/Statistics/Documents/coreindicators/Core-List-of-Indicators_March2016.pdf.

ITU - Organizations - "FAO catalog". Data.apps.fao.org. (2021). Retrieved 2021, from https://data.apps.fao.org/catalog/organization/itu.

&ensp;. (2021). Retrieved 2021, from https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42&sId=1479329334655.

Gender Inequality Index (GII) | Human Development Reports. Hdr.undp.org. (2021). Retrieved 2021, from http://hdr.undp.org/en/content/gender-inequality-index-gii.

Population density | Data Catalog. Datacatalog.worldbank.org. (2021). Retrieved 2021, from https://datacatalog.worldbank.org/population-density-people-sq-km-land-area.

Richardson, A., Mulder, T., & l Vehbi, T. (2018). Nowcasting New Zealand GDP using machine learning algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3256578

Predicting GDP growth in Malaysia using knowledge-based economy indicators : a comparison between neural network and econometric approaches - Sunway Institutional Repository. Eprints.sunway.edu.my. (2021). Retrieved 2021, from http://eprints.sunway.edu.my/9/.

Espinoza, R., Fornari, F., & Lombardi, M. (2011). The Role of Financial Variables in predicting economic activity. *Journal Of Forecasting*, *31*(1), 15-46. https://doi.org/10.1002/for.1212

Mallik, G. and Saha, S. (2016), "Corruption and growth: a complex relationship", *International Journal of Development Issues*, Vol. 15 No. 2, pp. 113-129. https://doi.org/10.1108/IJDI-01-2016-0001

Fischer, S. (1993). The role of macroeconomic factors in growth. *Journal Of Monetary Economics*, *32*(3), 485-512. https://doi.org/10.1016/0304-3932(93)90027-d

Bertay, A., Dordevic, L., & Sever, C. (2021). IMF. Retrieved 20 August 2021, from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwirzOG-ssDyAhVbyzgGHSiHATMQFnoECBwQAQ&url=https%3A%2F%2Fwww.imf.org%2F-%2Fmedia%2FFiles%2FPublications%2FWP%2F2020%2FEnglish%2Fwpiea2020119-print-pdf.ashx&usg=AOvVaw0OrU0BBAzkwrg0IbgRNyRQ.

WAQAR, J. (2021). *Impact of ICT on GDP per worker: A new approach using confidence in justice system as an instrument. : Evidence from 41 European countries 1996- 2010*. DIVA. Retrieved 20 August 2021, from http://www.diva-portal.org/smash/record.jsf?pid=diva2:931181.