How does corruption affect economic growth in developing countries? A machine learning approach [*]

Qian Zhang[†]

June 3, 2020

Abstract

This paper intends to uncover the impact of corruption on the economic growth among developing countries. To deal with the potential inconsistency of current corruption indicators, this paper uses Support Vector Machines (SVM) to construct a new index of corruption in a continuous series on the interval of [0, 1] for 140 developed and developing countries all around the world from 1996 to 2016. This new index mainly reflects the level of corruption control for each country, i.e. higher the index, better the control of its domestic corruption. With this newly-constructed corruption-control indicator, and by Dynamic Panel Data (DPD) model and two-step "difference" GMM estimations, this paper finds that there exists an optimal level of corruption control regarding economic growth for developing countries. Also, better regulator quality would enlarge the positive effect of corruption control on economic growth. For the prediction of economic growth, this paper compares multiple machine learning methods and finds that Random Forest model performs the best among all in terms of the out-of-sample Mean Squared Error (MSE).

keywords: Corruption control, economic growth, SVM, machine learning.

[*] I want to express my sincere gratitude to Dr. Richard Evans who provides me with continous support and helpful comments. I also want to thank all the classmates of MACS 30250 this Spring.

[†] MACSS program, University of Chicago, timqzhang@uchicago.edu.

1 Introduction

Corruption has long been existing in the human society, and it plays an important role in the economic growth. People keeps wondering how it affects growth, especially in developing countries, where corruption may occur with quite high probability. However, so far, both theories and empirical evidences have not reached a consensus on the impact of corruption in the developing countries, even the measurement of corruption itself is controversial. Therefore, this paper intends to provide some approaches to find out how corruption may affect economic growth in the developing countries.

The organization of the paper is as follows: Section 2 will provide a brief review of the theoretical and empirical evidences on the effect of corruption on growth, and discuss some potential contributions from this paper. Section 3 mainly introduces the methods for the reconstruction of corruption index via machine learning algorithms, and the estimation models to check the impact of corruption. Section 4 posts several hypotheses and unfolds the empirical results to test the hypotheses, and then analyzes the results with detailed discussion. Section 5 will further focus on the prediction of economic growth and select the best prediction model. Finally, Section 6 will summarize what this paper finds and makes the conclusion, as well as some policy implications. Also, limitations of this paper will be discussed in this section.

2 Literature review

2.1 Controversy over the effect of corruption

The impact of corruption on economic growth has long been discussed by academia, and there is a large and growing literature focusing it. However, answer to this topic is unclear theoretically, since scholars develop two contradicting theories and they haven't reached a consensus. One strand of the literature, represented by Leff (1964) and Leys (1965), claims that the corruptive actions like bribery are severe in countries mainly with poor quality of governance, where ill-functioning governments often promote sluggishness and officials do not have any incentives on work. Under such a scenario, corruption may provide a strong incentive on officials to fasten the decision-making process, enhance the overall efficiency and productivity, and thus improve economic performance. In other words, the control of corruption may hinder the growth pace. Such theory is also known as "grease the wheels" hypothesis, and it is echoed by Lui (1985), who argues that awards based on the scale of corruption could be provided and may attain Pareto-optimal allocation of the entire market.

The other side of the debate is called “sand the wheel” hypothesis, which argues that corruption never has positive effect on the economic growth, and advocates the strict control on corruption. Many voices are direct rebuttals to “grease the wheel” hypothesis. Myrdal (1972) claims that corruption only negatively influence the efficiency of governance, investment and then economic growth, as officials may often delay the working process to obtain time for corruptive activities. Kurer (1993) further argues that corruptive officials have incentive to make distortions in economy to sustain their illegal income. Therefore, corruption neither improves efficiency, nor it compensates for ill institutions, which makes “grease the wheel” hypothesis invalid. Moreover, Mauro (1995) points out that corruption will impede private investment, and then negatively influence economies.

Empirical studies reveal a more intense controversy with divergent results. Given that the scholars focus on distinct regions with different data source, multiple models and estimation methods, empirical papers result in even totally opposite conclusions.

Some papers show positive effect on corruption, thus supporting to loosen regulation on corruption. Biru (2010) finds that there exists a positive relationship between corruption and economic growth in Bangladesh using cross-sectional data and regression analysis. By utilizing the panel data of 69 countries (both developed and developing ones) from Penn World Table and two corruption indexes called Corruption Perception Index (CPI) and World Governance Indicator (WGI), Méon and Weill (2010) construct a Stochastic Frontier model estimated by MLE, and figures out that “grease the wheels” hypothesis is supported by empirical evidence. Also, according to Huang (2016), who assesses 13 Asia Pacific countries with panel data from 1997 to 2013, South Korea is one strong empirical evidence against the common perception that corruption is bad for economic growth.

However, more papers tend to conclude a negative relationship between growth and corruption, standing for a stricter corruption control. Swaleheen (2011) utilizes the data from World Development Indicator (WDI) and CPI to run GMM estimations, finding that corruption has a significant negative effect on the per capita real GDP. Farooq et al. (2013) check the case of Pakistan under a time series framework over the period of 1987 to 2009 and finds that the corruptive actions severely impede the GDP growth. Sharma and Mitra (2019) examine a sample of 103 countries around the world, further classify them by their income level (high, middle, low), and run the GMM estimation on a Dynamic Panel Data model with the data from International Country Risk Guide’s (ICRG) corruption index and WDI. They suggest that the corruption control within countries have positive effect on economic growth. Moreover, a Fixed Effect model is implemented by Frimpong et al. (2019) on countries of South African communities using data from WDI, CPI and polity IV index. They find that institutions should be proactive, otherwise anti-corruption policies would not exert positive influence on economies.

More conclusions are made other than simply positive or negative relations. Focusing on three particular developing countries, Wedema (1997) finds that corruption doesn't affect the economic growth pace. Glaeser and Saks (2006) find no significant effect of corruption on growth in the U.S. Treisman (2007) also finds no significant relations by modelling a cross-national panel data. More interestingly, Acemoglu and Verdier (1998) provide a general equilibrium approach. They find that it may be optimal to allow certain corruption to promote economies and reach an equilibrium for least developed countries. Ahmad et al. (2012) further echo this point by arguing that there exists a quadratic relation between corruption and growth in an inverse U shape way.

2.2 Potential contributions of this paper

Among all the literature, one thing to notice is that while there are some papers assessing the effect of corruption in developing countries, their focuses are too specific, usually on certain regions or even just one particular country. Also, their conclusions are inconsistent, with both positive (Biru, 2010) and negative (Frimpong et al., 2019) relations. Therefore, a general picture on the effect of corruption among all developing countries is still unclear and a further comparison of corruption's influence within the group of developing countries is still not disclosed.

With such motivations, this paper intends to include all the developing countries to figure out corruption - growth relations and make empirical inferences. The identification of developing countries will refer to the classifying criteria by United Nations (United Nations, 2019).

This paper also argues that the controversy in empirical evidences may mainly come from two reasons, namely specific model and estimation approaches, and the usage on different measurements of corruption indicator. This paper will contribute to tackle these two issues and make the result of this paper more justified and convincing.

The models used in the literature are various, including but not limited to basic regression model (Biru, 2010), Stochastic Frontier model (Méon and Weill, 2010), Dynamic Panel Data model (Sharma and Mitra, 2019), etc., with their corresponding estimation strategies of linear regression, MLE and GMM. This paper will mainly refer the Dynamic Panel Data model as it is the latest approach to focus on the relationship between corruption and growth.

While the data source for macroeconomic indicators is mainly WDI provided by World Bank, the sources of corruption index are different between empirical papers. For three most common used corruption indexes (Wraith and Simpkins, 2011), namely Transparency International's Corruption Perception Index (CPI), International Country Risk Guide's (ICRG) corruption index, and World Bank's World Governance Indicator (WGI), different papers tend to pick one or two of them as the proxy of corruption in the models as mentioned before. However, these indexes take distinct factors into account during the calculation. For example, CPI focuses more on the perception of corruption by integrating various perception indices (Transparency International, 2019), while WGI mainly evaluates the quality of governance (Kaufmann et al., 2010) by checking efficiency, regulator quality, etc. Such calculations based on different factors partially explain the different empirical conclusions. Also, the methodologies of some indices have once been changed, and the indices over time may not be comparable. Moreover, (Wraith and Simpkins, 2011) point out that the corruption itself is hard to measure, especially in developing countries.

To solve this problem, this paper will mainly reconstruct a corruption index indicating the corruption control, one indicator that reflects corruption level as suggested by Sharma and Mitra (2019). The measurement of corruption control will be done by a Support Vector Machine (SVM) method inspired by Gründler and Krieger (2016) and Lima and Delen (2020), and the new indicator will be a normalized sequence with continuous values ranging from 0 to 1. This paper will take six factors as input to SVM to generate the new corruption index. Using this new indicator, this paper will figure out a robust effect of corruption on growth in developing countries. Further discussion on the reconstruction of corruption index will be mentioned in Section 3.

3 Methods and Models

3.1 Reconstruction of corruption indicator

3.1.1 Theoretical framework of SVM

Inspired by Gründler and Krieger (2016), this paper will mainly consider the method of Support Vector Machines (SVM), in particular the Support Vector Regression to predict continuous values. While this paper mainly focuses on the group of developing countries, the reconstruction of corruption index will cover all countries available to make accuracy approximations. Due to the feature of data, the new corruption index will mainly reflect the control of corruption in each country.

The corruption index $c_{i,t} \in \mathcal{C} \subseteq \mathbb{R}$ for certain country i at period t can be expressed as a function $\mathscr{F} : \mathcal{X} \subseteq \mathbb{R}^m \to \mathcal{C} \subseteq \mathbb{R}$ of the extent to which the country-year pairs satisfy certain given conditions $\mathcal{X}$, where m denotes the number of conditions selected. Therefore, the corruption index could be expressed as

$$c_{i,t} = \mathscr{F}(x^1_{i,t}, \ldots, x^m_{i,t}), \forall(i,t).$$

This paper also wants to mention that because of some unobserved features and measurement errors, it is not feasible to provide a perfect fit, and the purpose of using Support Vector Regression is to compute a function that could greatly approximate the "true" function without losing essential information.

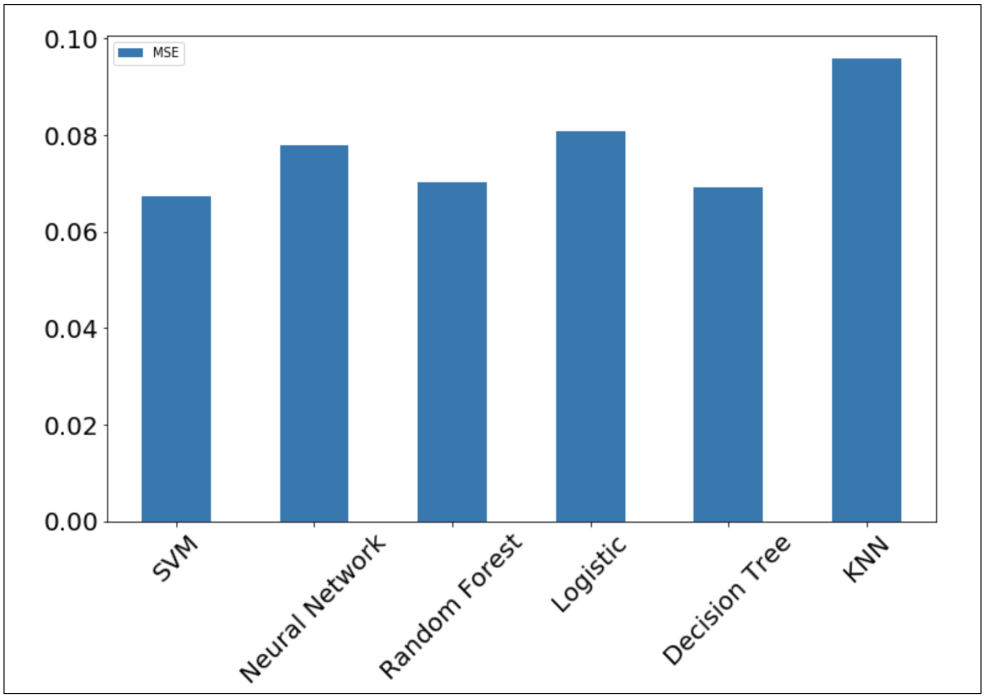
3.1.2 Algorithms of reconstruction

The logic of algorithm on reconstruction also mainly refers to Gründler and Krieger (2016). Firstly, some variables indicating features and characteristics need to be selected. From Graycar and Smith (2013), countries' political institutions and extent of law implementation should be considered in the measurement. Andersson and Heywood (2009) point out that culture patterns like the religious power in countries has strong relation with the perception of corruption. Also, socioeconomic conditions, internal conflicts and bureaucratic efficiency need to be included as factors suggested by Heywood and Rose (2013) to represent overall stability and governance in countries. Overall, this paper will take these six variables to form the feature space.

Secondly, this paper codes all the country-year pairs in the dataset that could be identified unambiguously as having good or bad corruption control. As there are several existing corruption indices, this paper mainly uses the corruption index from International Country Risk Guide (ICRG) in the model fitting and training as suggested by Treisman (2007) and Kaufmann et al. (2010), since the measurement methodologies of other indices have once been changed and are not suitable for cross-country panel data analysis. The corruption index from ICRG is coded continuously from 0 (most corruptive) to 6 (least corruptive). This paper takes advantage of this index setting, and considers that being most corruptive indicates the least strict corruption control, and being least corruptive indicates the most strict corruption control, as suggested by Sharma and Mitra (2019). As a result, without inversing the index, the ICRG index itself could imply the level of corruption control or the anti-corruption level for each country. The index is then divided in three parts. For country-year pairs with index larger than 3.6 (larger than 60% quantile), this paper marks them as having good control of corruption (code with 1), and those with index less than 2.4 (less than 40% quantile) are coded with 0 as having bad control of corruption. This paper thus keeps the country-year pairs with coding of 0 and 1 to form the sample set for later on modeling.

For the third step, this paper randomly splits this sample set to create the training set and uses it to fit the SVM model, approximating the function mentioned above. To further justify the SVM model, this paper controls all other conditions i.e. random seed, and runs some other machine learning methods. The comparison of MSE for each method is shown in Figure 1, and it shows that the usage of SVM is statistically convincing.

Figure 1: Comparison of MSE

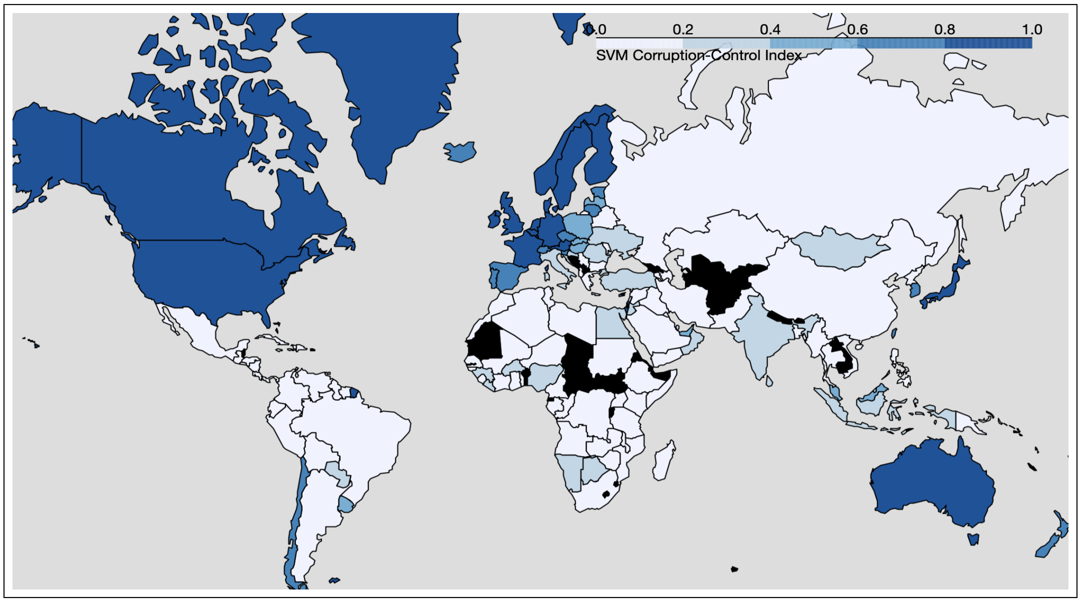


Lastly, this paper uses the estimated $\mathscr{F}$ to apply to all country-year pairs, and thus calculated the new corruption index $c_{i,t} \in [0, 1]$. Higher the index, better the corruption control for certain country. To have a robust calculation, the bootstrapping is used.

3.1.3 Overview of Corruption Control around the world.

The reconstruction of corruption index covers 140 countries from 1996 to 2016 due to the limitation of data available. Figure 2 visualizes an overview of the level of corruption control around the world in 2016. Deeper the color, stricter corruption control the country is, and countries marked black are the ones that the index does not cover. It shows that most countries in Europe, North America, and Australia have relatively good control on corruption, while many countries in Asia, Middle East regions, South America and Africa suffer from relatively bad corruption control. It could be also seen that one country's corruption-control level is probably related to its neighbor countries.

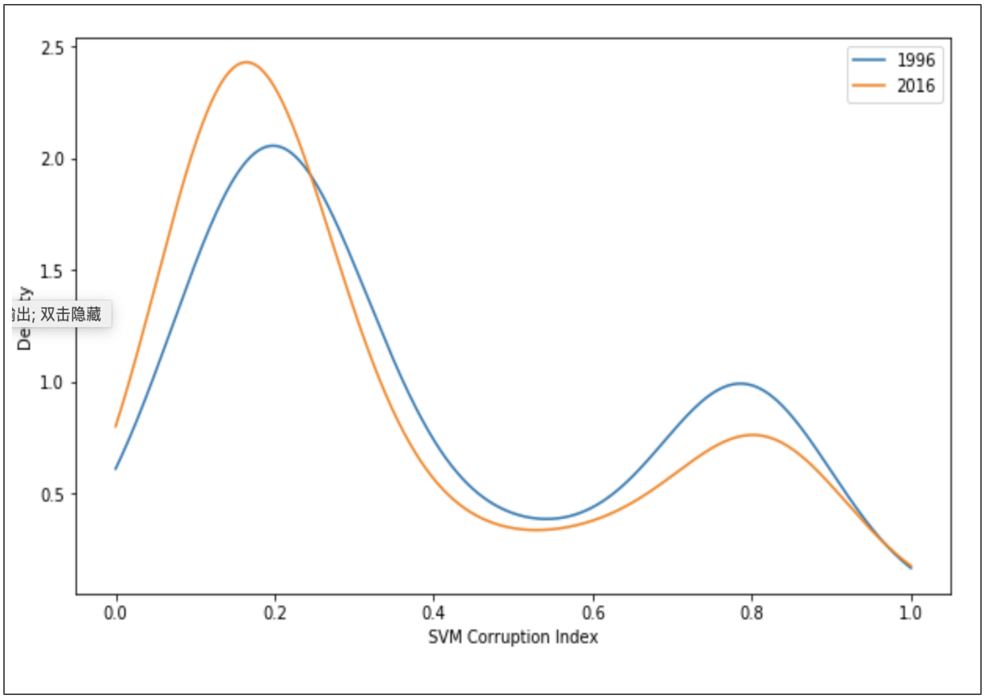
Figure 2: Corruption Control around the world, 2016



Another way to for visualization is shown in Figure 3, where the distribution of corruption index on corruption control in 1996 and 2016 are estimated by the Gaussian kernel. It shows a bimodal feature for both 1996 and 2016. The first peak is located at the low level of corruption control, and the second peak at the relatively high level of corruption control. From 1996 to 2016, it seems that many countries with good control on corruption in 1996 loosen the control on corruption over 20 years and become more corruptive in 2016. In the figure, the density of the second peak goes down while the density of the first peak goes up. Overall, this figure shows a worrying trend that over 20 years from 1996 to 2016, fewer countries keep high-level corruption control and enforce anti-corruption policies, and more countries become corruptive.

Overall, the check on the two visualizations justifies the reconstruction of corruption index by this paper.

Figure 3: Kernel estimates of Corruption Index



3.2 Model specification and data

3.2.1 Model specification

A Dynamic Panel Data (DPD) model is used to explore the effect of corruption on economic growth among developing countries as suggested by Sharma and Mitra (2019). The DPD model mainly includes the lagged dependent variable as the explanatory variable to specify the unobserved panel effects. The baseline model is shown as

$$y_{i,t} = \beta y_{i,t-1} + \gamma c_{i,t} + \theta \mathbf{X}_{i,t} + \mu_i + \xi_t + \epsilon_{i,t},$$

where $y_{i,t}$ denotes the log form of GDP per capita, $c_{i,t}$ denotes the corruption index, $\mathbf{X}$ denotes all other control variables, μ_i denotes country's fixed effect, and ξ_t denotes the time's fixed effect.

As for estimation method, Arellano and Bond (1991) suggest the "difference" GMM estimation for DPD model, in which the lagged dependent variable serves as the instrument and all variables are taken the first difference. After taking the first difference, the fixed effect μ_i will be wiped out. Blundell and Bond (1998) also propose a "system" GMM estimation, which aims to alleviate the problem of poor instruments by adding extra moment conditions. However, as mentioned by Roodman (2009), while both "difference" and "system" GMM have the estimation problem from instrument proliferation, such problem is particularly dangerous in “system” GMM, which could cause severe biases. Therefore, considering the data set and number of instruments that would be used, this paper will mainly consider the “difference” GMM for estimation.

3.2.2 Data and Variables

The data mainly comes from World Development Indicators (WDI) of World Bank, from 1996 to 2016 over 128 developing countries. The dependent variable is the GDP per capita and will be in the log form. The variable of corruption control is directly from the reconstructed corruption index. The index of Regulatory Quality is also included suggested by Dzhumashev (2014), which comes from World Governance Indicator by World Bank with continuous values, and it implies that higher the index, better the regulator quality of one country. This paper also includes some important control variables to capture the aspects of trade openness, government consumption, capital investment, natural resources, inflow of foreign direct investment (FDI), and inflation from WDI. The descriptive statistics is shown in Table 1.

Table 1: Summary of descriptive statistics

Variables	Obs	Mean	Std.Dev.	Min	Max
GDP PER CAPITA	2630	5448.30	8370.52	187.52	64864.71
CORRUPTION INDEX	1974	0.23	0.17	0.00	1.00
REGULATORY QUALITY	2646	-0.39	0.77	-2.63	2.26
INFLATION	2623	13.44	111.67	-36.57	4800.53
GOVERNMENT CONSUMPTION	2464	14.66	7.21	0.91	135.81
TRADE OPENNESS	2553	82.55	50.86	0.03	442.62
CAPITAL INVESTMENT	2464	22.51	8.23	-2.42	69.67
FDI INFLOW	2605	4.38	7.46	-37.16	161.82
NATURAL RESOURCE	2630	10.19	12.85	0.00	86.45

Note: GDP per capita is in constant 2010 US dollar. Inflation is in the percentage form. Government Consumption, Trade Openness, Capital Investment, FDI Inflow and Natural Resource are all in the form of percentage share of GDP. Particularly, Trade Openness is calculated by (imports + exports) / GDP, where imports and exports are the imports and exports of goods and services.

Table 2: Effect of corruption on growth, two-step ”difference” GMM

Dependent variable: LGDP	(1)	(2)	(3)	(4)
LAGGED LGDP	0.965***	0.967***	0.966***	0.964***
	(0.002)	(0.002)	(0.002)	(0.002)
CORRUPTION CONTROL(CC)	0.0865***	0.286***	0.246***	0.563***
	(0.007)	(0.027)	(0.031)	(0.072)
CC^2		-0.244***	-0.254***	-0.747***
		(0.037)	(0.039)	(0.102)
REG_QUALITY			0.0285***	0.000567
			(0.002)	(0.003)
REG_QUAL×CC				0.139***
				(0.015)
TRADE_OPEN	0.000134***	0.000125***	5.58e-05***	-5.45e-05***
	(1.13e-05)	(1.03e-05)	(1.36e-05)	(1.49e-05)
GOV_CONSUMPTION	0.00113***	0.00104***	0.000558***	0.000385
	(0.000193)	(0.000184)	(0.000185)	(0.000323)
INFLATION	-4.91e-05	-5.77e-05*	3.04e-06	-4.72e-06
	(3.73e-05)	(2.97e-05)	(3.54e-05)	(3.79e-05)
FDI_INFLOW	-0.000271***	-0.000212***	-0.000231***	-0.000244***
	(1.80e-05)	(2.47e-05)	(2.34e-05)	(1.78e-05)
CAPITAL_INVEST	0.00188***	0.00210***	0.00177***	0.00200***
	(0.000142)	(0.000126)	(0.000129)	(0.000125)
NATURAL_RESOURCE	0.000166***	9.70e-05	0.000605***	0.000567***
	(6.31e-05)	(5.91e-05)	(4.42e-05)	(8.78e-05)
Observations	1,738	1,738	1,738	1,738
Number of countries	94	94	94	94
Serial Correlation Test (p value)	0.720	0.643	0.551	0.501
Sargan Test (p value)	1.000	1.000	1.000	1.000

[1] Extra Control Variable is year dummies. See the full result table in A-1.

[2] LGDP is denoted as the log form of GDP per capita.

[3] p values in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

4 Empirical Result and discussion

4.1 Empirical result

This paper mainly uses two-step "difference" GMM estimation to get the robust estimates. Three hypotheses are going to be tested. (1) Stricter corruption control has a positive impact on the economic growth among developing countries. (2) Higher regulatory quality has a positive impact on the economic growth among developing countries. (3) Higher regulatory quality will enlarge the positive effect of corruption control on economic growth among developing countries. The empirical result is shown in Table 2. The first column shows the baseline model. The second column adds the squared corruption index to further check if there exists a quadratic relation between corruption control and economic growth, and the first two columns will test Hypothesis 1. The third column adds the regulatory quality to test Hypothesis 2. The fourth column further adds the interaction of regulatory quality and corruption control to test Hypothesis 3.

4.2 Discussion

Hypothesis 1: Stricter corruption control has a positive impact on growth.

This hypothesis echoes the "sand the wheel" hypothesis pointed out by Mauro (1995). From the baseline model, the coefficient of corruption control is significantly positive at 1% level. It verifies the hypothesis and suggests that with all other variables fixed, stricter corruption control will positively affect economic growth in developing countries. Moreover, an interesting result is found when adding the squared corruption control term. The coefficient of corruption control keeps positive and the coefficient of squared term is significantly negative. Such a relation keeps significant at column 4. It suggests that with all other variables equal, there exists a downward quadratic relation between corruption control and economic growth, and an optimal level of corruption control exists among developing countries. In other words, when countries make more efforts on anti-corruption, it will firstly improve economic growth, but when the corruption control goes over the optimal level, the over-control of corruption will gradually impede growth. Such result echoes the hypothesis and empirical evidence from Acemoglu and Verdier (1998). Also, the significantly positive coefficient of lagged LGDP shows that the past levels of corruption control have an impact on both current and future economic growth.

Hypothesis 2: Higher regulatory quality has a positive impact on growth.

Kinda et al. (2009) suggest that the regulatory environment and quality is crucial for economic performance in developing countries. From column 3, the coefficient of regulatory quality is highly positive at 1% level and in general justifies the hypothesis. However, some empirical evidence (Gani, 2011) shows a negative but insignificant impact of regulatory quality on growth. Therefore, further analysis on the impact of regulatory quality to economic growth could be done, but it is irrelevant to the major topic of this paper.

Hypothesis 3: Higher regulatory quality will enlarge the positive effect of corruption control on growth.

Aidt et al. (2008) develop a theory and provide the empirical evidence that the relationship between corruption and growth is related to the quality of regulation. The coefficients of corruption control, squared CC term and interaction term in column 4 are all significant on 1% level, and the overall marginal effect of corruption control on growth could be written as $(0.563 - 1.494 \times \mathrm{CC} + 0.139 \times \mathrm{REG\_QUAL})$. Such effect could be explained based on the level of corruption control. Given a country with mild corruption control, i.e. $0.563 - 1.494 \times CC > 0$ and $CC >> 0$, and all other variables fixed, if regulatory quality becomes better, i.e. one unit increase of regulatory quality index, the positive effect of corruption control will become larger, which verifies this hypothesis. Also, given a country with excessive corruption control, i.e. $0.563 - 1.494 \times CC < 0$, and all other variables fixed, if regulatory quality becomes higher, i.e. one unit increase of regulatory quality index, the negative effect of corruption control will become smaller or even transfer to the positive effect on growth. Overall, if the country enforces a proper level of corruption control (no excess control), keeping the current anti-corruption policies and enhancing regulatory quality will together give huge positive influence on growth.

5 More on economic growth: prediction

In this section, several statistical learning models will be implemented to predict the economic growth via the major variables used in the Section 4. The purpose of this section is to select the best models for the prediction on economic growth.

5.1 Prediction models

5.1.1 Ridge Regression

The main advantage of Ridge regression is that it alleviates the problem of multicollinearity by adding one more penalty in the cost function. Given m instances and n explanatory variables, and denote the coefficients to be estimated as w, the new cost function for Ridge regression could be written as

$$f(w) = \sum_{i=1}^{m}(y_i - \sum_{j=1}^{n} w_j x_{ij})^2 + \lambda \sum_{j=1}^{n} w_j^2.$$

By adding the extra constraint on w, Ridge regression improves the basic OLS regression. The penalty coefficient λ plays a role that when the coefficients take quite large values, the value of cost function will be penalized. Also, the Ridge regression will shrink the estimated coefficients.

5.1.2 Lasso Regression

The cost function of Lasso regression is similar to that of Ridge regression, except that the penalty part is changed from the squared form to the absolute form. It could be written as

$$f(w) = \sum_{i=1}^{m}(y_i - \sum_{j=1}^{n} w_j x_{ij})^2 + \lambda \sum_{j=1}^{n} |w_j| .$$

Lasso regression is another approach to improve the basic OLS regression, and it also helps the feature selection by forcing some coefficients equal to zero.

5.1.3 Support Vector Regression (SVR)

The SVR is one application of the overall Support Vector Machines. To apply SVM into regression, a margin of tolerance called epsilon is set in approximation to the SVM. The main idea for SVR is also quite similar to that of SVM, which is to look for the best hyperplane to minimize errors, keeping in mind that part of the error is tolerated. There is also a penalty parameter called C in SVR, which serves similar role to λ in Lasso and Ridge regression.

5.1.4 K Nearest Neighbor (KNN) Regression

The major difference between KNN regression and KNN classification is that the predicted value to a new data point by KNN regression is the mean value of its nearest neighbor points, other than the majority voting result of its nearest neighbor points for the discrete value prediction. To briefly summarize the core algorithm of KNN, it calculates the distance of each point, and find out k "nearest neighbor" point around each point. Then for the regression, the value of each point is assigned with the mean value of its neighbors. This paper mainly uses the Euclidean distance with the uniform weight.

5.1.5 Neural Network Regression

The Neural Network Regression, or the Multi-layer Perceptron Regression defines the input layer, hidden layers and the output layer to perform the deep learning process. The activation functions will be implemented to link different layers. For the regression, the activation function from the hidden layer to the output layer will be linear to fit continuous values. For a typical feed-forward neural network regression with just one hidden layer, the output could be written as

$$y = f_o[\alpha_o + \sum_{h=1}^{m} w_{ho} f_h(\alpha_h + \sum_{i=1}^{n} w_{ih} x_i)],$$

where n is the node number in the input layer, m denotes the number of nodes in the hidden layer, i denotes the input layer, h denotes the hidden layer, o denotes the output layer. f_h is the activation function from input to the hidden layer, and f_o is the activation function from hidden layer to output. For regression, f_o could be just linear that $f_o(x) = x$.

5.1.6 Decision Tree and Random Forest Regression

A single decision tree regression creates a model that predicts values by learning some decision rules from the data features. In the case of regression, the target value y is expected to have floating point values other than the integer values as in the Decision Tree classification. The key hyper-parameters for a decision tree are maximum depth of the tree (max_depth), minimum number of samples required to split an internal node (min_samples_split), minimum sample number for each leaf node (min_samples_leaf) and maximum features (max_features) used. Based on the decision tree model, the Random Forest regression is an ensemble model that takes multiple decision tree models instead of the single one to improve the accuracy. One more hyper-parameter will be considered, which is called n_estimators. It decides how many decision tree models will be included in the random forest.

5.2 Algorithms of prediction

For the six prediction models introduced above, this paper considers decision tree and random forest as two models, and adds OLS regression as the baseline model. Therefore, there are in total eight models to run prediction on the economic growth. The prediction has the following steps.

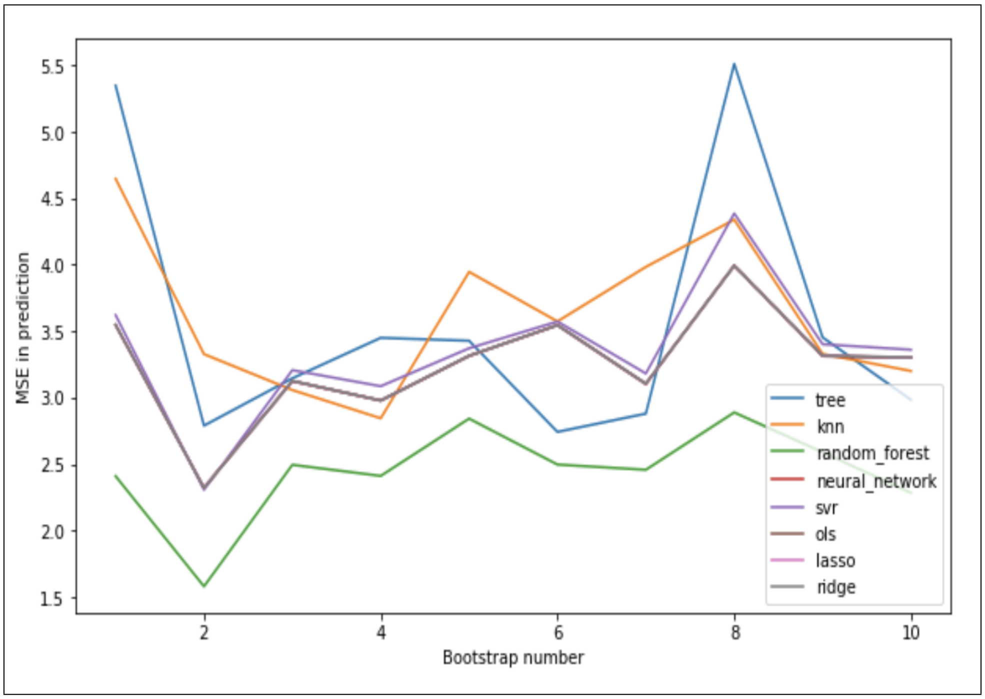
The first step sets the ranges of the key parameters to be tuned for each model in the randomized search cross validation (RSCV). The setting of RSCV keeps the same for each model to control the conditions, i.e. the number of processors used, the CV number, etc. Then the models are fitted for the training data sets and the parameters are tuned to get the best parameter settings by RSCV.

The second step uses the fitted model to predict the test data sets, and calculate the MSEs of predict values for each model.

The third step applies the bootstrapping trick to get robust comparison on the prediction accuracy of each model. To avoid heavy computing, this paper randomly selects 10 random-seeds to split the whole dataset into training and testing ones, and the 10 training sets will go through step 1 and 2 as mentioned above. The major criterion for comparison is the MSE in prediction.

5.3 Model comparison

Figure 4: Comparison of MSE in growth prediction



The plots of MSE of each model for 10 bootstraps are shown in Figure 4. The MSEs are rescaled to the range from 0 to 10. The figure shows that for the 10 randomly selected bootstraps, Random Forest performs the best in term of MSE on prediction. Therefore, this paper concludes that Random Forest could be considered as the best model to predict economic growth.

6 Conclusion

This paper mainly assesses the impact of corruption on economic growth among the developing countries, and having a reliable measurement of corruption level is necessary and significant to understand corruption itself and its impact on growth. Current corruption indices have discrete values and include distinct factors and methodologies, which drives scholars make controversial conclusions. Utilizing the machine learning approach of SVM, this paper reconstructs a more reliable corruption index that indicates the degree of corruption control in the form of continuous values for 140 countries from 1996 to 2016.

Based on the newly-constructed corruption index, this paper finds that there exists a downward quadratic relation between corruption control and economic growth within developing countries. In other words, an optimal level of corruption control exists in the developing countries to maximize the economic growth. Also, for countries with proper control of corruption which improves growth, higher regulatory quality in these countries will further enlarge the positive impact of corruption control on growth. These results imply that developing countries may find a good balance between corruption control and economic growth to reach the general equilibrium, and avoid the over-regulation and excessive control on corruption. However, realizing it may be quite difficult in reality. The results also indicate that for countries with loose control on corruption, when designing the anti-corruption policies, the government should also consider how to improve the regulatory quality in governance. Moreover, this paper concludes that Random Forest model is among the best prediction algorithms on economic growth, which may have some implications on growth prediction and relevant decision makings.

While this paper makes some contributions like the reconstruction of corruption index and selection of best machine learning models on economic growth, some limitations should also be considered. The first is that the difference GMM in estimation, although better than system GMM method, still generates the problem of Instrument Variable proliferation that would weaken the power of Sargan test (a test that checks the over-identification of IV) as the p values are quite high in the empirical result. Also, further work could be done to explore more approaches on the reconstruction of corruption index, since the measurement of corruption is still in controversy, and this paper just provides one approach to measure it accurately.

References

Acemoglu, Daron and Thierry Verdier, "Property Rights, Corruption and the Allocation of Talent: A General Equilibrium Approach," *The Economic Journal*, dec 1998, *108* (450), 1381–1403.

Ahmad, Eatzaz, Muhammad Aman Ullah, and Muhammad Irfanullah Arfeen, "Does corruption affect economic growth?," *Latin American Journal of Economics*, 2012, *49* (2), 277–305.

Aidt, Toke, Jayasri Dutta, and Vania Sena, "Governance regimes, corruption and growth: Theory and evidence," *Journal of Comparative Economics*, 2008, *36* (2), 195–220.

Andersson, Staffan and Paul M. Heywood, "The politics of perception: Use and abuse of transparency international's approach to measuring corruption," *Political Studies*, 2009, *57* (4), 746–767.

Arellano, Manuel and Stephen Bond, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 1991, *58* (2), 277–297.

Biru, Paksha Paul, "Does corruption foster growth in Bangladesh?," *International Journal of Development Issues*, jan 2010, *9* (3), 246–262.

Blundell, Richard and Stephen Bond, "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 1998, *87* (1), 115–143.

Dzhumashev, Ratbek, "Corruption and growth: The role of governance, public spending, and economic development," *Economic Modelling*, 2014, *37*, 202–215.

Farooq, Abdul, Muhammad Shahbaz, Mohamed Arouri, and Frédéric Teulon, "Does corruption impede economic growth in Pakistan?," *Economic Modelling*, 2013, *35*, 622–633.

Frimpong, Jennifer, Stepana Lazarova, and Samuel Asante Gyamerah, "Anti-corruption Instrument and Economic Growth: Evidence from SADC Member States," *Journal of Finance and Economics*, 2019, *7* (1), 14–22.

Gani, Azmat, "Governance and Growth in Developing Countries," *Journal of Economic Issues*, mar 2011, *45* (1), 19–40.

Glaeser, Edward L. and Raven E. Saks, "Corruption in America," *Journal of Public Economics*, 2006, *90*, 1053–1072.

Graycar, Adam and Russell G. Smith, "03_2011_Heinrich_Hodess_Measuring Corruption.pdf," in "Handbook of global research and practice in corruption" 2013, pp. 18–33.

Gründler, Klaus and Tommy Krieger, "Democracy and growth: Evidence from a machine learning indicator," *European Journal of Political Economy*, 2016, *45*, 85–107.

Heywood, Paul M. and Jonathan Rose, "Close but no Cigar: The measurement of corruption," *Journal of Public Policy*, 2013, *34* (3), 507–529.

Huang, Chiung Ju, "Is corruption bad for economic growth? Evidence from Asia-Pacific countries," *North American Journal of Economics and Finance*, 2016, *35*, 247–256.

Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi, "The worldwide governance indicators: Methodology and analytical issues," 2010.

Kinda, Tidiane, Patrick Plane, and Marie-Ange Véganzonès-Varoudakis, "Firms' productive performance and the investment climate in developing economies: an application to MENA manufacturing," *World Bank Policy Research Working Paper Series*, 2009, *4869* (March).

Kurer, Oskar, "Clientelism, Corruption, and the Allocation of Resources," *Public Choice*, 1993, *77* (2), 259–273.

Leff, Nathaniel H, "Economic Development Through Bureaucratic Corruption," *American Behavioral Scientist*, nov 1964, *8* (3), 8–14.

Leys, Colin, "What is The Problem About Corruption?," *The Journal of Modern African Studies*, 1965, *3* (2), 215–230.

Lima, Marcio Salles Melo and Dursun Delen, "Predicting and explaining corruption across countries: A machine learning approach," *Government Information Quarterly*, 2020, *37*, 101407.

Lui, Francis T, "An Equilibrium Queuing Model of Bribery," *Journal of Political Economy*, 1985, *93* (4), 760–781.

Mauro, Paolo, "Corruption and Growth," *The Quarterly Journal of Economics*, aug 1995, *110* (3), 681–712.

Méon, Pierre Guillaume and Laurent Weill, "Is Corruption an Efficient Grease?," *World Development*, 2010, *38* (3), 244–259.

Myrdal, Gunnar, "Corruption: its causes and effects," in "Asian Drama: an inquiry into the poverty of nations," The Penguin Press, 1972, pp. 166–174.

Roodman, David, "Practitioners' corner: A note on the theme of too many instruments," *Oxford Bulletin of Economics and Statistics*, 2009, *71* (1), 135–158.

Sharma, Chandan and Arup Mitra, "Corruption and Economic Growth: Some New Empirical Evidence from a Global Sample," *Journal of International Development*, 2019, *31*, 691–719.

Swaleheen, Mushfiq, "Economic growth with endogenous corruption: an empirical study," *Public Choice*, 2011, *146* (1), 23–41.

Transparency International, "Corruption Perception Index 2019," Technical Report, Transparency International 2019.

Treisman, Daniel, "What Have We Learned About the Causes of Corruption from Ten Years of Cross-National Empirical Research?," *Annual Review of Political Science*, 2007, *10*, 211–244.

United Nations, "World Economic Situation and Prospects 2019," Technical Report, UN Economic and Social Affairs 2019.

Wedema, Andrew, "Looters, Rent-Scrapers, and Dividend-Collectors: Corruption and Growth in Zaire, South Korea, and the Philippines," *The Journal of Developing Areas*, 1997, *31* (4), 457–478.

Wraith, Ronald and Edgar Simpkins, "Corruption in developing countries," 2011.

APPENDIX

A-1 Full empirical results of Table 2

Table 3 : Effect of corruption on growth, two-step GMM

Dependent variable: LGDP	(1)	(2)	(3)	(4)
LAGGED LGDP	0.965***	0.967***	0.966***	0.964***
	(0.002)	(0.002)	(0.002)	(0.002)
CORRUPTION CONTROL(CC)	0.0865***	0.286***	0.246***	0.563***
	(0.007)	(0.027)	(0.031)	(0.072)
CC^2		-0.244***	-0.254***	-0.747***
		(0.037)	(0.039)	(0.102)
REG_QUALITY			0.0285***	0.000567
			(0.002)	(0.003)
REG_QUAL×CC				0.139***
				(0.015)
TRADE_OPEN	0.000134***	0.000125***	5.58e-05***	-5.45e-05***
	(1.13e-05)	(1.03e-05)	(1.36e-05)	(1.49e-05)
GOV_CONSUMPTION	0.00113***	0.00104***	0.000558***	0.000385
	(0.000193)	(0.000184)	(0.000185)	(0.000323)
INFLATION	-4.91e-05	-5.77e-05*	3.04e-06	-4.72e-06
	(3.73e-05)	(2.97e-05)	(3.54e-05)	(3.79e-05)
FDI_INFLOW	-0.000271***	-0.000212***	-0.000231***	-0.000244***
	(1.80e-05)	(2.47e-05)	(2.34e-05)	(1.78e-05)
CAPITAL_INVEST	0.00188***	0.00210***	0.00177***	0.00200***
	(0.000142)	(0.000126)	(0.000129)	(0.000125)
NATURAL_RESOURCE	0.000166***	9.70e-05	0.000605***	0.000567***
	(6.31e-05)	(5.91e-05)	(4.42e-05)	(8.78e-05)
y2015	-0.00217**	-0.00126	-0.00152**	-0.000901
	(0.000900)	(0.000955)	(0.000666)	(0.000806)
y2014	0.00516***	0.00774***	0.00558***	0.00720***
	(0.00103)	(0.00118)	(0.00113)	(0.00104)

Continued on next page

Table 3 : Effect of corruption on growth, two-step GMM

Dependent variable: LGDP	(1)	(2)	(3)	(4)
y2013	0.0128***	0.0158***	0.0135***	0.0161***
	(0.00142)	(0.00160)	(0.00192)	(0.00182)
y2012	0.0131***	0.0167***	0.0138***	0.0179***
	(0.00142)	(0.00155)	(0.00174)	(0.00178)
y2011	0.0149***	0.0193***	0.0158***	0.0191***
	(0.00135)	(0.00145)	(0.00175)	(0.00166)
y2010	0.0200***	0.0240***	0.0216***	0.0258***
	(0.00190)	(0.00181)	(0.00169)	(0.00182)
y2009	-0.0212***	-0.0191***	-0.0199***	-0.0182***
	(0.00175)	(0.00188)	(0.00174)	(0.00147)
y2008	0.00945***	0.0123***	0.0106***	0.0124***
	(0.00155)	(0.00137)	(0.00213)	(0.00172)
y2007	0.0244***	0.0270***	0.0255***	0.0272***
	(0.00178)	(0.00153)	(0.00215)	(0.00199)
y2006	0.0249***	0.0281***	0.0256***	0.0272***
	(0.00172)	(0.00151)	(0.00217)	(0.00205)
y2005	0.0189***	0.0219***	0.0201***	0.0233***
	(0.00131)	(0.00144)	(0.00170)	(0.00176)
y2004	0.0223***	0.0256***	0.0248***	0.0262***
	(0.00171)	(0.00148)	(0.00224)	(0.00206)
y2003	0.00622***	0.00915***	0.00756***	0.00806***
	(0.00157)	(0.00152)	(0.00184)	(0.00198)
y2002	-0.00123	0.00135	-0.000102	-0.000170
	(0.00159)	(0.00146)	(0.00209)	(0.00209)
y2001	-0.00553***	-0.00258	-0.00346*	-0.00468**
	(0.00202)	(0.00190)	(0.00197)	(0.00188)
y2000	-0.000925	0.000597	-0.00160	0.000226
	(0.00223)	(0.00160)	(0.00170)	(0.00191)
y1999	-0.00782***	-0.00693***	-0.00914***	-0.00549***

Continued on next page

Table 3 : Effect of corruption on growth, two-step GMM

Dependent variable: LGDP	(1)	(2)	(3)	(4)
	(0.00199)	(0.00189)	(0.00146)	(0.00182)
y1998	-0.0169***	-0.0177***	-0.0194***	-0.0232***
	(0.00249)	(0.00241)	(0.00278)	(0.00300)
Constant	0.206***	0.160***	0.203***	0.183***
	(0.0122)	(0.0143)	(0.0167)	(0.0204)
Observations	1,738	1,738	1,738	1,738
Number of countries	94	94	94	94
Serial Correlation Test (p value)	0.720	0.643	0.551	0.501
Sargan Test (p value)	1.000	1.000	1.000	1.000