Investigating Globalization by Country Attributes

Vid Chan, Anderson Monken, Douglas Neumann, Nicole Yoder

Georgetown University

Course: Statistical Learning For Analytics

Spring 2020

Table of Contents

Introduction, Audience, and Motivation

Throughout history, the elements that define a country change at a nearly constant rate. Borders, political system, and population size are some of the first that come to mind. However, perhaps the most important indicator of the general health or future of a country lies with the economic trends that country is experiencing. The International Monetary Fund (IMF), Organization for Economic Co-operation and Development (OECD), and World Bank (WB) all update datasets concerning different yearly economic indicators for respective countries. These datasets provided both the ability to review historical data as well as the necessary information to model how the different indicators interact. By cleaning and combining data sets from these different sources, our group is focused on identifying economic and societal trends through the lens of globalization. Specifically how globalization, defined in our research as the number of trading partners, is tied to a wide range of indicators: from quality of life variables, such as tourism, access to cell phones, and life expectancy, to fundamental economic statistics.

Economies are more connected than they have ever been. They share common interests in systems like shipping and production. These commonalities provide for a wider web to be cast, allowing more countries, independent of size or military might, to take part in the world's economy. Globalization is defined as "the development of an increasingly integrated global economy marked especially by free trade, free flow of capital, and the tapping of cheaper foreign labor markets" (Merriam-Webster, n.d.). Inevitably, there are gains and losses to entry in such a system. Some economies will suffer, while others stand to gain enormously. Many of the positive impacts of globalization are well known, such as the exchange of information and connectivity, rise in per capita income, and an increase in civil liberties. However, there are also some negatives. Other countries are able to exert pressure on trade partners, once-abundant environmental resources may be depleted, and long-held jobs could be outsourced to other countries (Hufbauer, 2019).

The audience for this research is those who have an interest in global trade; specifically, organizations like the United Nations, International Monetary Fund, and World Bank. These organizations would gain from direct knowledge of how changes within a country impact globalization. Also, the leaders of countries have a vested interest in understanding the complex changes that economies experience upon entry in the world economy.

The motivation for this research lies with providing information to the leaders of countries who may have an interest in shifting their economy to a more globalized stance, or perhaps may see a benefit in a reduction of globalization. Our research offers those leaders information on how aspects of their society and economy interact based on their decision.

High-Level Problem Definition and Goal

The research for this project focuses on understanding how changes in various attributes of a country relates to their level of globalization. The goal for our research is not focused on prediction but inference and interpretability through identifying the attributes most closely associated with globalization and modeling how they interact.

Related Work

Globalization is an often-studied aspect of the world economy. There has been a number of studies that looked at the relationship between different predictors and outcomes concerning a more interconnected economy. Two articles had a similar focus as our research. The first was a report by the World Bank on “Globalization's Impact on Gender Equality.” The second was a paper written by Marco Moretti titled “Globalization of Mobile and Wireless Communications: Bridging the Digital Divide.”

The 2011 report from the World Bank offered some insight into how gender roles are changing. Some of the key points of the report were to explain how costly it is for economies to not integrate genders and therefore limiting participation in their economy. Other points were made about how globalization causes countries to enter in trade agreements that limit gender discrimination. Lastly, globalization makes wages for all members of the workforce more fair (World Bank, 2012).

Moretti's 2010 paper on the wider access to cell phone use offers insight into how globalization has increased access to information and communication technology. Countries that take part in the world economy are more likely to have access to information. Conversely, countries that haven't had goods and service costs reduced by the benefits of globalization suffer from lack of access (Moretti, 2010).

The Data

Globalization can be measured in a variety of ways usually related to trade. However, the volume of trade (the sum of imports and exports of goods and services) cannot be used directly to measure globalization, since it is not standardized for the size of the country. We considered three other options: the number of trade partners a country has, a country's trade measured as a percentage of their GDP, or the percent change of a country's trade volume from year to year. However, while the second two options seem to be standardized, they did not make sense for our analysis under further scrutiny. The measures give insight into how a country's trade volume is changing and how important trade is to their economy, but do not make two countries' levels of globalization discernable in comparison.

Table A1 demonstrates this by comparing the United States to Kazakhstan over a few years. The first measure clearly indicates that the US is more globalized than Kazakhstan. The second measure shows how the countries are changing, but only gives insight into how the countries trade is changing over time, rather than relative to other countries. The third measure demonstrates how important trade is to a country's economy, but countries with large economies like the US may have smaller relative trade percentages, even if they are globalized. Therefore, we decided to use the number of trade partners as the globalization measure and our target variable.

The data for the number of trade partners came from the World Integrated Trade Solution website, which is a partnership between the World Bank, the UN, and other international trade organizations. Originally, it was split by import partners and export partners, but the two numbers were very similar for each country and some of the export partners' data was missing. Therefore, we decided to use the max value from those two columns as the singular trade partners variable. The other variables used as predictors, as well as their data sources, are summarized in Table A2.

The variables had a range of years for available data, some going back all the way to 1960, but most did not start until the 1980's or 90's. We decided to limit the years to 1990-2016, since that would reduce the amount of missing data significantly, and we only kept countries that showed up in all four data sources. However, many of the variables were missing some years for certain countries. Countries were removed entirely if they were missing more than 20% of their values. Otherwise, the missing values in each column were imputed by country using the pandas function interpolate. "Interpolate" filled in missing values between two available years linearly, while missing values at the beginning or end were filled in with the next closest value. Table A3 shows an example. After cleaning was completed, we were left with data for 99 countries over 27 years (1990-2016) each.

Exploratory Data Analysis

Figures A1 and A2 show the trends of countries in the top ten and bottom ten number of trading partners in selected years. Figure A3 shows the ranking of the average number of trading partners for top ten and bottom ten countries. The United Kingdom comes in first in the average ranking. Kyrgyzstan comes in last. All three figures (A1, A2, and A3) provide insights into some of these economies and their upward or downward movement into the top ten and bottom ten lists. Specifically, Figure A1 suggests some incredible upward movements of different economies into the top ten most trading partner status. Countries like Thailand, South Korea, Mexico, and Poland have shown remarkable engagement in global trade as measured by the number of trading partners. Thailand was in 7th position amongst the top ten in 2009. It then jumped to the 6th and 4th in 2012 and 2016 respectively. South Korea was in the 9th and then 8th positions in 2009 and 2016. These impressive achievements deserve further investigation, which is shown in Figures A5 and A6.

Figure A4 provides a correlation matrix between dependent and independent variables in the dataset. Blue circles represent positive correlation, while red circles represent negative correlation. The plot shows a strong positive correlation between the number of trading partners (Max_Partners) and percentage of mobile phone subscription (mobilesub_per100peeps). Some other noticeable variables such as number of international tourist arrivals, agricultural percentage of GDP, GDP per capita, and GDP per CO2 unit, show some correlations (positive and negative) with the main output variable. These variables are then plotted in Figures A5 and A6 for the two selected countries--Thailand and South Korea.

For Thailand, five out of six plots show consistent positive trends between the number of trading partners and independent variables. However, the scatter plot for agricultural percentage of GDP does not clearly show a trend. All six plots in the South Korea case show consistent trends. With an exception of the agricultural percentage plot, the rest of the plots in Figure A6 show positive trends. These plots and the correlation matrix provide further insights of the relationship between some of these variables and the output variable. This will be investigated in further detail later in the modeling section.

A principal component analysis is conducted to examine the relationship between all independent variables and to see any additional trends between different economies. In Figure A7, each number represents one specific country. For example, number 42 represents India. Many countries seem to cluster around the zero values of PC1 and PC2 according to the plot. This indicates that there are few isolated countries.

Modeling Methods and Results

Different modeling techniques are employed to help understand the relationship between the number of trading partners and various predictors. The entire dataset is split into training and testing sets using random sampling (75%-25% respectively). The train set was then used to build various models. Mean squared error (MSE) for the test set is used to evaluate the effectiveness of models.

The first model considered was linear regression. Three different model selections such as best subset, forward stepwise, and backward stepwise selections are used to find optimal number of predictors. Figures A8, A9, and A10 were the results of the methods. Cp, BIC, Adjusted R-squared are metrics used to compare model fits on the training data to the best model subset of the predictor variables. Cp and adjusted R-squared for all three model selection methods show that eight predictors would provide the lowest test MSE. Those eight predictors are year, GDP per unit of energy-related CO2 emissions, purchasing power parity conversion rate, government revenue, GDP per capita, agricultural percentage of GDP, international tourist arrival, and rural population percentage. These eight variables were then used as predictors in subsequent models.

Ridge and lasso regressions were performed using all the predictors derived from the three model selection methods above. Cross-validation was then performed to pick tuning parameter lambda for each regression method. Ridge regression has an optimal log(lambda) around 2.19 (Figure A11) while lasso regression's optimal log(lambda) is around 0.057 (Figure A12). The relative performances of their test MSEs are discussed at the end of this section.

The Q-Q plot in Figure A13 shows that linear regression does well in the middle part, but around the boundaries, the linear relationship between the number of trading partners and those predictors are not too stable. Figure A13 also shows the regression outputs for the linear regression model. Overall, estimates of the coefficients of the two regressions are comparable. The signs of all coefficients are as expected. According to the regression output, an increase of the number of international tourists by 10 million, 9.96 number of trading partners will be increased.

Figures A14 and A15 used decision tree regression models instead of linear regression to examine the relationship between the output and the predictive variables with the goal of predicting the number of trading partners. The unpruned tree (Figure A14) shows international tourist arrival (root node), mobile phone subscriptions, and agricultural percentage of GDP are the three highest nodes in the model. The pruned tree (Figure A15) shows an approximately even split of results across four leaves. The unpruned tree significantly improves the test MSE to 581.05, while pruning the tree raises the test MSE to 793.94.

A number of decision tree ensemble methods were considered to improve on the performance of the decision tree regression. Figure A16 shows the relationship between five-fold cross validation test MSE and the number of trees for random forest models for p=16 (bagging), p/2=8, and sqrt(p)=4 variables available at each tree node. The forest models were grown from one to 1,000 trees, with only 500 trees shown. The test MSE was found to level out at 150 trees at all variable choices, with the random forest employing four variables for each tree node having the lowest test MSE among the three. Figures A17 and A18 display the depth and variable importance for the bagging and random forest models.

Gradient boosting trees analysis considered a variety of shrinkage values (0.001,0.01,0.1,0.5), trees (1-2000), and interaction depth (1-8). The models were evaluated using 5-fold cross validated test MSE. Figure A19 and figure A20 show plots of the model performance results. Generally more trees and larger interaction depth have better model performance. Shrinkage value of 0.01 provides the smoothest decline of CV test MSE as trees are added to the model. The best model has hyperparameters trees = 2000, interaction depth = 8, and shrinkage = 0.01. Figure A21 shows the variable importance for the gradient boosting model.

After running each of these models mentioned in this section, test MSE is obtained by applying the train model to the test set. Table A4 summarizes all of the different models with their corresponding test errors. The standard linear regression has the largest test MSE. Ridge and lasso regressions have similar error values. Decision tree method’s test MSE is in the middle relative to the rest. Bagging, random forest, and the gradient boosted trees have the three lowest test MSE values compared to the rest. The best model, with the lowest test MSE, is the gradient boosted model.

Conclusion and Discussion

The modeling methods in this paper broadly agree on the most important variables that are linked to globalization. The decision tree, random forest, and boosted tree models all have international tourism as the most important variable to predict the number of trading partners. Linear regression also puts international tourism as an important variable. Figure A18 shows that in both bagging and random forest models that international tourism is the most important predictor. Bagging places almost all importance on international tourism, while random forest provides additional views on the data and shows other important variables like mobile subscriptions and gross domestic product by purchasing power parity.

Globalization is a driving force of the international economy, and understanding the dynamics behind its growth will help countries examine globalization's effects. Based on our findings, the most important takeaways for developing countries that want to globalize and international development organizations are to encourage and promote tourism, and maximize cell phone access, raise life expectancy, and shift their economy to non-agricultural industries. Other variables that cannot be easily changed but are also important for globalization's reach include gross domestic product based on purchasing power parity and population size. All of these aspects of society and the economy serve as the best indicators of the level of globalization.

Future Work

Based on recent world events the trajectory of globalization could potentially be altered. The Covid-19 crisis poses a threat to the continued expansion of globalization. Governments could move away from international trade deals and focus more on self-reliance (Stiglitz et al., 2020). While this may alter the scope of the future of the world economy, the relationship between predictors of globalization done in our research was not impacted by the current crisis.

The crisis offers the potential to provide support to some of our findings. Specifically, how citizens' quality of life is impacted if an economy were to shift from a trade focus to a self-reliance focus. In addition, looking into how a decrease in global travel impacts other aspects may potentially reveal more unexplored relationships. Given a more robust data set in the coming decade, those statistics could be used to further address some of our conclusions.

References

Collins, M. (2015, May 6). *The Pros And Cons Of Globalization.* https://www.forbes.com/sites/mikecollins/2015/05/06/the-pros-and-cons-of-globalization/#45332af8ccce

Merriam Webster. (n.d.). *Globalization*. https://www.merriam-webster.com/dictionary/globalization

Hufbauer, G.C. (2019, June 10). *Globalization Facts and Consequences.* https://www.piie.com/commentary/speeches-papers/globalization-facts-and-consequences

World Bank. (2012). Globalization's Impact on Gender Equality: What's Happened and What's Needed. *World Development Report 2012, September 2011*, 254-278. https://doi.org/10.1596/978-0-8213-8810-5

Moretti, M. (2010). Globalization of Mobile and Wireless Communications: Bridging the Digital Divide. *Globalization of Mobile and Wireless Communications Signals and Communication Technology*, 19–29. https://doi.org/10.1007/978-94-007-0107-6_3

Stiglitz, J. E., Shiller, R. J., Gopinath, G., Reinhart, C. M., Posen, A., Prasad, E., … Mahbubani, K. (2020, April 15). *How the Economy Will Look After the Coronavirus Pandemic.* https://foreignpolicy.com/2020/04/15/how-the-economy-will-look-after-the-coronavirus-pandemic/

Data Sources

International Monetary Fund (World Economic Outlook):

https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/download.aspx

World Integrated Trade Solution:

https://wits.worldbank.org/Default.aspx?lang=en

https://wits.worldbank.org/CountryProfile/en/Country/BY-COUNTRY/StartYear/1988/EndYear/2017/Indicator/NMBR-MPRT-PRTNR

https://wits.worldbank.org/CountryProfile/en/Country/BY-COUNTRY/StartYear/1988/EndYear/2017/Indicator/NMBR-XPRT-PRTNR

Organization for Economic Co-operation and Development:

https://stats.oecd.org/

World Bank:

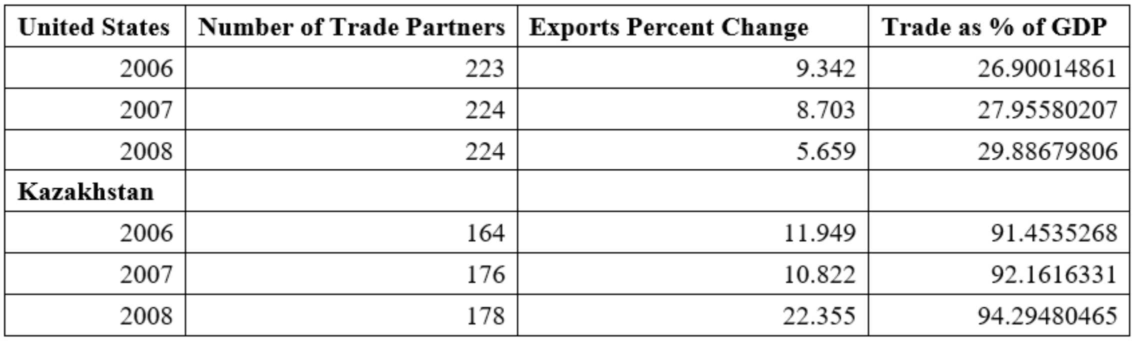
https://data.worldbank.org/indicator

Link to Code Repository

https://github.com/AndersonMonken/ANLY512_Spring2020_Project

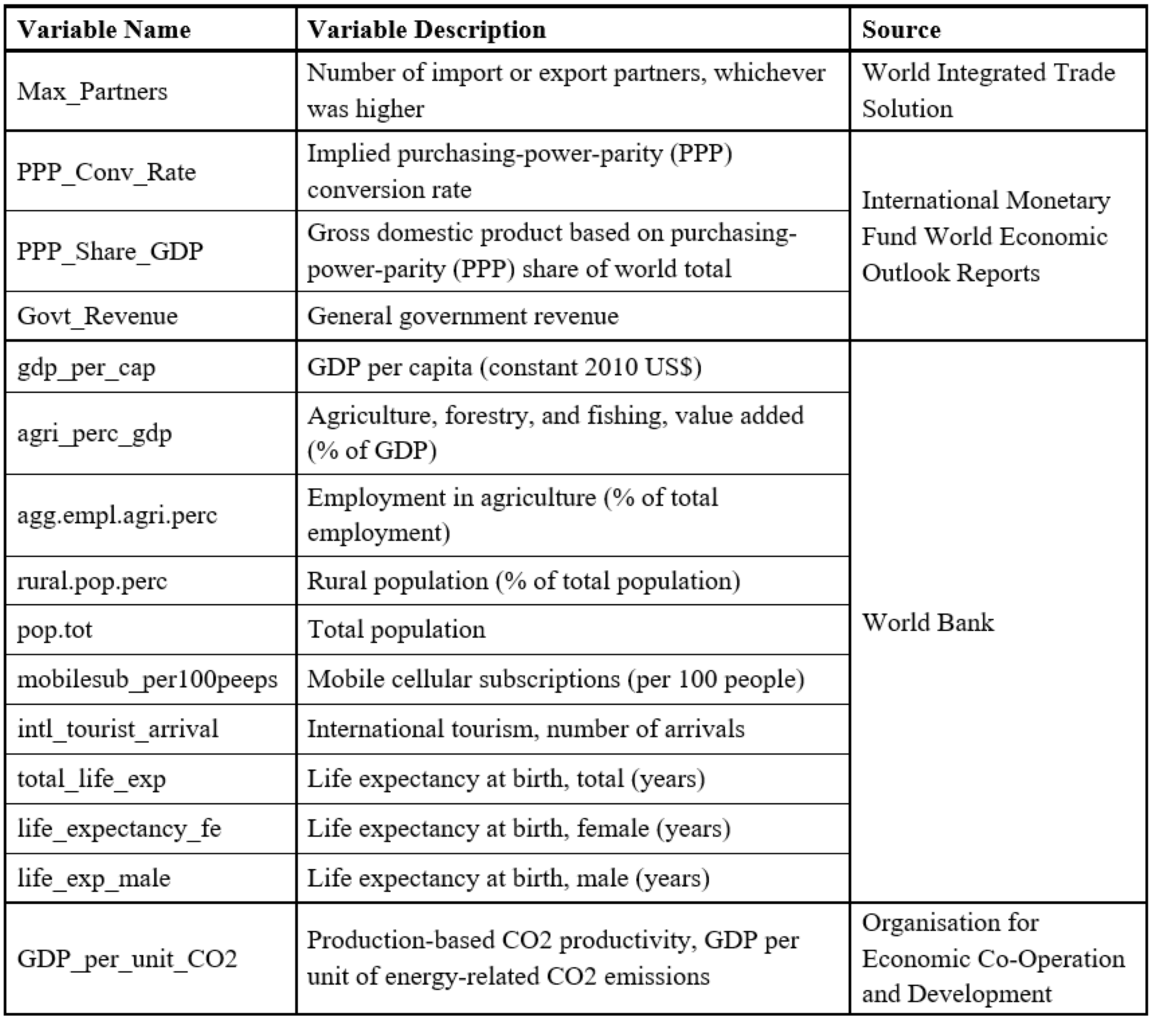
Appendix

Table A1: Comparison of Possible Globalization Measures



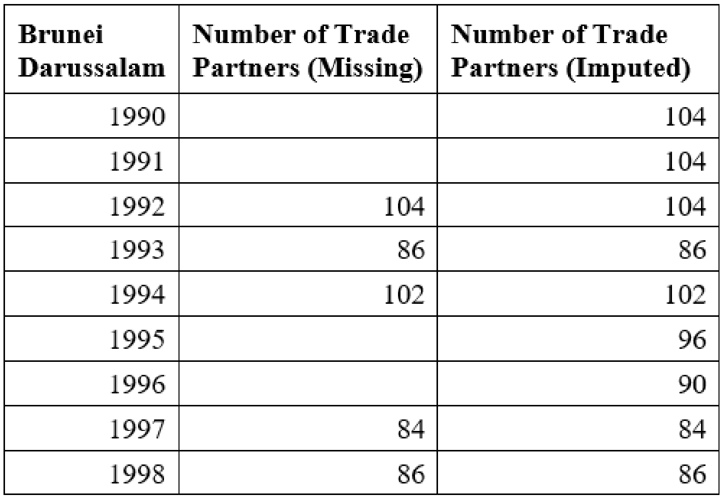
United States	Number of Trade Partners	Exports Percent Change	Trade as % of GDP
2006	223	9.342	26.90014861
2007	224	8.703	27.95580207
2008	224	5.659	29.88679806
Kazakhstan			
2006	164	11.949	91.4535268
2007	176	10.822	92.1616331
2008	178	22.355	94.29480465

Table A2: Variable Descriptions and Sources



Variable Name	Variable Description	Source
Max_Partners	Number of import or export partners, whichever was higher	World Integrated Trade Solution
PPP_Conv_Rate	Implied purchasing-power-parity (PPP) conversion rate	International Monetary Fund World Economic Outlook Reports
PPP_Share_GDP	Gross domestic product based on purchasing-power-parity (PPP) share of world total	
Govt_Revenue	General government revenue	
gdp_per_cap	GDP per capita (constant 2010 US$)	World Bank
agri_perc_gdp	Agriculture, forestry, and fishing, value added (% of GDP)	
agg.empl.agri.perc	Employment in agriculture (% of total employment)	
rural.pop.perc	Rural population (% of total population)	
pop.tot	Total population	
mobilesub_per100peeps	Mobile cellular subscriptions (per 100 people)	
intl_tourist_arrival	International tourism, number of arrivals	
total_life_exp	Life expectancy at birth, total (years)	
life_expectancy_fe	Life expectancy at birth, female (years)	
life_exp_male	Life expectancy at birth, male (years)	
GDP_per_unit_CO2	Production-based CO2 productivity, GDP per unit of energy-related CO2 emissions	Organisation for Economic Co-Operation and Development

Table A3: Example of Data Imputation



Brunei Darussalam	Number of Trade Partners (Missing)	Number of Trade Partners (Imputed)
1990		104
1991		104
1992	104	104
1993	86	86
1994	102	102
1995		96
1996		90
1997	84	84
1998	86	86

Table A4: Test MSE for the Models

Model	Note	Test MSE
Ridge Regression	Best lambda is 2.19.	851.71
Lasso Regression	Best lambda is 0.056.	845.30
Linear Regression	Q-Q plot shows non-linear relationship at the top and bottom quantiles of the output. (Figure A13)	909.15
Decision Tree Regression	Figure A14 shows nodes for this tree. (unpruned)	581.05
Bagging Trees	Best model: trees = 150	268.69
Random Forest	Best model: trees = 200, mtry = 4	239.98
Boosted Trees	Best hyperparameters: shrinkage=0.01, ntree=2000, interaction.depth=8	218.72

Figure A1: Top 10 Countries for Selected Year

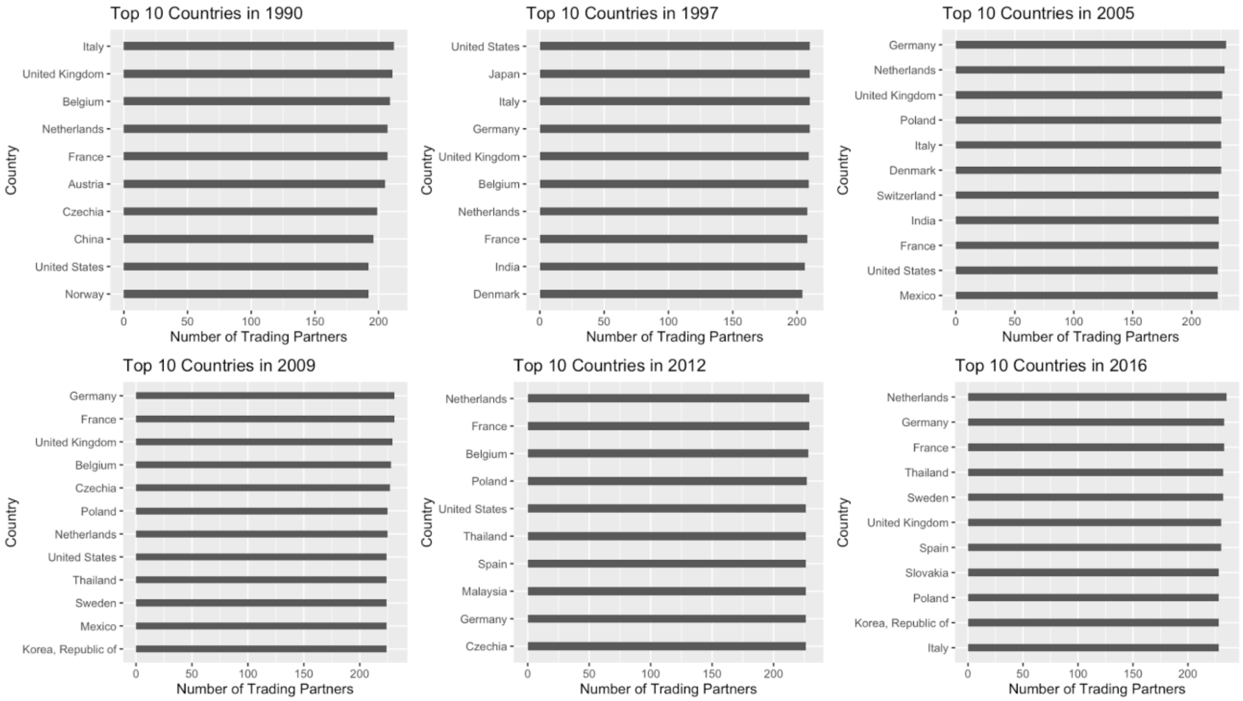


Figure A2: Bottom 10 Countries for Selected Year

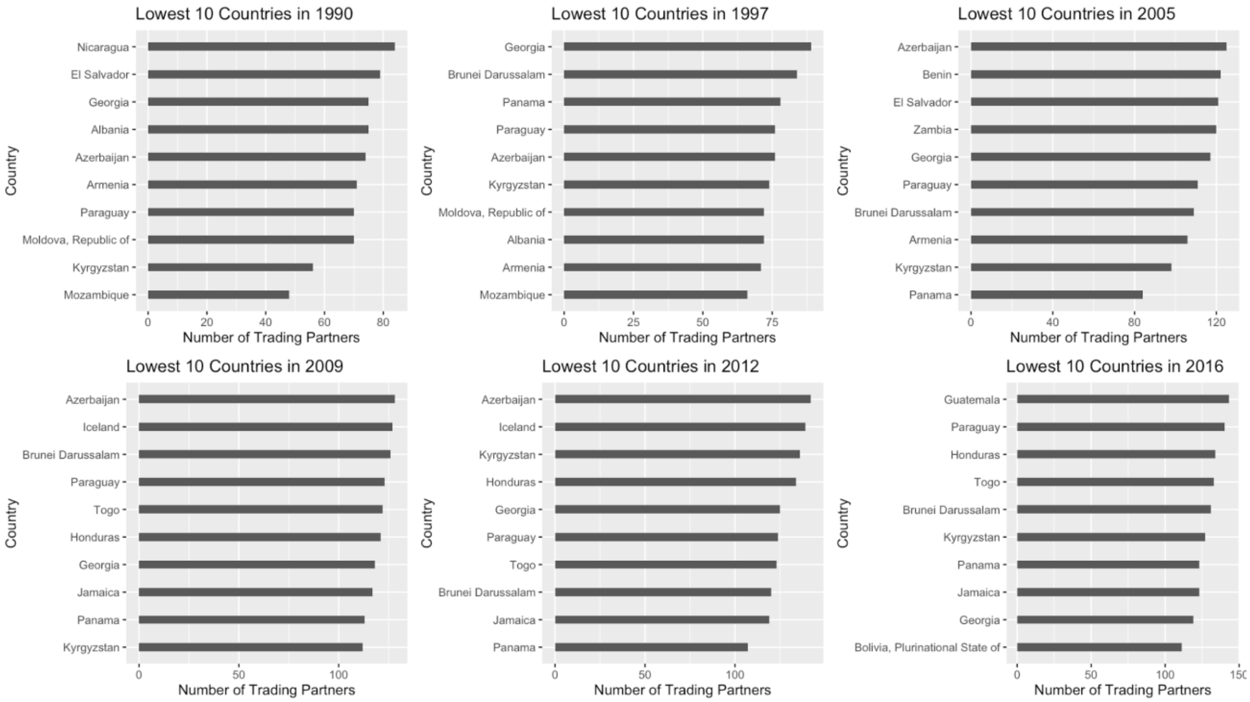


Figure A3: Average Number of Trading Partners from 1990-2016

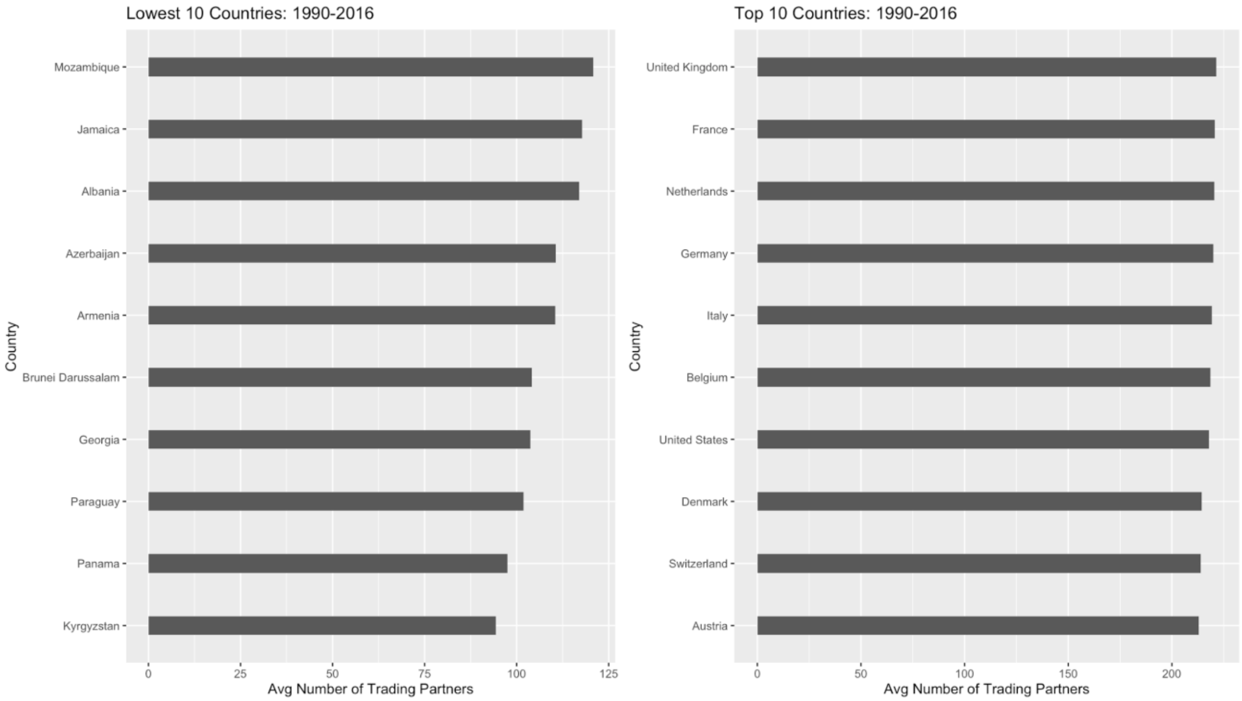


Figure A4: Correlation Plot

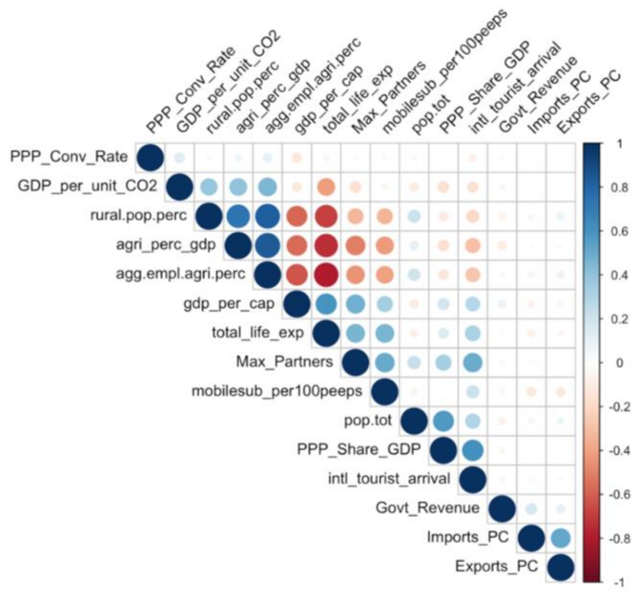


Figure A5: Thailand Case

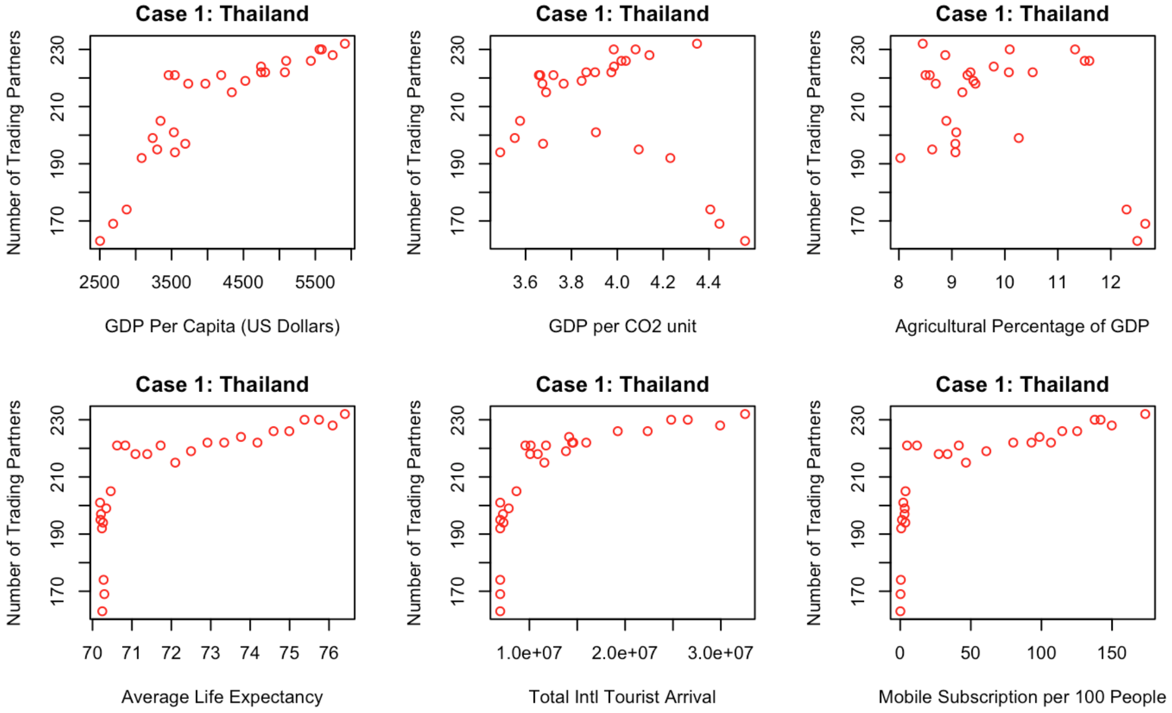


Figure A6: South Korea Case

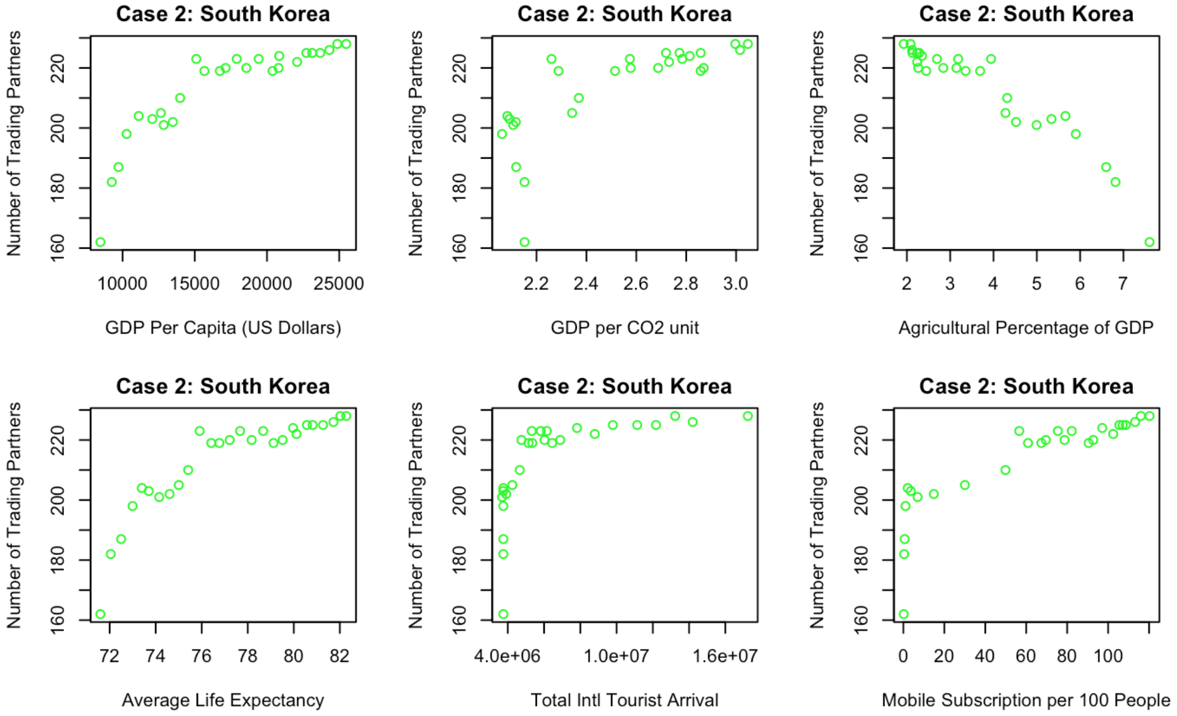


Figure A7: Principal Component Analysis

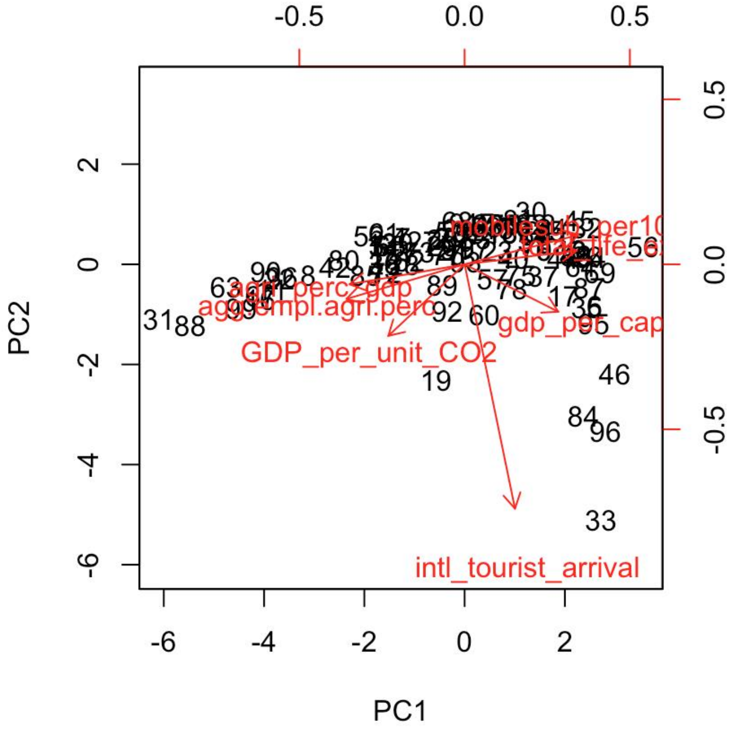


Figure A8: Best Subset Selection

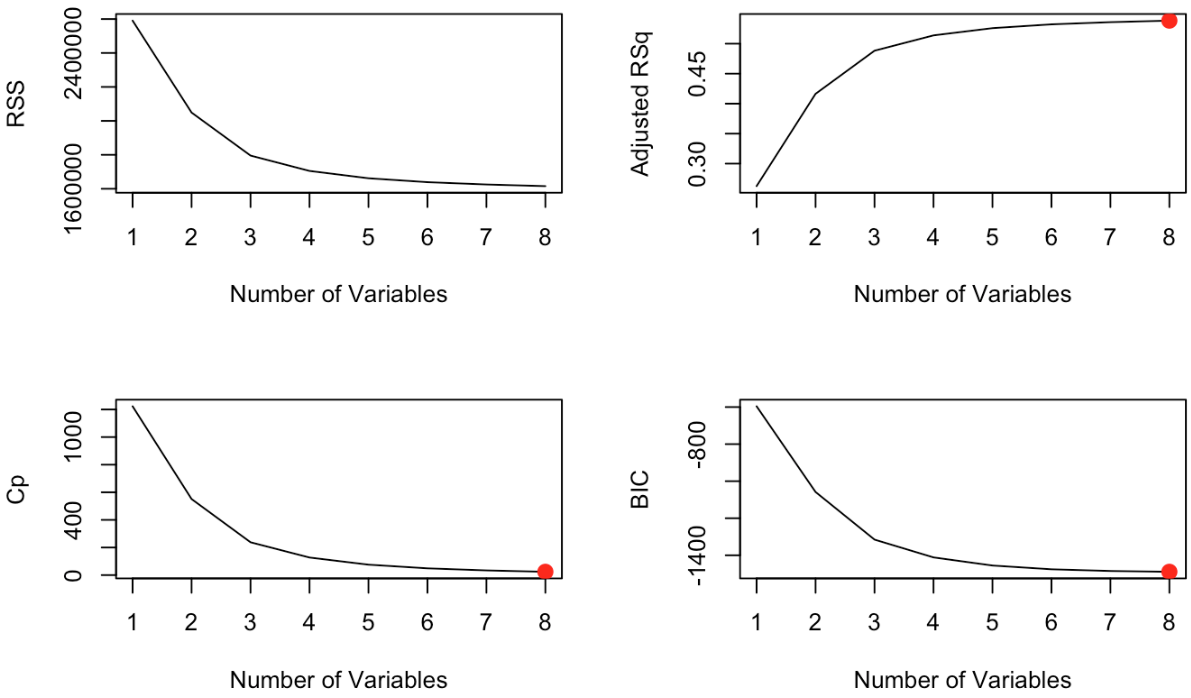


Figure A9: Forward Stepwise Selection

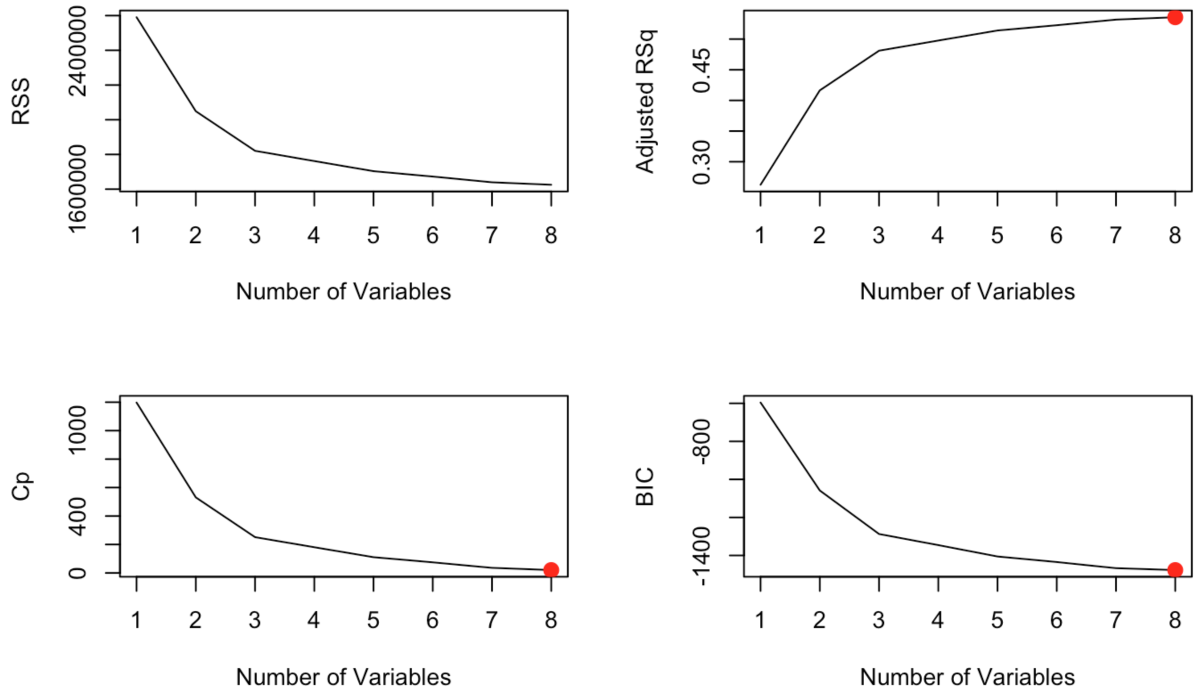


Figure A10: Backward Stepwise Selection

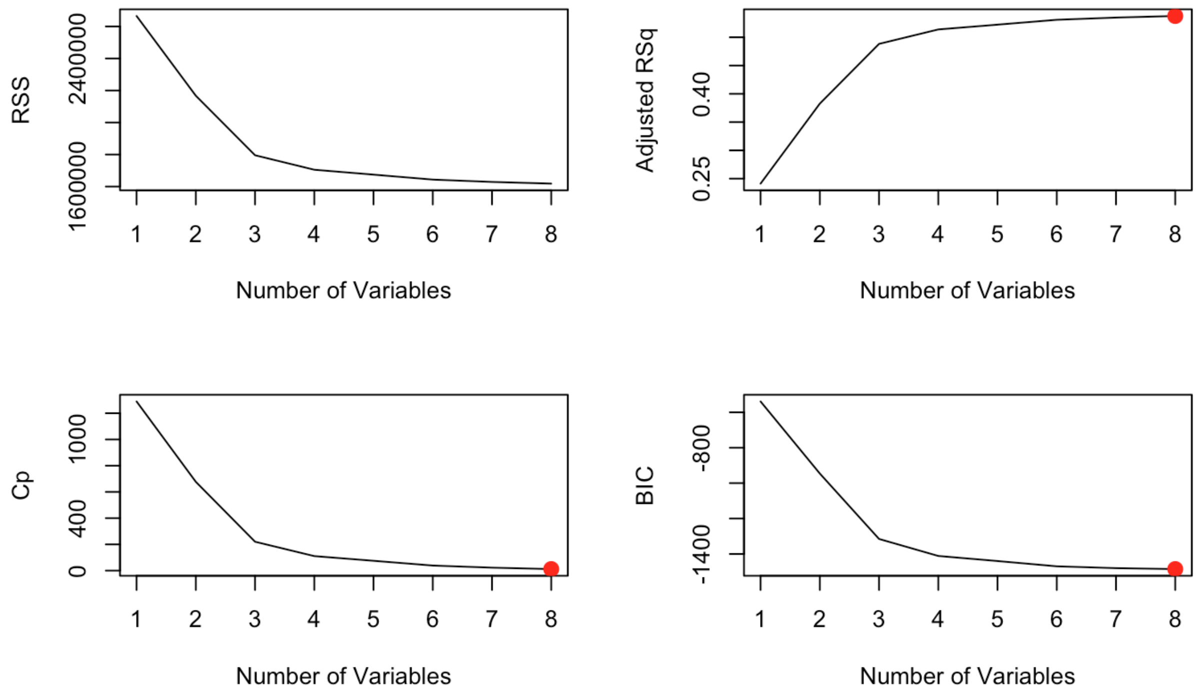


Figure A11: Ridge Regression

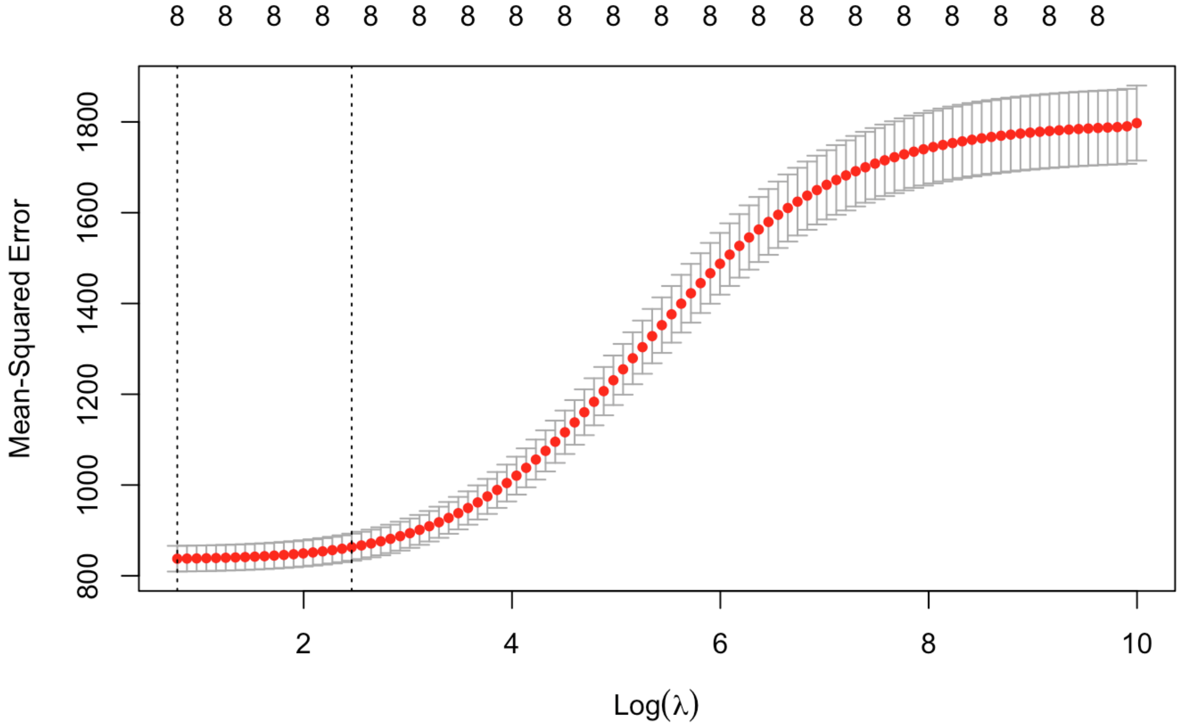


Figure A12: Lasso Regression

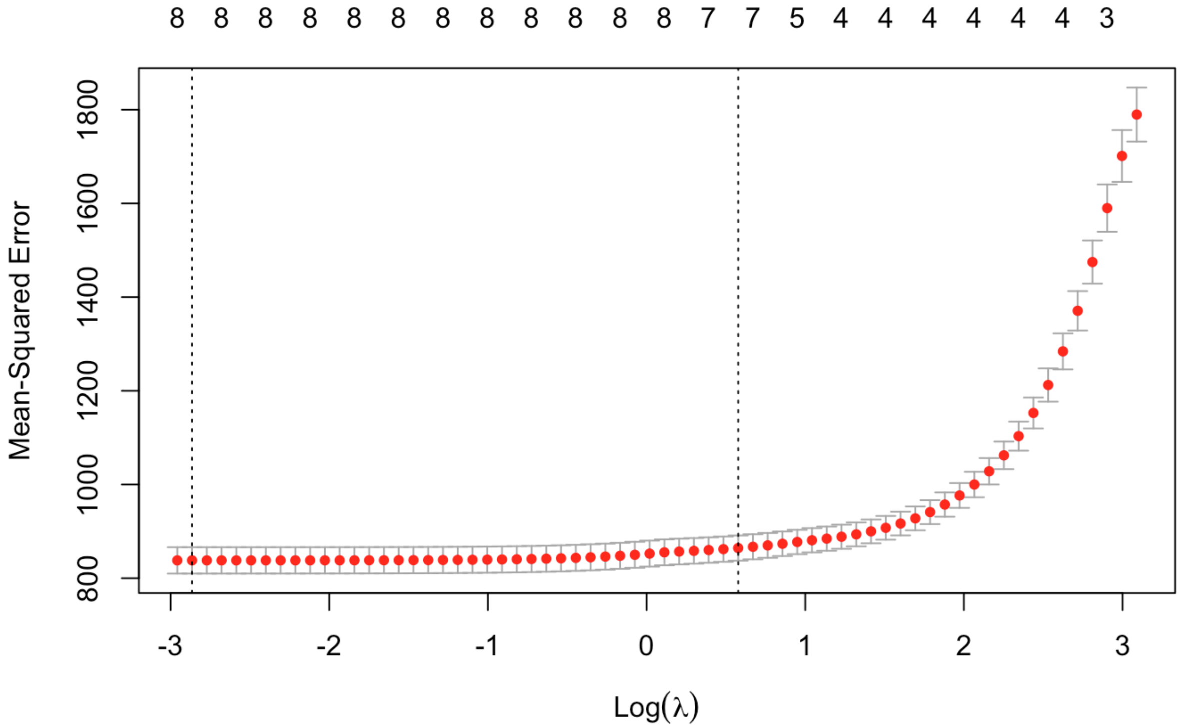
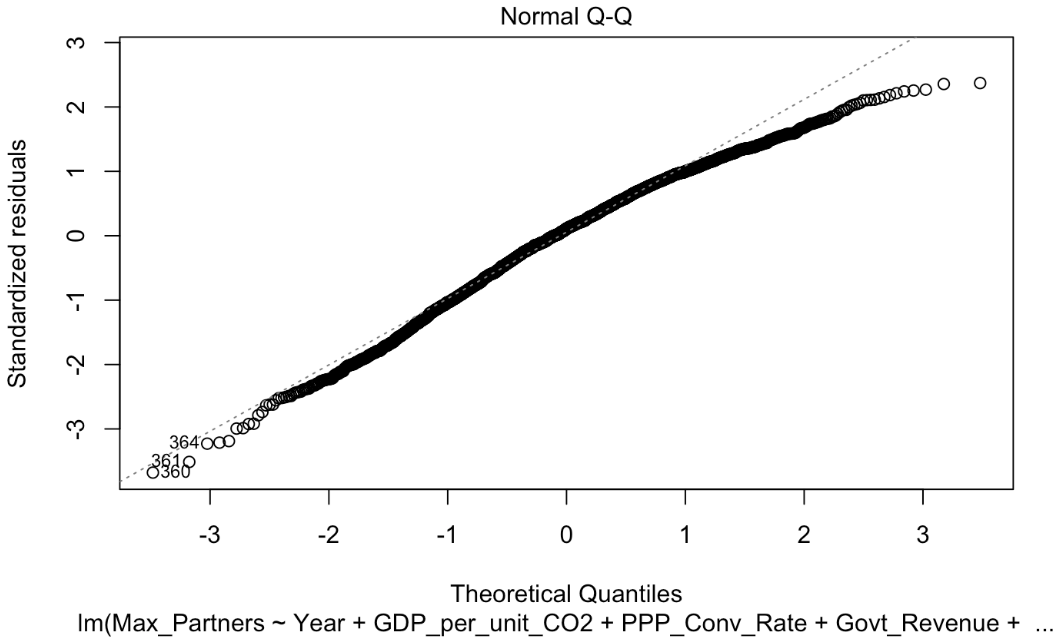


Figure A13: Linear Regression Q-Q Plot and Output



Linear Regression Output:

	Coefficient	**Std. Error**	**t value**	**Pr(> |t|)**
Intercept	-3.001e+03	1.734e+02	-17.306	< 2e-16 ***
Year	1.573e+00	8.658e-02	18.173	< 2e-16 ***
GDP_per_unit_CO2	-7.403e-01	2.065e-01	-3.585	0.000345 ***
PPP_Conv_Rate	1.350e-02	1.925e-03	7.012	3.21e-12 ***
Govt_Revenue	-5.815e-01	1.494e-01	-3.983	0.000102 ***
GDP_per_Cap	6.870e-04	4.276e-05	16.068	< 2e-16 ***
Intl_tourist_arrival	9.966e-07	5.301e-08	18.798	< 2e-16 ***
Agri_perc_GDP	-1.209	1.084e-01	-11.148	< 2e-16 ***
Rural_pop_perc	3.736e-01	5.368e-02	6.960	4.59e-12 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1 Multiple R-Squared: 0.5395, Adjusted R-Squared: 0.5376 F-Statistic: 292.1 on 8 and 1995 DF, p-value: < 2.2e-16				

Figure A14: Unpruned Tree Model

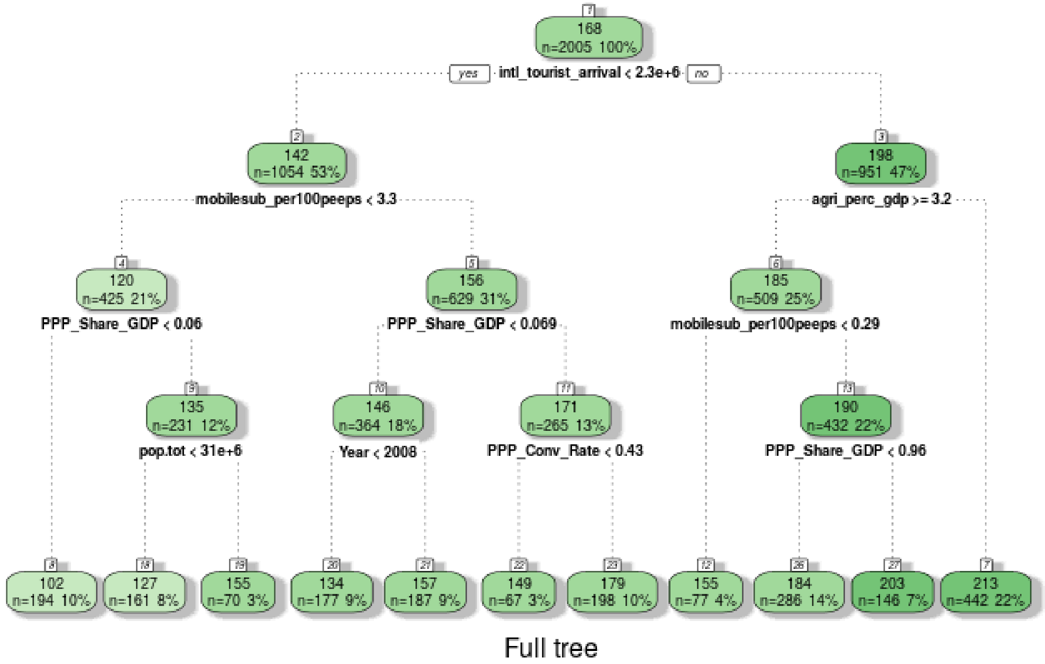


Figure A15: Pruned Tree Model

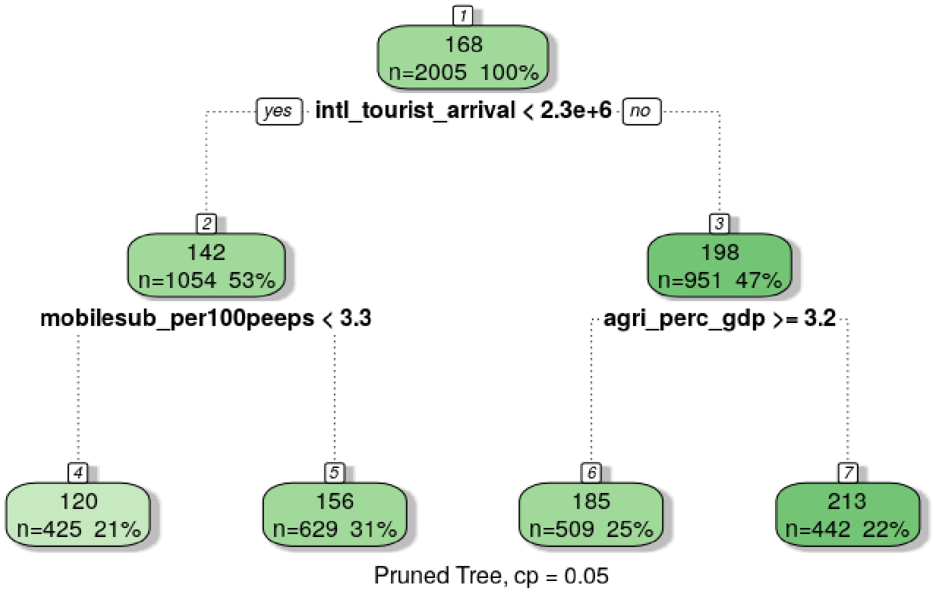


Figure A16: Random Forest Models - Test MSE and Number of Trees

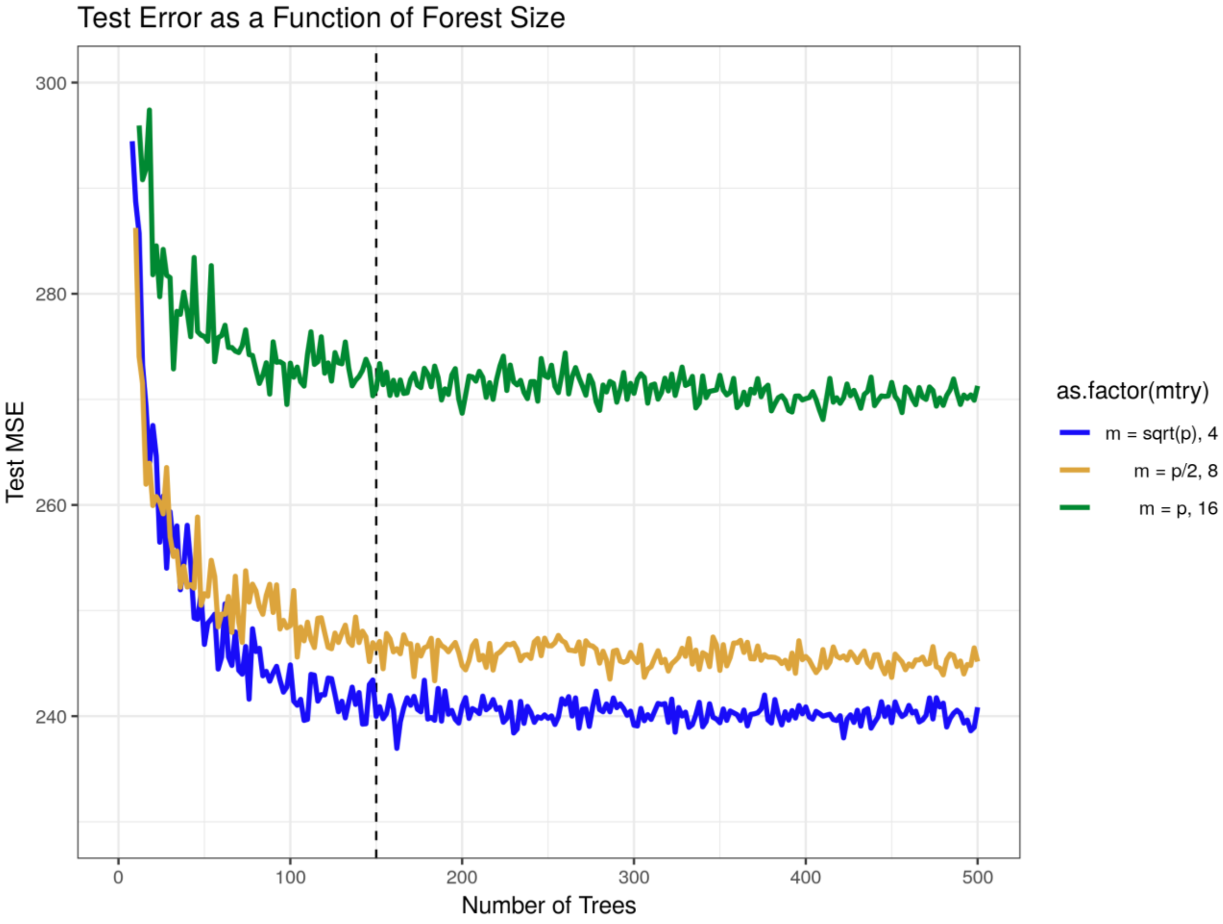


Figure A17: Comparison of Bagging and Random Forest Variable Tree Depth

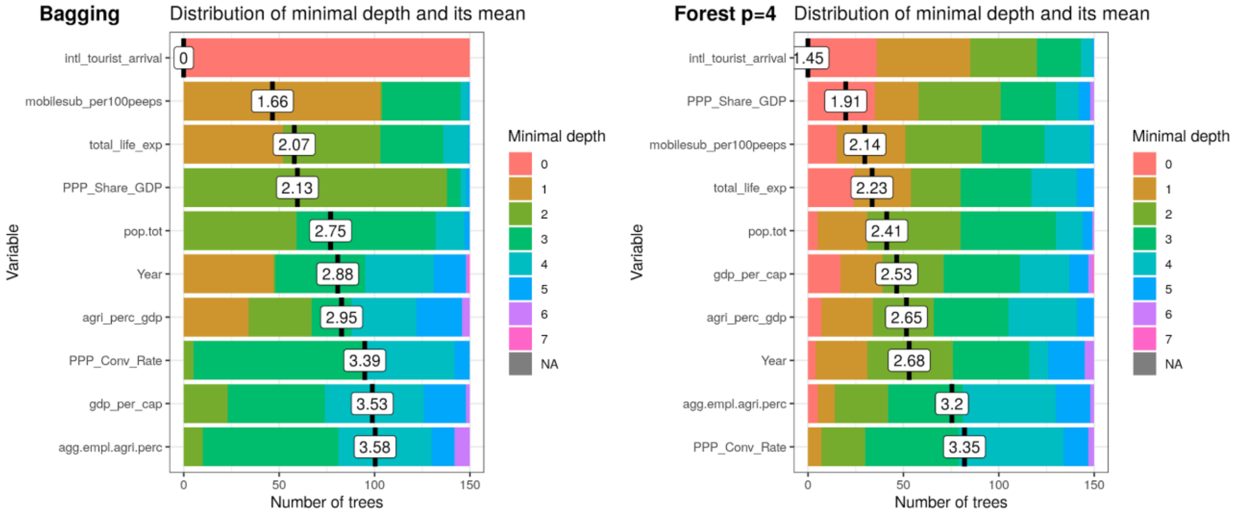


Figure A18: Comparison of Bagging and Random Forest Variable Importance

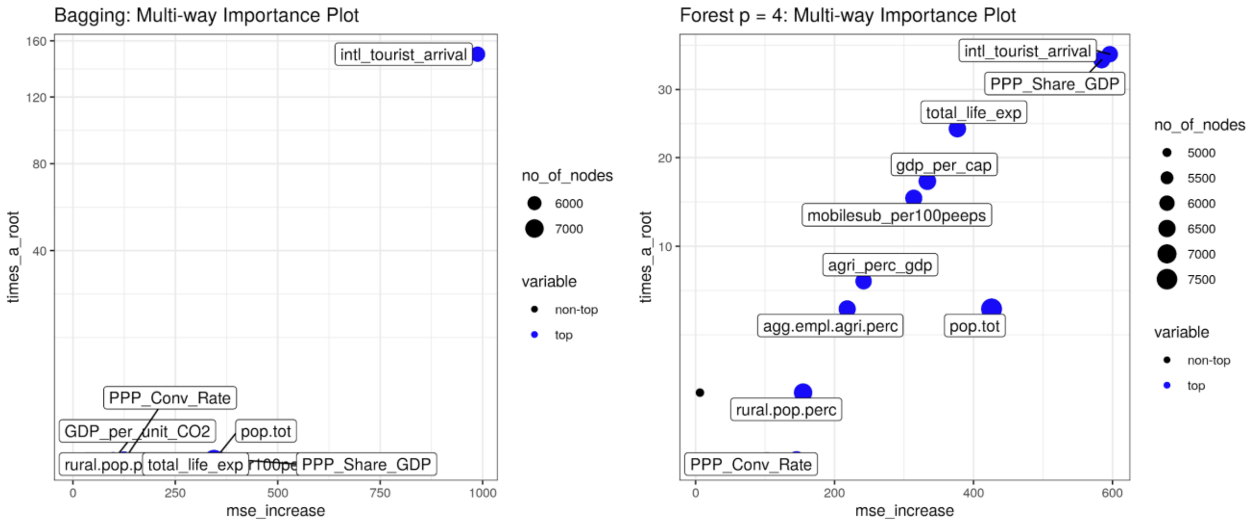


Figure A19: GBM Test MSE Comparison

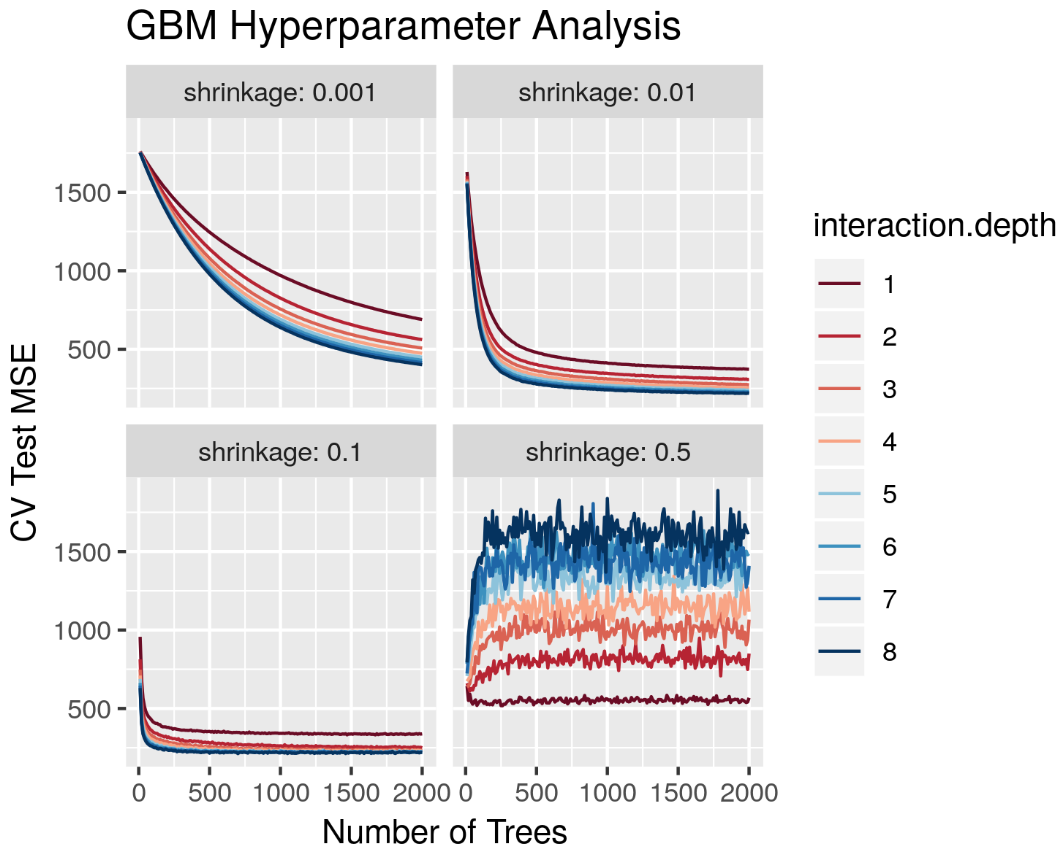


Figure A20: Best GBM Hyperparameter Close-Up

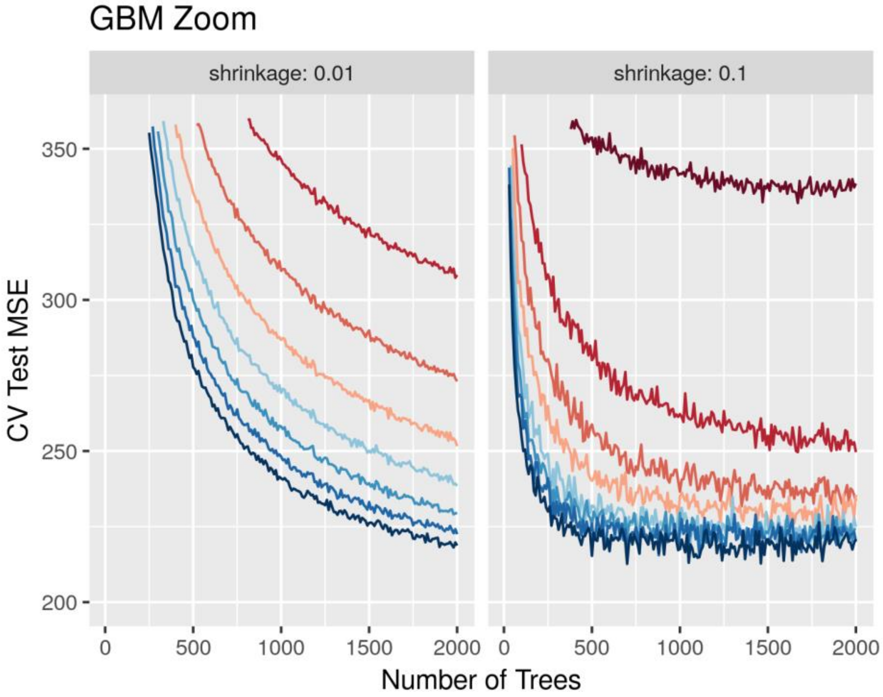
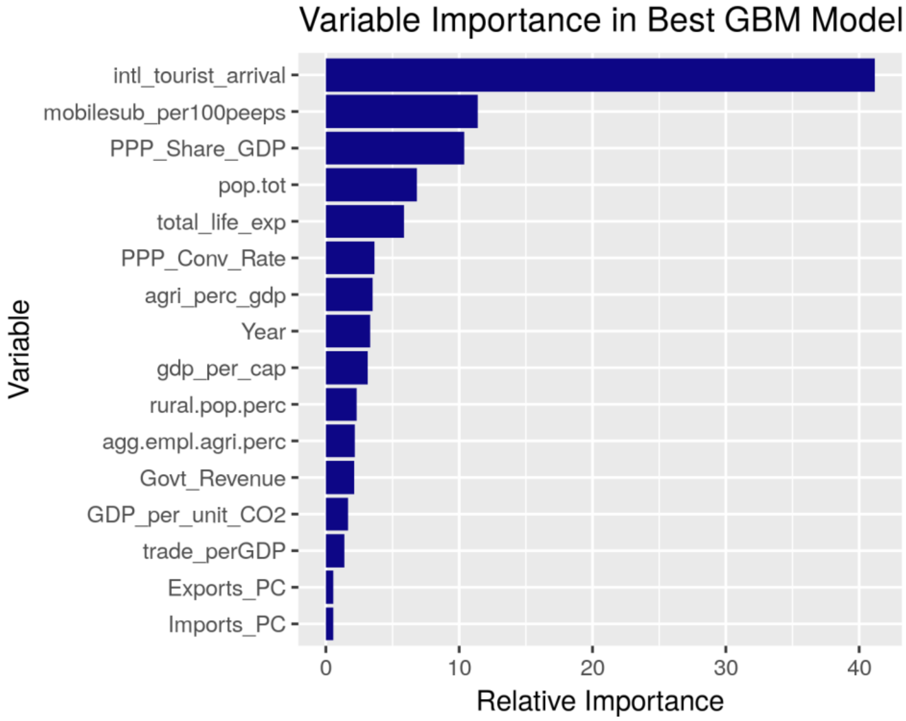


Figure A21: GBM Variable Importance Plot


Variable Importance in Best GBM Model
intl_tourist_arrival
mobilesub_per100peeps
PPP_Share_GDP
pop.tot
total_life_exp
PPP_Conv_Rate
agri_perc_gdp
Year
gdp_per_cap
rural.pop.perc
agg.empl.agri.perc
Govt_Revenue
GDP_per_unit_CO2
trade_perGDP
Exports_PC
Imports_PC
Variable
Relative Importance