Final Project: Identifying Predictors of Political Corruption in Malaysia

Alif Iskandar Zabidi

10/4/2020

Contents

1 An Analysis of Political Corruption in Malaysia: Identification of Key Predictors of Corruption

1.1 Introduction

According to Transparency International, Malaysia has often occupied the position of the second least corrupt nation in Southeast Asia. But corruption and the public perception of such corruption has become a major political issue in Malaysia in recent years, a result of public outcry and the steady growth of opposition to Malaysia's long-running former ruling coalition, Barisan Nasional (or in English, the National Front). Barisan Nasional was the longest running ruling coalition in Malaysia's history, having administered the country since its independence in 1957. It began as a popular coalition made up of racially oriented political parties including UMNO (the United Malays National Organisation), the MIC (Malaysian Indian Congress) and the MCA (Malaysian Chinese Association) amongst others, and was for many decades relatively unchallenged in its dominance in Malaysia's parliament (Berman 2016, pg 145-147).

However, in the past decade, Malaysia faced a tumultuous period under the government of the former Prime Minister Najib Razak, with multiple scandals and allegations of graft and other forms of corruption being leveled against the Prime Minister and Barisan Nasional (Edwards 2018, pg 11-12). One of the largest of these scandals relates to the 1 Malaysia Development Fund, often abbreviated as 1MDB, in which a national development fund was spearheaded by Najib Razak and was used to siphon massive quantities of national funding into the personal wealth of politicians and other elites within the ruling coalition (Edwards 2018, pg 9-10).

Najib Razak had gained power in 2009, and ruled the country for nearly a decade, during which time increasing evidence of extensive government corruption came to light as a result of media coverage and investigations by both domestic and international monitoring agencies. Public outcry continued to grow to the point that it gave Barisan Nasional's political opponents sufficient momentum to mount a credible threat to BN's mandate. This came to fruition in May of 2018 during Malaysia's 14th General Election, when the country's major opposition coalition Pakatan Harapan (the Alliance of Hope), managed to overturn Barisan Nasional's long-running majority in parliament, and became the defacto ruling coalition of Malaysia.

It is of interest to note that Pakatan Harapan was spearheaded by former Prime Minister Tun Mahathir Mohamad, who was Malaysia's 4th and longest reigning democratically elected leader, during his term in office between 1981 and 2003. Prime Minister Mahathir, though largely popular in his time in office, was also no stranger to allegations of corruption (Edwards 2018, pg 9). The most recent change, which occurred on the 1st of March 2020 was essentially a political coup de grace, in which the Deputy Prime Minister Muhyiddin Yassin was able to secure enough support from political parties within and without the Pakatan Harapan coalition, and wrest sufficient parliamentary seats away from PM Tun Mahathir Mohamad. Due to how recent this shift in leadership was, it suffices to say that there is yet to be sufficient data for this to be included in the intended analysis.

Given the tumultuous political situation Malaysia now finds itself in, it is imperative that research should be dedicated to political corruption in Malaysia and how greatly it's effect has differed across the various governments that have ruled the country. It is the intention of this paper to identify key covariates and predictors of political corruption in Malaysia using machine learning methods. Using the "Varieties of Democracy" data-set, hosted by the V-Dem Institute in Gothenburg, Sweden, it is hoped that variable selection, decision trees and random forests can be designed to hone in on the key indexes in the VDem Data-set, to provide a suitable framework for future data analysis on political corruption. Variables that are identified will prove useful in the prediction of corruption trends not only within Malaysia, but with the additional objective of creating a generalisable approach for other nations.

The initial country based analysis will cover the time period between 1980 to 2020, as within this period the country was ruled by three individuals, PM Tun Mahathir Mohamad (who was in office as Prime Minister twice, as Malaysia's 4th and 7th national leader), PM Abdullah Badawi, PM Najib Razak and PM Muhyiddin Yassin. This time frame was chosen due to the upward trend of political corruption in Malaysia, and should allow the machine learning algorithms employed to identify important variables that can also be applied to other nations that have experienced growth in political corruption. After these predictors have been identified, a brief application will be applied to a neighbouring nation in Southeast Asia, the Philippines, with the expectation that the models will be able to achieve a similar (if somewhat) lower capacity to explain the variability of corruption in another nation.

This study is primarily an exploratory data analysis, to identify key indicators using an ensemble of machine learning techniques. These identified variables will then be compared to qualitative sources in the discussion section, to confirm their importance to studies on political corruption, or even to take note of unexpected variables that were chosen. Limited testing of the identified variables through a linear model will be applied to the Philippines in later sections, to explore the external validity of the approach. The groundwork of this approach will essentially lay the foundation for future modeling studies that will incorporate regional data, perhaps to identify trends in political corruption in all of the nations of Southeast Asia, and if the generalisability of the approach is sound at this point, could be used to develop a more complex model on political corruption that includes other regions.

1.2 Research Questions

- What covariates can be identified as key predictors of the changes witnessed in Malaysia's political corruption index in the past four decades?
- Do the identified covariates align with theories on corruption and good governance within the greater literature?
- Are these predictors also significant in modelling corruption when applied to a regional neighbour, the Philippines?"

1.3 Literature Review

Following a brief literature review, I have found that academic resources related to analyses of corruption in Malaysia remain largely limited to studies of institutional level corruption (and these studies are often related to specific public sector agencies), and with few sources conducting data based analyses on corruption. In a study by Kapeli & Mohamed for instance, they consider the public sector to be the largest concern in terms of corruption, and argued that low political will, ignorance of the causes of corruption, duplication of anti-corruption initiatives, and low public support for battling corruption as the primary reasons that institutional corruption has worsened in Malaysia in the past two decades (Kapeli and Mohamed 2019, pg 552-554).

In an earlier article by the same authors, Kapeli & Mohamed note that the continued worsening of corruption perceptions in Malaysia in the past decade were also a result of the continued failure of anti-corruption oversight institutions, such as the Malaysian Anti-Corruption Commission (MACC), which had been established in 1997 originally as the Anti-Corruption Agency (Kapeli and Mohamed 2015, pg 534). One of their primary arguments in the article is centred on the fact that although Malaysia has actually developed extensive and elaborate frameworks and strategies to tackle corruption (through agencies such as the MACC), these systems are obstructed by inherent defects in the country's overarching political systems, cultures and institutions (Kapeli and Mohamed 2015, pg 528)

This assessment is supported by Hashim, who argues that efforts to tackle institutional corruption in Malaysia have failed due to a development of a "double standard" in the value system of key public officials in Malaysia, where key national leaders "allow [and] tolerate pilfering and pillaging by public officials, to persist." (Hashim 2017, pg 560). This is essentially a result of top-down leadership in the country failing to address the development of an increasingly corrupt culture within government institutions, as they themselves develop intricate webs of public and private sector actors that conduct corrupt activities, making any efforts to detect and investigate such activities difficult and tedious (Hashim 2017, pg 559).

Unfortunately, the recent 2018 election in which Barisan Nasional was removed from power has not generated a large body of academic study on the subject, though it is likely we will see more research on the matter in the coming years. However, in an article by Ben Bland from 2019, he succinctly describes the ousting of former PM Najib Razak's government at the hands of Pakatan Harapan and Tun Dr. Mahathir Mohamad:

"The coalition rode a wave of anger against the perceived corruption of the Najib government. It represents one of the most dramatic, bloodless rejections of authoritarian rule in recent global politics, with the United Malays National Organisation — in power for more than 60 years — ousted at the ballot box. Even more remarkable is the fact that this reformist revolution was led by former strongman Mahathir Mohamad, prime minister from 1981 to 2003, in alliance with his protégé-turned-nemesis Anwar Ibrahim." (Bland 2018)

Bland goes on to note that while repression, electoral manipulation and allegations of corruption reached their zenith under PM Najib Razak, that critics of the newly elected government remain wary, and have concerns that a government under Tun Dr. Mahathir may return to a state of reduced transparency and a continuation of pervasive public sector corruption

(Bland 2018, pg 2). It is apparent that while Mahathir has appointed some leading civil society advocates and technocrats to economic and institutional reform committees, Bland notes that certain appointments have been regarded with suspicion, such as Daim Zainuddin as a key economic adviser given that many Malaysians doubt that the former finance minister (Zainuddin) is likely to be a genuine reformer (Bland 2018, pg 2).

In the larger context of corruption literature in the field of political science, there are many predictors of corruption that have been identified by various sources. In an analysis of four member states of the Association of Southeast Asian Nations (ASEAN), Sari et al (Sari, Cahaya, and Joseph 2020, pg 11-13) highlights the differing capacities for corruption disclosure practices in Indonesia, Thailand, the Philippines and Vietnam as key predictors for corruption levels domestically. They also note that structural organisation of domestic anti-corruption agencies has a major role to play in allowing an effective framework to fight political and other forms of corruption, and hence is a key indicator of corruption levels. Unfortunately, a measure with this level of specificity does not exist in the VDem Data-set, but can be accounted for by one of the measures for rule of law, such as executive and judicial accountability.

On the issue of rule of law, and its many principles and components being a key part of reducing political corruption levels, it has been argued that in the case of Malaysia, there exists extensive legal frameworks for monitoring and combating corruption, but enforcement is completely lacking (Siddiquee 2005, pg 126). This issue of enforcement is a key part of determining the true extent that rule of law is being upheld, as any country can have extensive anti-corruption mechanisms, but completely falls short on implementation and enforcement.

In a multiple regression analysis by Shim & Eom, rule of law was found to be a significant predictor for corruption in three of four models they had developed, and especially in the sense of law enforcement on corruption (for which they used an overarching rule of law index as a proxy) (Shim and Eom 2008, pg 310). They also argued that, aside from law enforcement, two key components that were often cited as key anti-corruption measures were establishing professionalism amongst civil servants, and enhancing bureaucratic quality overall, which they found to be significant as predictors of corruption in their models as well. In light of these theories, it is expected that a number of indicators covering transparency, accountability, law enforcement and other key parts of good governance will be identified by later variable selection methods.

1.4 Data Source and Methodology

The data is sourced from the V-Dem Institute (Coppedge, Gerring, Knutsen, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Luhrmann, Marquardt, Paxton, et al. 2020, 2020), which is a private research institute based in Gothenburg, Sweden. The data-set that the institute maintains is the "Varieties of Democracy" data-set, which features policy data, measures of democratisation and other key covariates that relate to the political structure and institutions of the world's nations, with data reaching back over a century. As they state, the Varieties of Democracy data-set is a "multidimensional and dis-aggregated data-set that reflects the complexity of the concept of democracy as a system of rule that goes beyond the simple presence of elections. The V-Dem project distinguishes between five high-level principles of democracy: electoral, liberal, participatory, deliberative, and egalitarian, and collects data to measure these principles." (Coppedge, Gerring, Knutsen, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Luhrmann, Marquardt, Paxton, et al. 2020, 2020)

The V-Dem data-set is a truly extensive data-set, featuring 4108 variables, with over 27,000 observations when taking into account the historical data of all observed nations. In the case of Malaysia, data goes back as far 1900, when the nation was officially administered as the Federated Malay States, a protectorate of Great Britain. The recorded data for Malaysia in the period of interest (between 1980 to 2020) is also extensive and appears to be frequently updated between years for variables of interest. While there is some missing data, I found that many these gaps were limited to variables that were unrelated to the question of political corruption, and will be eschewed for the purposes of this study.

The methodological strategy of this study is to first clean and filter the data as much as possible, by which I will aim to reduce the risk of having variables with high collinearity being fed into the variable selection models. This is certainly a risk given that many of the 4000 over variables in VDem are re-coded and re-scaled versions of primary versions of the data. Furthermore, a filter will be applied to remove variables that lack high levels of differentiation, that is, where observations have the same recorded value for intervals longer than three years. This is a necessary step to remove variables in which data collection by the VDem Institute and their associates is more limited, for instance when the last recorded is given for a period of 5 years due to infrequent surveying and data collection.

Although some predictors of interest may be lost due to this procedure, it also serves the purpose of removing variables with single factor labels that would interfere with variable selection algorithms, while also keeping only variables with higher levels of differentiation in the data allowing the algorithms to learn and fit trends more accurately.

Variable selection will be conducted first, which will take the remaining three hundred variables (after filters and data cleaning), and locate the six most important. Variable selection will be accompanied with decision trees to minimise the number of variables further, though in balance with maintaining a high amount of the variation in the outcome variable explained by these predictors. Following this, a linear model will be fitted to the data using the selected variables, to confirm the statistical significance of the selection.

This two step process is useful in that it applies two machine learning algorithms to narrow down variables to the smallest number needed in explaining a large proportion of variance in the outcome variable, but it can be somewhat cumbersome to apply two processes to a set of data. There is also the issue that decision trees, specifically regression trees, can be relatively expensive and time consuming computationally, though this is somewhat negated by the earlier variable selection. A definite weakness however is the higher risk of overfitting data, due to decision trees not employing regularisation and instead relying on tree pruning. This may lead to lower testing accuracy when the chosen variables are applied to a different context.

The next approach will involve Random Forest algorithms, which will automatically choose variables by their importance in explaining the highest proportion of variation in the outcome, and can do so with the breadth of all the variables included. The most important of these variables will then be used to fit a linear model, to identify the statistical significance of chosen variables. This random forest based method as compared to the previous two step method (from now on referred to as the tree based method) utilises the inbuilt ranking of variables based on importance in explaining variation, and is more straightforward to apply.

The primary advantage of random forests is that they essentially build a "strong learner" from a group of "weak learners", with these individual learners of course being separate decision trees. Random forests are quick to conduct and are often produce more reliable results than individual trees as a result of in-built random sampling, however, there also have the inherent weakness of overfitting especially in data-sets that are particularly noisy.

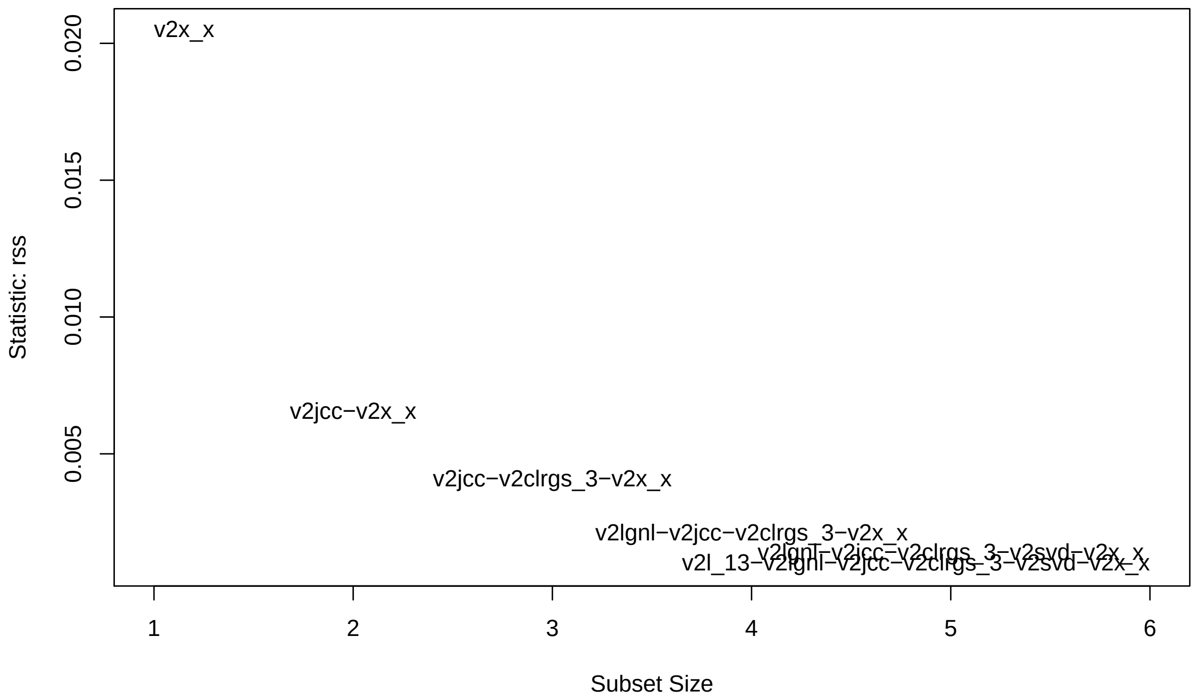
As a final step, both sets of predictors will be used to conduct a brief test of generalisability on a regional neighbour of Malaysia, the Philippines. The Philippines has been chosen not on the basis of having similar development levels, which the two countries do not have in common, but because of the widespread allegations of political corruption that exists in the country. Sari et al compared the levels of corruption disclosure practices (essentially public disclosure and reporting on incidents of corruption) in four nations within ASEAN, and found the Philippines to have the most pervasive level of corruption in its private sector, while also having the weakest anti-corruption agencies of the four. For this reason, the Philippines should prove to be an effective comparative subset of the data in terms of high levels of corruption.

1.5 Variable Selection and Decision Tree Approach

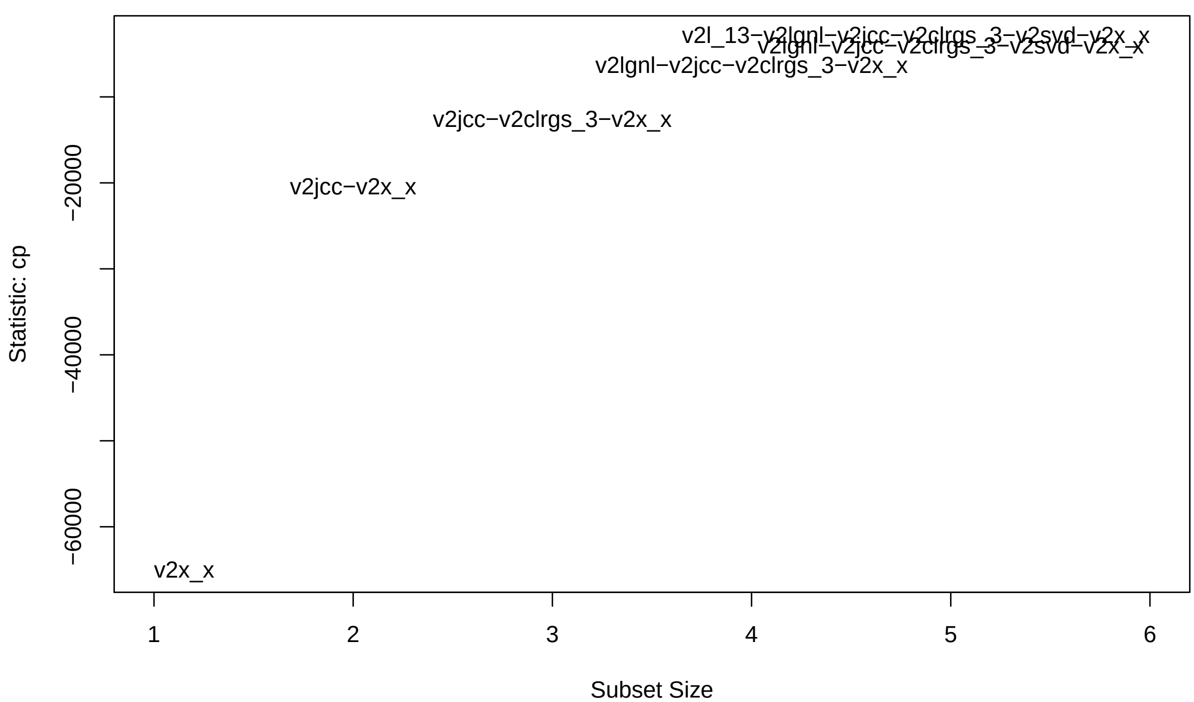
Using the subsets function, the earlier regsubsets variable selection that narrowed down the number of variables from hundreds to just six can be easily translated to plots that also show the Mallows Cp, Bayesian Information Criteria, Adjusted R-squared and Residual Sum of Squares of the chosen subset. These six variables can then be input into a decision tree, so that the algorithm can determine the smallest number of regressors that can achieve the highest accuracy in predicting variability in political corruption. This is at the heart of why a combination of variable selection and decision trees were chosen as a joint method, as I can easily identify a small set of predictors that could prove worthwhile for predictive models, down from a list of over three hundred indicators.

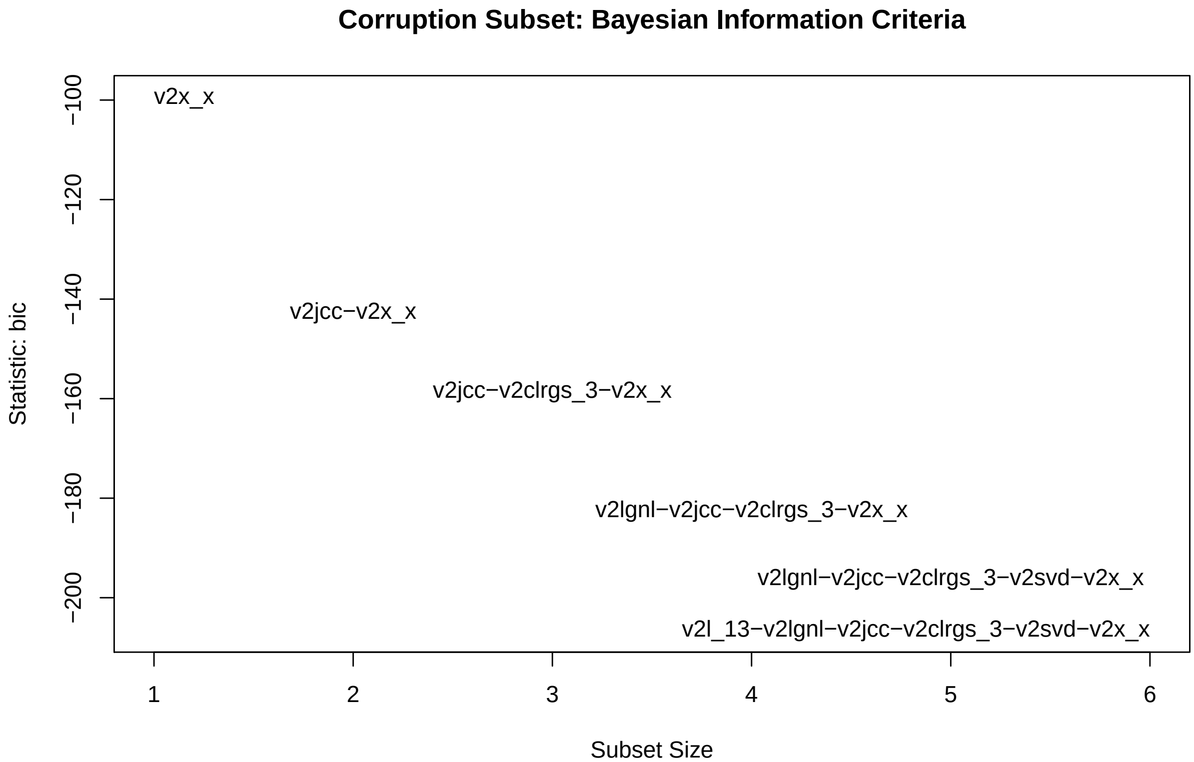
```
## Reordering variables and trying again:
```

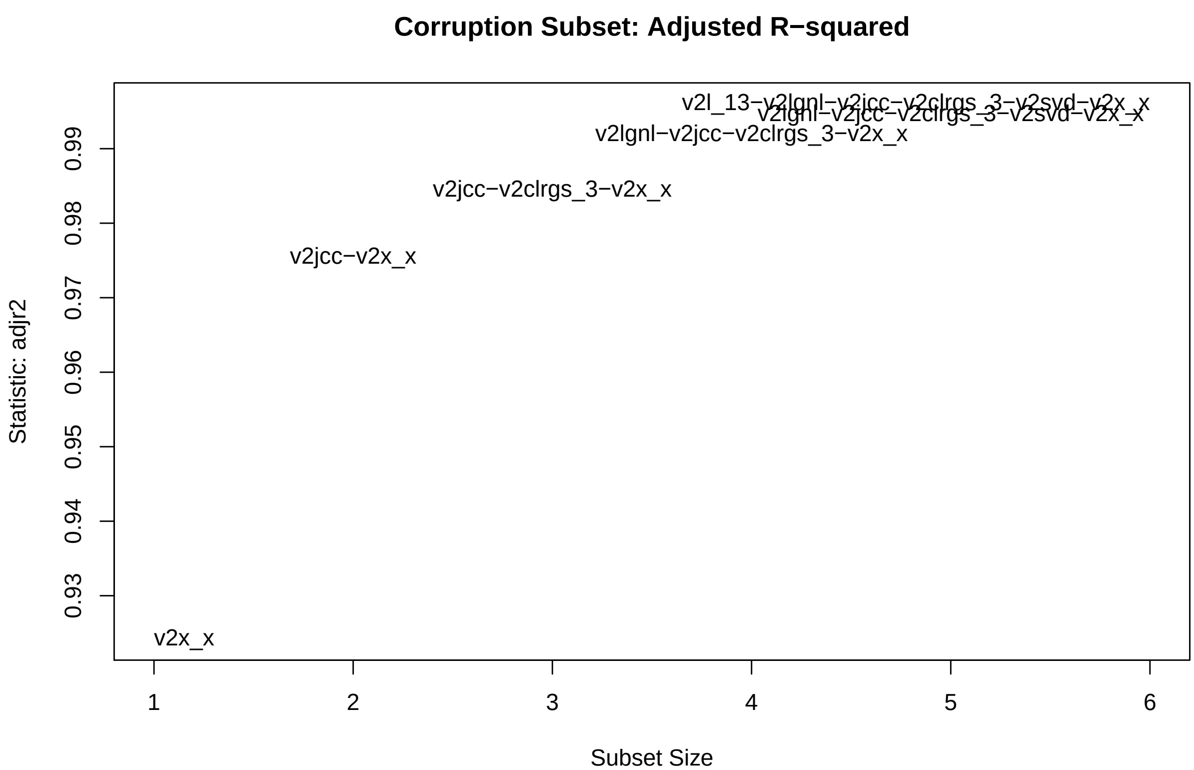
Corruption Subset: Residual Sum of Squares



Corruption Subset: Mallow's Cp



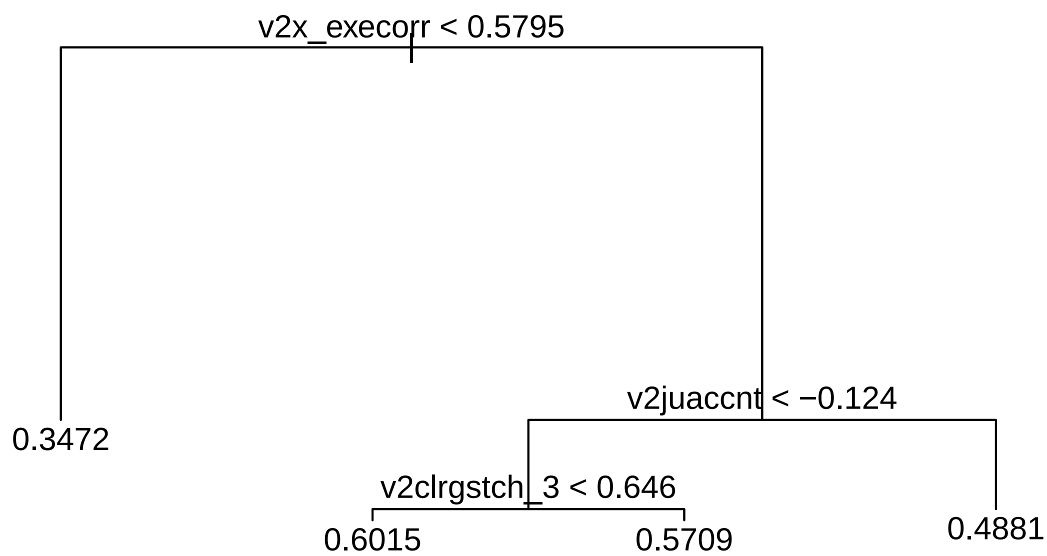

Corruption Subset: Bayesian Information Criteria
Statistic: bic
−100
−120
−140
−160
−180
−200
v2x_x
v2jcc−v2x_x
v2jcc−v2clrgs_3−v2x_x
v2lgnl−v2jcc−v2clrgs_3−v2x_x
v2lgnl−v2jcc−v2clrgs_3−v2svd−v2x_x
v2l_13−v2lgnl−v2jcc−v2clrgs_3−v2svd−v2x_x
1
2
3
4
5
6
Subset Size


Corruption Subset: Adjusted R−squared
Statistic: adjr2
0.99
0.98
0.97
0.96
0.95
0.94
0.93
v2l_13−v2lgnl−v2jcc−v2clrgs_3−v2svd−v2x_x
v2lgnl−v2jcc−v2clrgs_3−v2svd−v2x_x
v2lgnl−v2jcc−v2clrgs_3−v2x_x
v2jcc−v2clrgs_3−v2x_x
v2jcc−v2x_x
v2x_x
1
2
3
4
5
6
Subset Size

I then concluded that the optimal subset size as determined by the previously mentioned criteria is six variables, and includes the following indicators (names were initially abbreviated by the function to ease computation):

1. v2l_13 = Sub-national election area less free and fair characteristics (C) (v2elsnlfc), Areas that are remote (difficult to reach by available transportation, for example). (0=No, 1=Yes)
2. v2lgnl = Percentage of indirectly elected legislators upper chamber (A) (v2lginelup)
3. v2jcc = Judicial accountability (C) (v2juaccnt)
4. v2clrgs_3 = Stronger civil liberties characteristics (C) (v2clrgstch) = 3: Areas that are more economically developed. (0=No, 1=Yes)
5. v2svd = Domestic autonomy (C) (v2svdomaut)
6. v2x_x = Executive corruption index (D) (v2x_execorr)

```
##
## Regression tree:
## tree(formula = v2x_corr ~ v2elsnlfc_13 + v2lginelup + v2juaccnt +
##     v2clrgstch_3 + v2svdomaut + v2x_execorr, data = clean_data)
## Variables actually used in tree construction:
## [1] "v2x_execorr"  "v2juaccnt"    "v2clrgstch_3"
## Number of terminal nodes:  4
## Residual mean deviance:  0.0004593 = 0.017 / 37
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.0702000 -0.0011430 -0.0008947  0.0000000  0.0025000  0.0588000
```



From the above decision tree summary and plot, I was given the three most important predictors from this method, that will be taken and applied to a linear model on corruption. It should be noted that the three variables identified are the executive corruption index (v2x_execorr), judicial accountability (v2juaccnt) and category 3 responses of the stronger civil liberties characteristics index (v2clrgstch). Only these three predictors will be needed, as the express purpose of utilising a decision tree was to minimise the number of predictors needed in explaining the largest proportion of variance.

We can also note the specific thresholds given by terminal node splits in the tree, with the highest threshold in political corruption resulting from executive corruption greater than 0.579, judicial accountability lower than -0.124 and civil liberties less than 0.646 resulting in the data points with the highest corruption. This is a key strength of the approach, as not only are we given the three most important variables in the smaller subset, but also the exact hierarchy of importance in the determination of predicted corruption levels. Random forests employ trees in much the same way, but due to the complexity of the models, can be much harder to communicate and interpret.

```
##  v2x_execorr    v2juaccnt v2clrgstch_3
##     2.527917     2.588810     1.412769
```

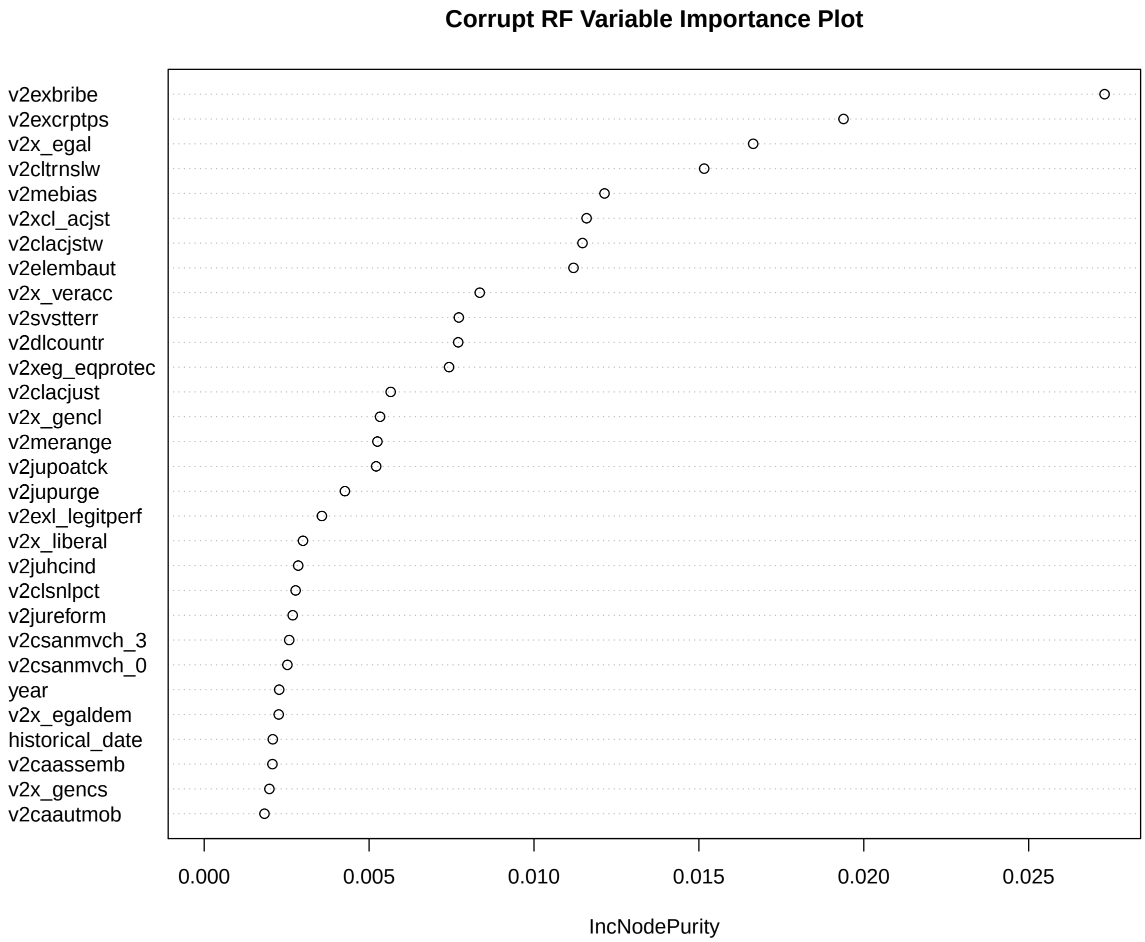
The variance inflation factor of the three selected variables is slightly high, though acceptable as they have not exceeded the value 3. Although these VIF values are by no means ideal, it has to be taken into account that many of the variables in the VDem data-set are constructed from other indexes, making it difficult to sift through the hundreds of indicators included. Any variable with a VIF higher than 3 will be deemed unacceptable, and will be discarded through another filtration process, until a set of variables with acceptable VIF and Adjusted R-squared values has been reached. This specification of a VIF value no greater than 3 will also be applied to the Random Forest approach.

1.6 Random Forest Approach

To take into account multicollinearity, I conducted a number of variance inflation factor tests on the variables that were highlighted to be of importance. The Rule of Law index has been removed due to its value being aggregated from a number of other indicators that can be seen as components of rule of law principles, such as judicial and executive accountability. Another set of examples would be the public sector and legislative corruption indexes, which were removed due to their VIF values being over 40. The variables removed can be seen in the formula call of the corrupt_rf model below. As mentioned earlier, any variables with a VIF higher than 3 were removed to prevent multicollinearity issues.

```
##
## Call:
##  randomForest(formula = v2x_corr ~ . - v2x_rule - v2x_neopat -      v2x_execorr - v2l
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 94
##
##           Mean of squared residuals: 0.001335147
##                     % Var explained: 80.26
```

The random forest has incorporated all of the remaining predictors from the filtered list, and created a model that has arrived at a very low mean squared error of 0.00133, though with the over 280 predictors remaining a lot of noise has been introduced that has reduced its predictive capacity. Hence, we are given approximately 80% of the variance in political corruption being explained by the selection. Once the number of predictors has been reduced greatly, we should be able to construct a model with higher predictive accuracy.



From the above variable importance plot, the three most important predictors from this random forest model will be taken and applied to a linear model on corruption. It should be noted that the three variables that have the highest node purity identified are the Egalitarian component index (v2x_egal), executive bribery and corrupt exchanges (v2exbribe) and public sector corrupt exchanges (v2excrptps). Only these three will be selected, in the interest of choosing a small subset of similar size to that chosen by the tree method. Using the forest method as compared to the tree method, we can easily choose the top three variables by importance rather than having to apply multiple steps.

```
##  v2exbribe   v2x_egal v2excrptps
##   2.231124   2.110221   1.645017
```

In comparison to the tree method's model, the predictors chosen by the random forest method appear to have lower VIF values on average when fit to a linear model. Two of these predictors, the egalitarian index and executive bribery, have VIF values greater than 2 which is not ideal.

1.7 Model Comparison

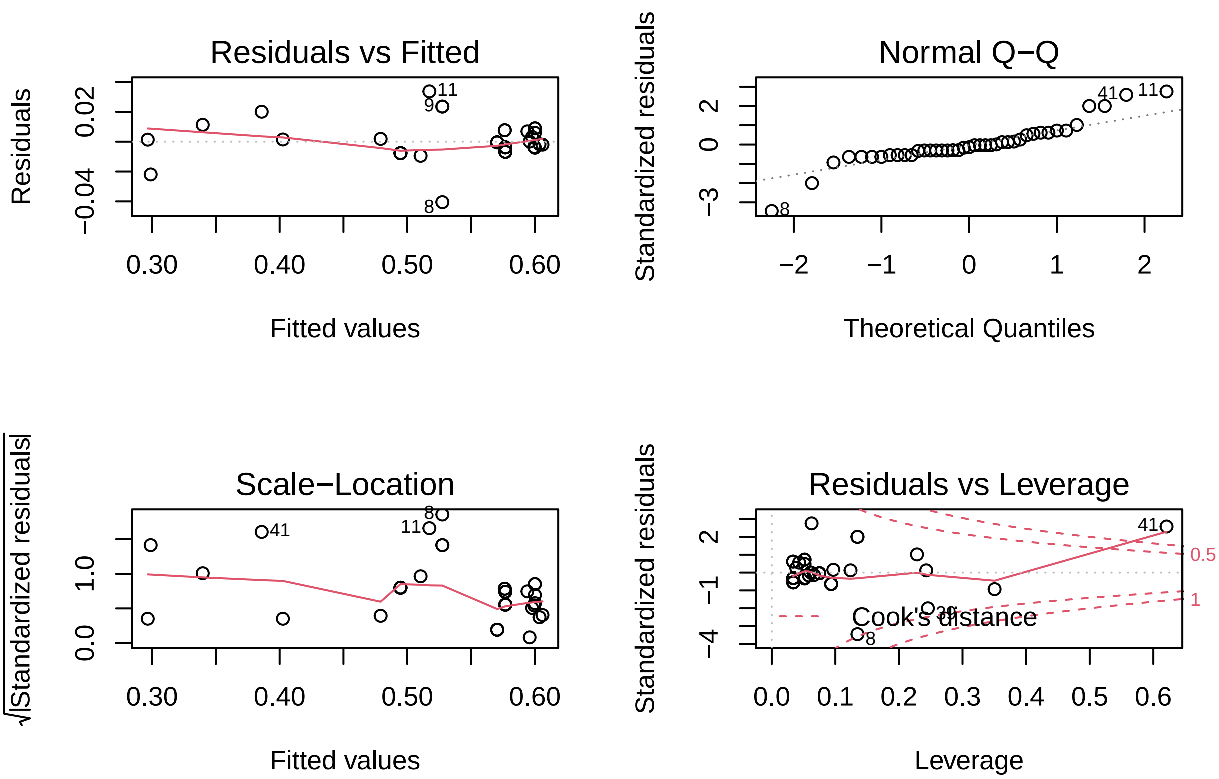
Below can be found two summaries for the models that were fit using variables selected by either method. In the first, we see the random forest's model, which included executive bribery, egalitarian index and public sector corrupt exchanges as explanatory variables. We are given a multiple R-squared of 0.979 and an adjusted R-squared which is only slightly reduced, at 0.977. These values are only important in the context of this study in comparing the amount of variance explained by the competing predictors chosen by one of the two variable selection methods. It is also worth stating that we can observe a marked increase in variability explained (80% with the full set of predictors, compared to 97.9% with only the three most important) by the random forest approach by shrinking the number of predictors included.

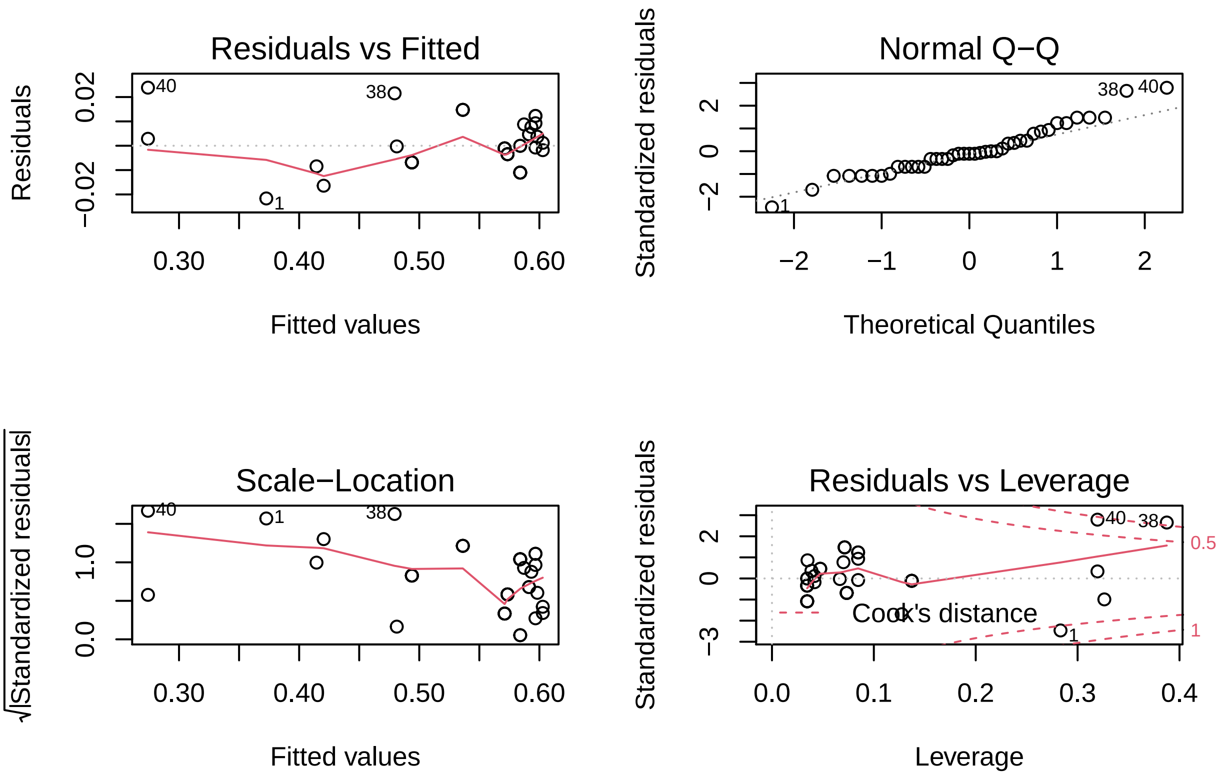
These summaries are most valuable in the sense that it can give us an indication of the statistical significance of the chosen predictors. When we observe the t-values and p-values, I can surmise that all three are highly statistically significant, and that the overall model is also significant given the minuscule size of its p-value. As such, we can deduce that these three variables are indeed highly valuable predictors of political corruption in the context of Malaysia.

```
##
## Call:
## lm(formula = v2x_corr ~ v2exbribe + v2x_egal + v2excrptps, data = clean_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.040507 -0.006818 -0.001692  0.005922  0.033662
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.699991   0.059019  11.860 3.60e-14 ***
## v2exbribe   -0.098396   0.005862 -16.784  < 2e-16 ***
## v2x_egal    -0.462219   0.093698  -4.933 1.73e-05 ***
## v2excrptps  -0.110033   0.009618 -11.440 1.04e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01263 on 37 degrees of freedom
## Multiple R-squared:  0.9787, Adjusted R-squared:  0.977
## F-statistic: 567.2 on 3 and 37 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = v2x_corr ~ v2x_execorr + v2juaccnt + v2clrgstch_3,
##     data = clean_data)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.021647 -0.006844 -0.001077  0.004664  0.023854
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.247772   0.019969  12.408 9.41e-15 ***
## v2x_execorr   0.502891   0.023823  21.110  < 2e-16 ***
## v2juaccnt    -0.086907   0.008662 -10.034 4.19e-12 ***
## v2clrgstch_3 -0.076110   0.015854  -4.801 2.61e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01038 on 37 degrees of freedom
## Multiple R-squared:  0.9856, Adjusted R-squared:  0.9845
## F-statistic: 846.2 on 3 and 37 DF,  p-value: < 2.2e-16
```





In the first of the two residual plots, the variables chosen by the random forest approach generated a model with residuals presented in a somewhat skewed pattern. But the assumption of linearity seems to hold, as there is no apparent presence of a pattern in the residuals, while the qqplot appears relatively normal, although the presence of heavy tails is noted. The residuals vs leverage plot reveals a handful of outliers, however it is less of a concern compared to the decision tree approach. Furthermore, we can see that across the plots for the two models, that the outliers highlighted are in fact observations 38, 40 and 41, which are amongst the most recent years in which variability of the political corruption index in Malaysia has been fluctuating due to perceptions that corruption has decreased under the new regime.

In the decision tree method's customised model, we can see from the residual plots that the assumption of linearity appears to be at question, as the residuals are concentrated towards the right. The qqplot appears to indicate normality, though with heavy tails also putting the assumption at risk. The leverage plot also shows at least 3 points with high leverage, which would be a concern if we did not already know of the rapid decline in perceived corruption levels in those years. Overall, it seems that a linear model will suffice for the moment, though a larger sample of data (perhaps multiple countries and a longer time-frame) would be necessary to see if these assumptions hold.

1.8 Test of Generalisability

At this stage of the analysis, the variables selected by both the tree and forest based methods will be applied to the Philippines. To create a similar subset of the VDem data-set for the Philippines, I will utilise the same procedure applied for filtering and cleaning the original Malaysian subset of data.

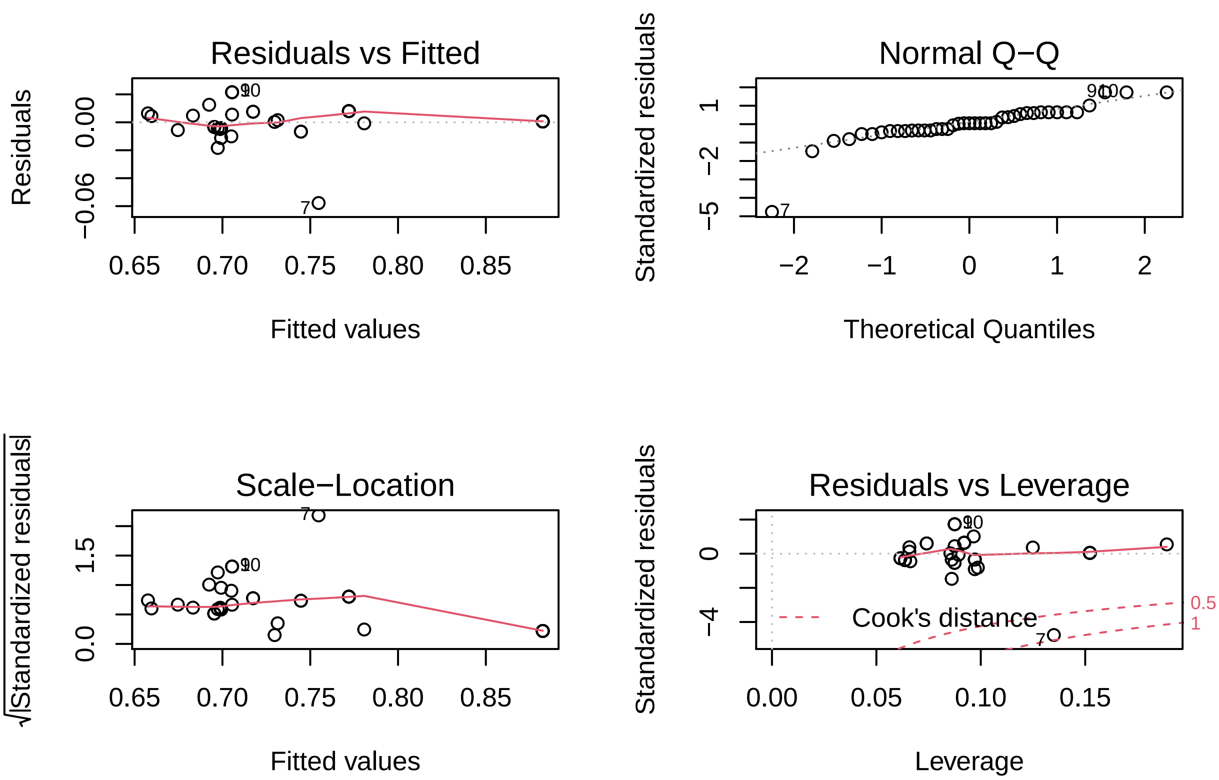
```
##
## Call:
## lm(formula = v2x_corr ~ v2x_execorr + v2juaccnt + v2clrgstch_3,
##     data = clean_data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057773 -0.004349  0.000577  0.007552  0.021503
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.495952   0.033646  14.740  < 2e-16 ***
## v2x_execorr   0.483147   0.025898  18.656  < 2e-16 ***
## v2juaccnt     0.013726   0.005853   2.345  0.02450 *
## v2clrgstch_3 -0.080124   0.022540  -3.555  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01304 on 37 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9635
## F-statistic: 353.3 on 3 and 37 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = v2x_corr ~ v2exbribe + v2x_egal + v2excrptps, data = clean_data2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.038275 -0.014123  0.000801  0.012130  0.040877
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.79756    0.03545  22.499  < 2e-16 ***
## v2exbribe   -0.03900    0.01960  -1.990   0.0540 .
## v2x_egal    -0.33292    0.07035  -4.732 3.22e-05 ***
## v2excrptps  -0.05627    0.02345  -2.399   0.0216 *
## ---
```

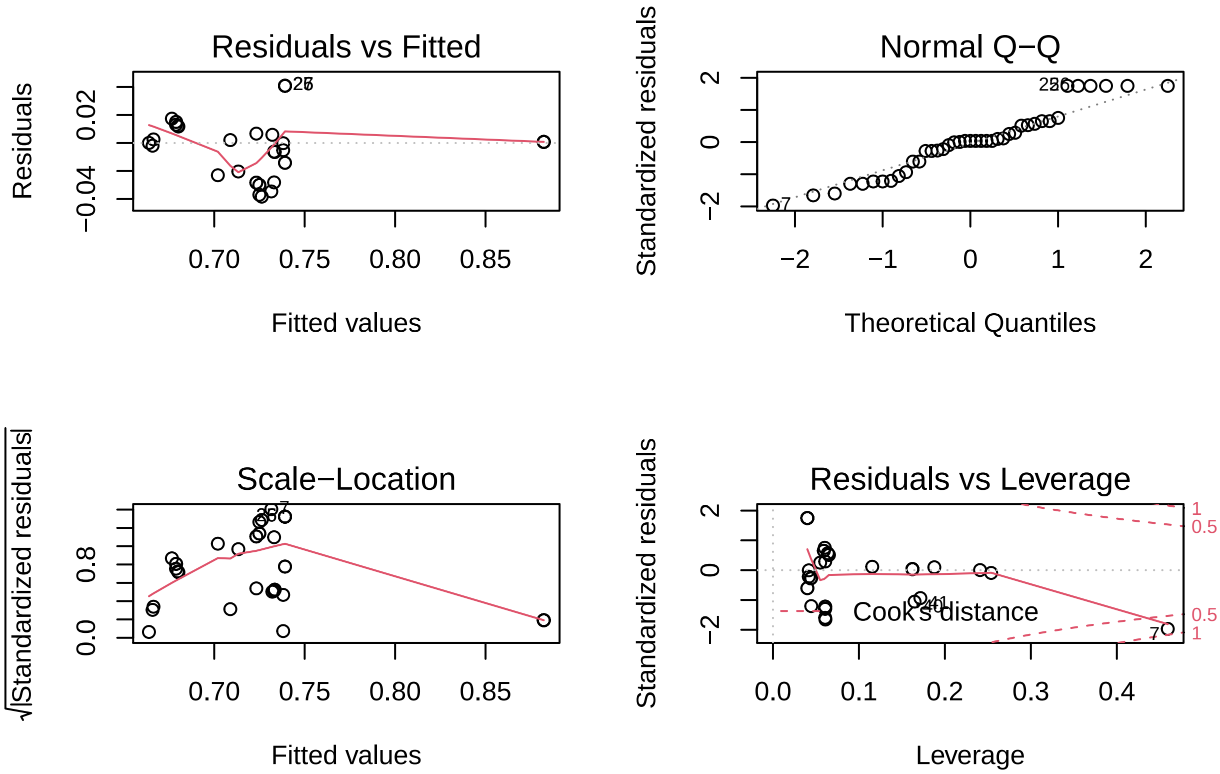
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02384 on 37 degrees of freedom
## Multiple R-squared:  0.8873, Adjusted R-squared:  0.8781
## F-statistic: 97.09 on 3 and 37 DF,  p-value: < 2.2e-16
```

In the summaries of the two methods, we can immediately see that the variables selected by the decision tree method provide a model with a better fit, in comparison to the random forest selected variables. Both models can be considered statistically significant, given overall model p-values for both are very small. However, if we compare the individual p-values of coefficients, we find that one of the predictors in the forest based model cannot be considered significant at the 0.05 level of significance, while another (public sector corruption, v2excrptps) is almost at the point of passing this significance threshold as well.

In comparing the multiple and adjusted R-squared values, the predictors chosen by the tree based method definitively explain a greater proportion of the variability of political corruption in the Philippines, with at the very least 96.4% of variance in corruption explained, in comparison to the alternative model, which achieved an adjusted R-squared of 0.878, or 87.8% of the variance explained by included indicators.



This is particularly good, considering that the predictors for executive corruption, judicial accountability and stronger civil liberties characteristics (label 3, meaning respect for civil liberties are stronger in areas that are more economically developed), could explain a similar level of variability of corruption in Malaysia (with the same model explaining 97.7% variance in Malaysian corruption). These results bode well for further application to future research that is inclusive of all nations within the greater ASEAN regional grouping.



The residuals plots of both models largely confirm the initial findings, in that it appears that the variables selected via decision tree in the first plot have created a better fitting model with more cleaner looking residuals. We see that OLS assumptions are on the whole held in the case of the tree based model, while in the second plot for the forest method we can see irregularities in the residuals vs fitted plot and scale-location plot that would suggest non-linearity, while the qqplot also shows a number of observations as violating the assumption of normality. Overall, it seems that without applying non-linear terms to the model, it is more appropriate to go with the decision tree based linear model.

1.9 Discussion

From the two separate methods utilised, we have arrived at a much smaller selection of predictor variables that can be used to explain a large proportion of the variance in political corruption in Malaysia. In the decision tree method, which utilised variable subset selection as an initial filter, three variables were identified as key regressors for political corruption, including the executive corruption index, judicial accountability and stronger civil liberties characteristics. In the random forest method, the three most important predictor variables that were selected include the egalitarian component index, executive bribery and corrupt exchanges and public sector corrupt exchanges. The two sets of variables are also listed below:

- Decision Tree Method Selected Variables

1. Executive corruption index (v2x_execorr)
2. Judicial accountability (v2juaccnt)
3. Stronger civil liberties characteristics (v2clrgstch)

- Random Forest Method Selected Variables

1. Egalitarian component index (v2x_egal)
2. Executive bribery and corrupt exchanges (v2exbribe)
3. Public sector corrupt exchanges (v2excrptps)

While both methods yielded models that could be considered statistically significant, and do in fact account for a large proportion of variance in corruption, the decision tree based method yielded a model that had greater success in its application to a different country setting, achieving approximately 10% higher explanatory power than the other. The results are somewhat unexpected in the case of the application of both models to the Philippines. In the Random Forest variable model, two of the predictors, public sector corruption and executive bribery were found to have lower statistical significance than some of the variables identified by the decision tree method.

The coefficient for public sector corruption specifically, was found to have breached the 0.05 level of significance threshold, rendering its statistical reliability as questionable for further use. It seems that certain aspects of understanding growth of political corruption in the Malaysia-specific literature does not carry over to different contexts. Although public sector corruption was highlighted as a key component of the pervasive corruption experienced in Malaysia in the past few decades, as argued by Kapeli (Kapeli and Mohamed 2019, pg 552-553), it appears that executive corruption is a better indicator in the case of the Philippines.

The difference in the two variables comes down to the exact action undertaken, as the VDem Institute distinguishes executive bribery as being an measure of how often a country's political executive (cabinet members, head of state etc.) "grant favors in exchange for bribes, kickbacks, or other material inducements" (Coppedge, Gerring, Knutsen, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Luhrmann, Marquardt, McMann, et al. 2020, pg 112-113). This variable is also on a positive ordinal scale from 0 to 4, where 0 is the response, "It is routine and expected", while 4 is the response, "It never, or hardly ever, happens".

As such, we can see that as the value of the executive bribery variable goes up, we would expect a decrease in the political corruption index. Executive corruption on the other hand, has an added component, as while its definition includes the granting of favours in exchange for bribes and kickbacks, this measure also includes stealing, embezzlement and misappropriation of public funds and state resources for personal or family use. This added component is gives us clear indication that the executive corruption index is a far more complete metric, hence its prioritisation by the more accurate tree based method. With both Malaysia and the Philippines having coefficients close to 0.5, this indicates that an increase in executive corruption has a significant associated effect on increased political corruption.

Executive corruption is a highly important predictor for the overall pervasiveness of political corruption within a country, as political executives tolerating corruption and bribery under their watch allows illicit behaviour and practices to develop as part of bureaucratic culture and structures, as was the case in Malaysia in the past two decades (Hashim 2017, pg 559). This top-down effect of corrupt activities has wide-ranging implications, as by tolerating pervasive corruption at lower bureaucratic structures in the governmental hierarchy, there is also the propensity for corruption to become ingrained in private sector practices as well.

This is in consideration of the fact that civil servants have a duty not only to safeguard governance structures from political influence and corruption, but also to prevent private sector entities from exerting their own influence to secure special interests (Shim and Eom 2008, pg 300). With no safeguard against these interests, corruption at the executive level can easily exacerbate overall political corruption by an inability to deter smaller scale bribery and favouritism between public and private sector entities. To see this variable chosen as a key predictor for corruption via variable selection is no surprise, and could very well be applicable to most countries facing pervasive corruption within political and bureaucratic structures.

Judicial accountability is seen to have a negative relationship with political corruption, a straightforward and widely accepted notion. This relationship is captured well by a statement from Berggren & Bjornskov, where they argue that "When corruption is present, and judicial accountability is compromised, "unfair procedures" in public governance tend to become endemic" (Berggren and Bjørnskov 2020, pg 2-3). In the case of Malaysia, we do indeed observe a major fall in corruption perceptions as a result of judicial accountability rising rapidly in the past 5 years, as can be seen in the plot of Figure 1 below. The estimated coefficient from the model, with a value of -0.0869 for Malaysia, indicates that there is a reasonable negative effect on political corruption with a unit increase in judicial accountability, and with the minuscule p-value should be considered a statistically significant relationship.

However, the coefficient for judicial accountability is somewhat puzzling in the case of the Philippines, as it has a positive coefficient of 0.0137, implying greater rule of law and better governance oversight in greater judicial accountability is associated with higher levels of corruption. Judging from the plot in Figure 2, this could be a result of the major variability in accountability relative to stable trends seen in political corruption, meaning that the model may not have accurately captured the relationship between the two variables. This is given credence by the somewhat high p-value of 0.02450, though this estimate means that it should still be considered significant at the 0.05 level of confidence.

Indeed, Berggren & Bjornskov found that the majority of research on the interaction of judicial accountability and corruption was closely related, succinctly stating that "corruption is the outcome of monopoly plus discretion minus accountability", and that various features of democracy like judicial accountability may help stifle corrupt behaviour in society overall (Berggren and Bjørnskov 2020, pg 3). They also argue that there is a two-sided element to this relationship, as governance structures are not infallible and may be affected by the consequences of corruption, specifically shaping which policies are instituted in the first place (Ibid), which may in part explain the unexpected relationship for the Philippines.

Stronger civil liberties characteristics, which is specifically the perception of stronger respect for civil liberties by government officials in wealthier districts, was found to have a negative relationship with political corruption, on the order of -0.0761 for Malaysia and -0.0801 for the Philippines. The implication of this finding seems to be that the public perceives political corruption to be lower when there is evidence of rule of law being upheld, in the sense that individual civil liberties are being protected in wealthy areas. This could be as a response to the wealthy and other elites being left unmolested by unjust government interference and that their rights to freedom of action and speech are also not repressed.

Some research on the relationship between civil liberties and corruption exists, with Roca & Alidedeoglu-Buchner identifying a significant statistical relationship between corruption and political rights, which they argue are closely related to civil liberties, as being inversely related to corruption perception (Roca and others 2010, pg 8-10). In their paper however, the focus is on the maturity of a democracy and does not include any element of the economic wealth of citizens or districts being related to stronger civil liberties. Further investigation on this specific indicator is suggested, as its role as a predictor of corruption remains somewhat abstract at this juncture, given its specificity to the wealth of a region playing a role. It is however, a statistically significant predictor, given the p-value for the coefficient was 0.00106, and as such is worth maintaining for future research.

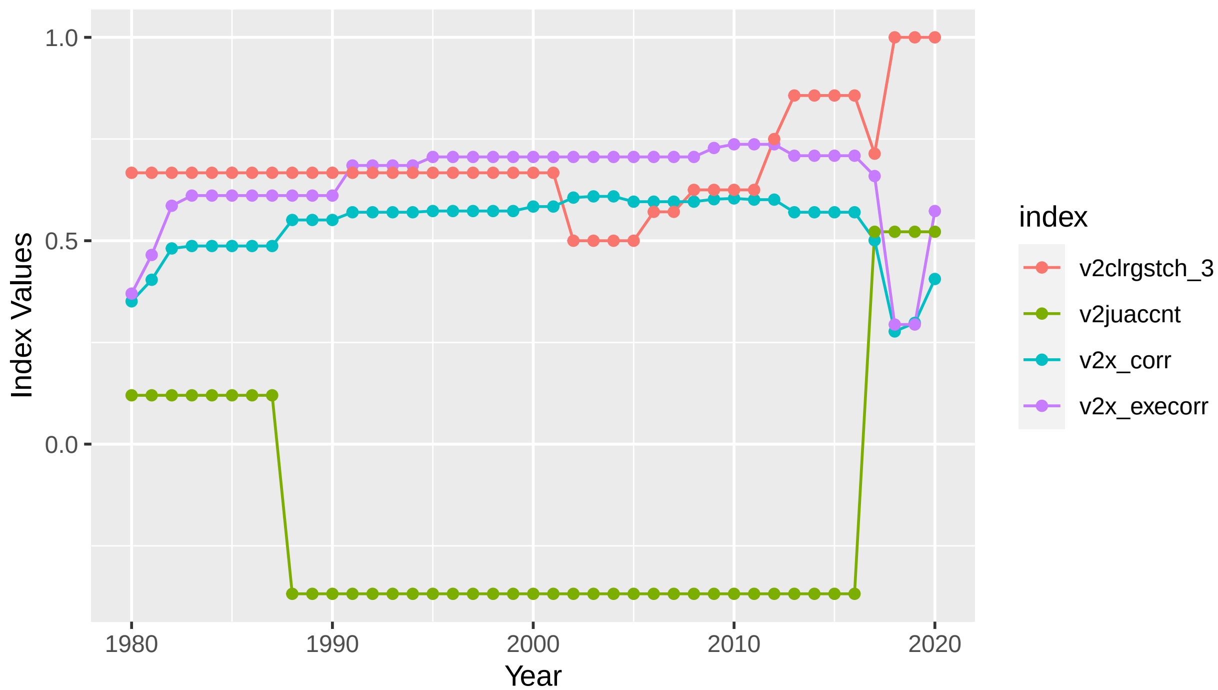
Regarding the variables chosen via the forest based method, and discarded thereof, the selection of public sector corrupt exchanges by the random forest approach was very much in line with existing literature, as both Kapeli (Kapeli and Mohamed 2019, pg 552-554) and Hashim (Hashim 2017, pg 559) highlighted the inadequacies of reporting and law enforcement against corrupt activities by public officials and civil servants. In the case of the latter, Hashim noted that the Malaysian Anti-Corruption Commission, or MACC, had a total of 3533 arrests for charges of corruption between the year 2011 and 2015. Of these individuals arrested, 1408 of them were public officials (which is nearly 40% of the total), demonstrating the high level of public sector corruption in recent years within the country. It is perhaps as a result of the high levels of Malaysian public sector corruption that the random forest prioritised this indicator relative to judicial accountability or stronger civil liberties in wealthier districts.

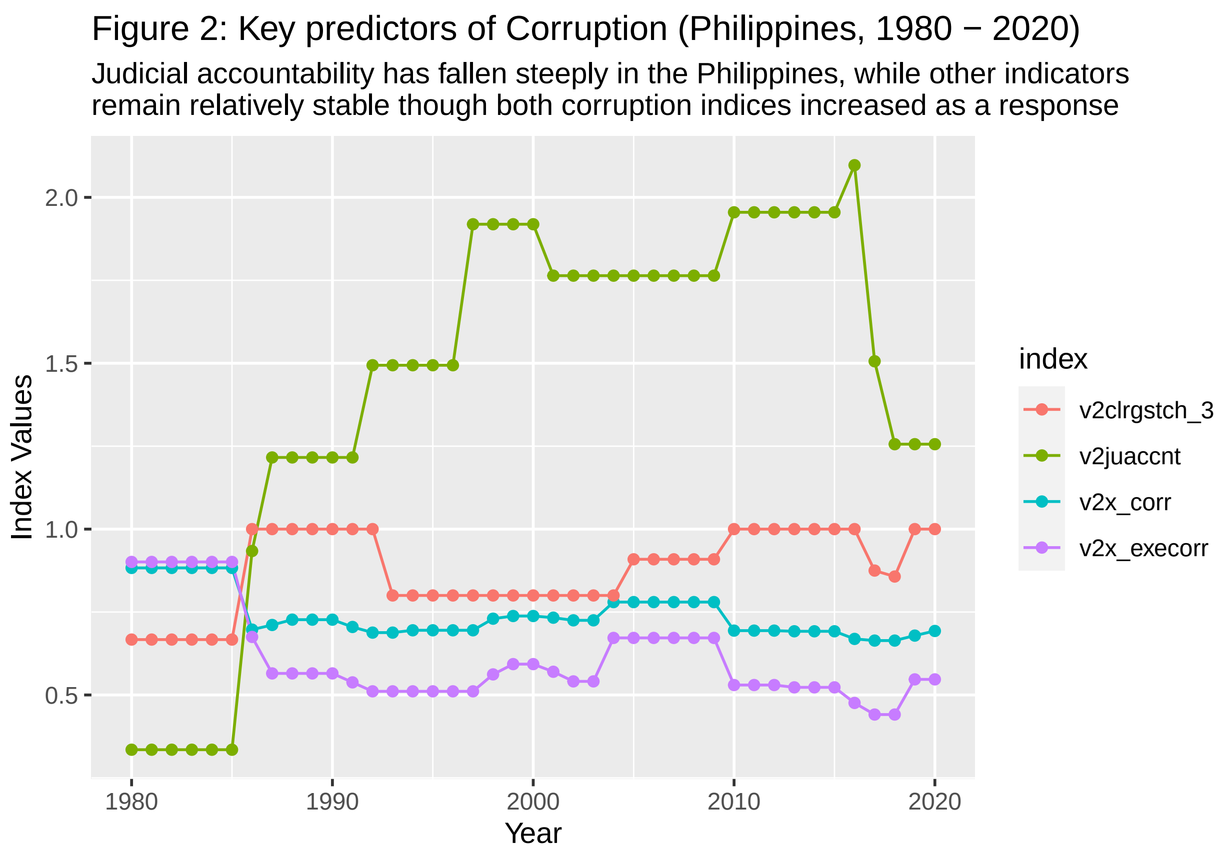
In the broader literature, it has been posited that the "wage rate of civil servants relative to that in the private sector has an impact on the incidence of corruption. Relatively low wages in the public sector will make the benefit of a given bribe seem greater and the cost of losing the government job if the bribe is discovered seem less" (Elbahnasawy and Revier 2012, pg 314). As such, it was quite unexpected that this indicator was not found to be more significant in the case of the Philippines, and may be a reflection of a more disciplined or well-enforced public sector relative to the downward trend in Malaysia, or perhaps that there is higher parity in the wage rate of Philippine civil servants relative to the private sector.

This demonstrates that either of the two methods may me somewhat vulnerable to overfitting and hence have higher variance. This would of course mean that the models are reliant upon the training data as well as the chosen indicators, and hence may have problems with external validity/generalisation due to high testing error rates. While the tree based method achieved high accuracy for both of these countries, we can see that this could be due to the relative similarity in the variability of indicators between the two countries in the plots below, the main distinguishing difference being the major increase in judicial accountability in Malaysia recently.

Figure 1: Key predictors of Corruption (Malaysia, 1980 − 2020)

Judicial accountability and civil liberties rose steeply, in step with a rapid decline in both corruption indices. However, a return to previous levels has occurred only a few years later



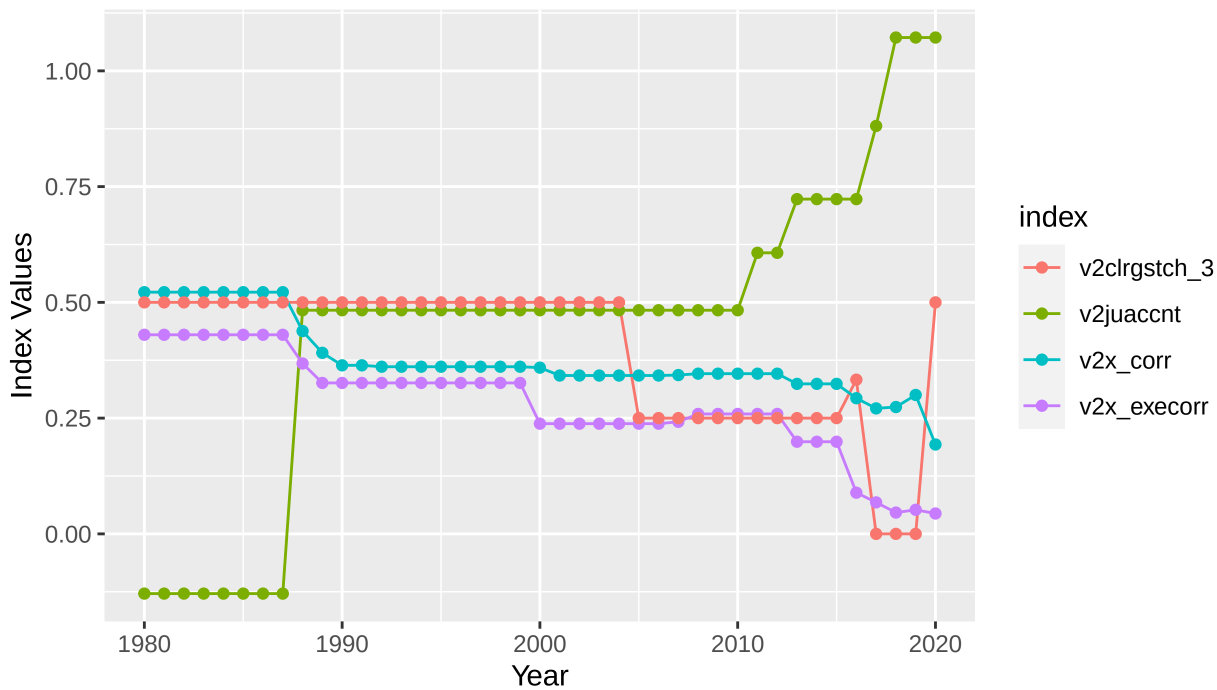


In order to confirm the external validity of this model, a wider net has to be cast, with additional nations and more data variation being investigated with the same set of variables chosen, that is executive corruption, judicial accountability and stronger civil liberties. The real test of generalisability at this stage, would then be to apply the model to countries that did not have high levels of corruption (or at least moderate levels). Although such a test goes beyond the scope of this study, a brief look at the variability in the same set of indicators for much less corrupt country in Asia, Taiwan, shows that there could be a possibility that these indicators can also be used for nations with different characteristics.

Interestingly, variance in judicial accountability seems to be tied to changes political corruption in all three plots, providing evidence that this could definitely be a universal predictor of corruption that warrants further investigation. However, it also appears that rapid increases in judicial accountability and stronger civil liberties has a less noticeable effect on reducing political and executive corruption in Taiwan relative to Malaysia and the Philippines. This may indicate that the chosen indicators have a lesser role as predictors of political corruption in perhaps higher incomes states, or more democratic states like Taiwan, though further testing of models would be required to verify this hypothesis.

Figure 3: Key predictors of Corruption (Taiwan, 1980 − 2020)

Taiwan has lower corruption indices overall, and major increases in civil liberties and judicial accountability seem to have a lower impact on reducing corruption, so the relationship between indicators may differ by context



Finding such contrary results would not necessarily disprove the utility of the tree based method's accuracy in explaining variance in corruption. It could simply be the case, that the chosen variables are indeed key predictors of corruption, but only for a nations with a specific set of conditions. For instance, the similarities between Malaysia and the Philippines would be that they are both members of ASEAN, have had high levels of political corruption historically, and are middle income countries. There could be other possible unidentified features that are not within the VDem data-set that could make this model applicable to other states, for instance, high levels of religious belief, a relatively young population and even tropical weather. If these were indeed characteristics that help explain the model's accuracy, then it would certainly be the case that it could be applied to other nations in Asia, Africa or Latin America. As such, future iterations of this process may merge additional data-sets to expand the pool of possible indicators of interest for explaining variation in political corruption.

1.10 Conclusion

To conclude, three predictors have been found to be the most important explanatory variables in explaining variation in political corruption in Malaysia. Of the two methods constructed, it is the variable selection and decision tree based method that has resulted in the most accurate predictive linear model. This is true of not only political corruption trends in Malaysia, but also in the Philippines, as was found during a brief generalisability test. Similar levels of variance were accounted for in both countries by the model created with variables chosen by the decision tree method. Furthermore, the chosen set of predictors remained statistically significant at all conventional levels of significance for both the Malaysian and Philippine data.

The identified variables include executive corruption, stronger civil liberties characteristics and judicial accountability, with these indicators falling largely in line with existing theories in political corruption literature. Although the importance of specific indicators changed somewhat depending on the country in question, expected relationships were found to be largely in agreement. But in observing the data of an unrelated nation, Taiwan, there is some indication that some of the variables identified may be weaker in predicting corruption in wealthier, democratic states.

This process of narrowing down a set key variables could prove invaluable in future research, not only on trends of political corruption in Malaysia, but also for its neighbours in the Association of Southeast Asian Nations. Perhaps a similar approach could be applied to other regions of the world, with further specifications that properly incorporate the unique characteristics of different contexts. Further testing would be required, but considering the extensive nature of the VDem Data-set, it is well within the means of other researchers to apply similar variable selection methods and predictive models as have been developed in this study, to their chosen context.

1.11 Appendix

```
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE)

library(tidyverse)
library(ggplot2)
library(bibtex)
library(tree)
library(dplyr)
library(leaps)
library(car)
library(randomForest)
library(knitr)
# Read in Vdem Dataset
vdem_data <- readRDS("Data/V-Dem-CY-Full+Others-v11.1.rds")
# Removal of NAs and filtering year to 1980 onwards
malaysia_data <- vdem_data %>%
  filter(country_name == "Malaysia")

coredata <- malaysia_data %>%
  filter(year >= "1980")

coredata <- coredata %>%
  select_if(~ !any(is.na(.)))

unique_data <- sapply(lapply(coredata, unique), length)
unique_prune <- coredata[ , unique_data >= 3]
# Regex to be used to filter additional versions of variables included in Vdem
# Ordinal scale, mean, standard deviation measures etc.

versions <- "(_osp|_ord|_codelow|codehigh|_sd|_mean|_nr|_3C|_4C|_5C|e_|commnt)"

prune_ver <- grep(versions, colnames(unique_prune),
                  value=TRUE, ignore.case =F)

clean_data <- unique_prune %>%
  select(-prune_ver)
# Subset Selection to identify significant variables.
# v2xnp_regcorr removed due to high collinearity with v2x_corr
set.seed(333)
corrupt_subset <- regsubsets(v2x_corr ~. -v2xnp_regcorr,
                             data = clean_data, method = "forward")
```

```r
corsubset_rss <- subsets(corrupt_subset, statistic="rss",
                         legend = TRUE, max.size = 6,
                         main = "Corruption Subset: Residual Sum of Squares")
corsubset_cp <- subsets(corrupt_subset, statistic="cp",
                        legend = TRUE, max.size = 6,
                        main = "Corruption Subset: Mallow's Cp")
corsubset_bic <- subsets(corrupt_subset, statistic="bic",
                         legend = TRUE, max.size = 6,
                         main = "Corruption Subset: Bayesian Information Criteria")
corsubset_adjr2 <- subsets(corrupt_subset, statistic="adjr2",
                           legend = TRUE, max.size = 6,
                           main = "Corruption Subset: Adjusted R-squared")
newtree <- tree(v2x_corr ~ v2elsnlfc_13 + v2lginelup +
                  v2juaccnt + v2clrgstch_3 + v2svdomaut +
                  v2x_execorr, data = clean_data)
summary(newtree)
plot(newtree, main = "Political Corruption Decision Tree: Malaysia")
text(newtree)

set.seed(777)
treevar.fit <- lm(v2x_corr ~ v2x_execorr + v2juaccnt +
                    v2clrgstch_3, data = clean_data)
# Variance inflation factor for tree based method
vif(treevar.fit)
set.seed(369)
corrupt_rf <- randomForest(v2x_corr ~. -v2x_rule -v2x_neopat
                           -v2x_execorr -v2lgcrrpt -v2x_pubcorr
                           -v2jucorrdc -v2juhccomp -v2xnp_regcorr
                           -v2juaccnt -v2jupack -v2exembez
                           -v2clrspct -v2xcl_rol -v2jucomp
                           -v2x_jucon -v2mecrit -v2xnp_pres
                           -v2clacjstm -v2xcl_prpty -v2clprptym
                           -v2x_clpriv,
                           data = clean_data)
corrupt_rf
varImpPlot(corrupt_rf, main = "Corrupt RF Variable Importance Plot")
set.seed(0101)
forest.fit <- lm(v2x_corr ~ v2exbribe + v2x_egal +
                   v2excrptps, data = clean_data)
# Variance inflation factor for forest based method
vif(forest.fit)
summary(forest.fit)
summary(treevar.fit)
# Random Forest model's residual plots
```

```r
par(mfrow = c(2,2))
plot(forest.fit)
# Decision Tree model's residual plots
par(mfrow = c(2,2))
plot(treevar.fit)
# Removal of NAs and filtering year to 1980 onwards,
# Country selection of the Philippines only
philippines_data <- vdem_data %>%
  filter(country_name == "Philippines")

testdata <- philippines_data %>%
  filter(year >= "1980")

testdata <- testdata %>%
  select_if(~ !any(is.na(.)))

# No columns with fewer than 3 unique values will be included,
# to prevent modelling errors.
# unique_phil created as a vector to allow filtering
unique_data2 <- sapply(lapply(testdata, unique), length)
unique_phil <- testdata[ , unique_data2 >= 3]
# Regex ("versions") will be applied,
# to filter different versions of key indices
# including ordinal, mean and others.

versions2 <- "(_osp|_ord|_codelow|codehigh|_sd|_mean|_nr|_3C|_4C|_5C|e_|commnt)"

# Filter variables with non-unique/low variability values
prune_phil <- grep(versions2, colnames(unique_phil),
                   value=TRUE, ignore.case =F)

clean_data2 <- unique_phil %>%
  select(-prune_phil)
# Linear Model using tree method's selected  variables
set.seed(121212)
tree.fitphil <- lm(v2x_corr ~ v2x_execorr +
                     v2juaccnt + v2clrgstch_3,
                   data = clean_data2)
# Linear Model using random forest method's selected variables
set.seed(212121)
forest.fitphil <- lm(v2x_corr ~ v2exbribe +
                       v2x_egal + v2excrptps,
                     data = clean_data2)
```

```
summary(tree.fitphil)

summary(forest.fitphil)
# Decision Tree method based model, residual plots
par(mfrow = c(2,2))
plot(tree.fitphil)
# Random Forest method based model, residual plots
par(mfrow = c(2,2))
plot(forest.fitphil)
# Plot showing differentiation between Malaysia & Philippines
# for the variability between indicators

# Malaysian Indicators
mal_plot <- clean_data %>%
  select(year, v2x_corr, v2x_execorr, v2juaccnt, v2clrgstch_3) %>%
  pivot_longer(cols = c(v2x_corr, v2x_execorr, v2juaccnt, v2clrgstch_3),
               names_to = "index", values_to = "values") %>%
  ggplot(aes(year, values, col = index)) +
  geom_line() +
  geom_point() +
  labs(title = "Figure 1: Key predictors of Corruption (Malaysia, 1980 - 2020)",
       subtitle = "Judicial accountability and civil liberties rose steeply, in step wit
       x = "Year",
       y = "Index Values")

# Philippine Indicators

phil_plot <- clean_data2 %>%
  select(year, v2x_corr, v2x_execorr, v2juaccnt, v2clrgstch_3) %>%
  pivot_longer(cols = c(v2x_corr, v2x_execorr, v2juaccnt, v2clrgstch_3),
               names_to = "index", values_to = "values") %>%
  ggplot(aes(year, values, col = index)) +
  geom_line() +
  geom_point() +
  labs(title = "Figure 2: Key predictors of Corruption (Philippines, 1980 - 2020)",
       subtitle = "Judicial accountability has fallen steeply in the Philippines, while
       x = "Year",
       y = "Index Values")
par(mfrow = c(1,2))
mal_plot + scale_fill_discrete(labels = c("Civil Liberties",
                                "Judicial Accountability",
                                "Political Corruption",
                                "Executive Corruption"))
phil_plot + scale_fill_discrete(labels = c("Civil Liberties",
```

```r
                                  "Judicial Accountability",
                                  "Political Corruption",
                                  "Executive Corruption"))
# Removal of NAs and filtering year to 1980 onwards,
# Country selection of the Taiwan only
tai_data <- vdem_data %>%
  filter(country_name == "Taiwan")

newdata <- tai_data %>%
  filter(year >= "1980")

newdata <- newdata %>%
  select_if(~ !any(is.na(.)))

# Plot showing variability of indicators

# Taiwanese Indicators
tai_plot <- newdata %>%
  select(year, v2x_corr, v2x_execorr, v2juaccnt, v2clrgstch_3) %>%
  filter(year >= "1980") %>%
  pivot_longer(cols = c(v2x_corr, v2x_execorr, v2juaccnt, v2clrgstch_3),
               names_to = "index", values_to = "values") %>%
  ggplot(aes(year, values, col = index)) +
  geom_line() +
  geom_point() +
  labs(title = "Figure 3: Key predictors of Corruption (Taiwan, 1980 - 2020)",
       subtitle = "Taiwan has lower corruption indices overall, and major increases in c
       x = "Year",
       y = "Index Values")
tai_plot
```

Bibliography

Berggren, Niclas, and Christian Bjørnskov. 2020. "Corruption, Judicial Accountability and Inequality: Unfair Procedures May Benefit the Worst-Off." *Journal of Economic Behavior & Organization* 170: 341–54.

Berman, Evan M. 2016. *Public Administration in Southeast Asia: Thailand, Philippines, Malaysia, Hong Kong, and Macao.* CRC Press.

Bland, Ben. 2018. "Malaysia: The Obstacles to Dismantling the Old Regime." *FT.com*.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, et al. 2020. "V-Dem Codebook v7.1." Varieties of Democracy (V-Dem) Project. https://www.v-dem.net/en/data/data-version-10/.

———. 2020. "V-Dem Country-Year/Country-Date Dataset v10." Varieties of Democracy (V-Dem) Project. https://www.v-dem.net/en/data/data-version-10/.

Edwards, Scott. 2018. *Malaysia's Elections: Corruption, Foreign Money, and Burying-the-Hatchet Politics.* Al Jazeera Centre for Studies.

Elbahnasawy, Nasr G, and Charles F Revier. 2012. "The Determinants of Corruption: Cross-Country-Panel-Data Analysis." *The Developing Economies* 50 (4): 311–33.

Hashim, NOREHA. 2017. "Development Efforts and Public Sector Corruption in Malaysia: Issues and Challenges." *Journal of Sustainability Science and Management* 12 (2): 253–61.

Kapeli, Nur Shafiqa, and Nafsiah Mohamed. 2015. "Insight of Anti-Corruption Initiatives in Malaysia." *Procedia Economics and Finance* 31: 525–34.

———. 2019. "Battling Corruption in Malaysia: What Can Be Learned?" *Journal of Financial Crime*.

Roca, Thomas, and others. 2010. "Corruption Perceptions: The Trap of Democratization, a Panel Data Analysis." *Group d'economie Lare-Efi Du Développement Working Paper No. DT/161/2010*.

Sari, Tiyas Kurnia, Fitra Roman Cahaya, and Corina Joseph. 2020. "Coercive Pressures and Anti-Corruption Reporting: The Case of Asean Countries." *Journal of Business Ethics*, 1–17.

Shim, Dong Chul, and Tae Ho Eom. 2008. "E-Government and Anti-Corruption: Empirical Analysis of International Data." *Intl Journal of Public Administration* 31 (3): 298–316.

Siddiquee, Noore Alam. 2005. "Public Accountability in Malaysia: Challenges and Critical Concerns." *International Journal of Public Administration* 28 (1-2): 107–29. https://doi.org/10.1081/PAD-200044546.