Abstract:

The purpose of this report is to record and explain the process used to create a predictive model, which can be used to determine if significant patterns appear in the 6 World Governance Indicators for Brazil. Patterns which are evident will be modeled and scored for further use. Predictive models produced will provide a foundation and a jumping off point for further research into this area such as comparison for future models to see if short term patterns are any indication of long term patterns for this country and also to be able to compare this country with other countries to see if there are correlative relationships between the policy choices of different countries within certain regional, economic, cultural, or governance groups. The scoring engines produced will help make further refinements easier as more data is collected in future years.

Organization and purpose

This study is being done to benefit the government and people of Brazil. Specifically the government entities which make decisions and create policies which affect the 6 World Governance Indicator scores which are assigned to them each year. This model will help them to know which areas they need to focus on improving the most. Results from the model created can be used to refocus policy making into troublesome areas so that the most benefit can be gained from decision making in the least amount of time.

The scoring engine produced will be useful in the future to compare future scores and see if they are following the current predicted paths or if they are improving or degrading. The scoring engines might also be used to judge the effectiveness of different regional governments within Brazil, given that the WGI scores can be computed in the same way as for this dataset.

Summary of selected data

Dataset: World Governance Indicators

Country selected: Brazil (it is important to note that Brazil will often be referred to throughout this report by it's official World Bank Code: BRA)

Independent variable: 16 years (1996, 1998, 2000-2014)

Dependant variables: 6 total: Voice and Accountability (VA), Political Stability (PS), Regulatory Quality (RQ), Governance Effectiveness (GE), Rule of Law (RL), Control of Corruption (CC)

Research question

The research question of this analysis is: do any of the variables or combinations of variables in the WGI for Brazil exhibit any statistically significant short term trends which might be used for predictive analysis? If so, what do these patterns indicate? This question can be answered by performing analysis on each of the variables to see if any have significant patterns. Answering this question will serve as a stepping stone for further research into the patterns present in world government choices and their consequences. International organizations and world governments will benefit from such perspectives on their effectiveness and can be put to use in encouraging positive choices and behaviors which benefit the people of this nation and dissuading negative choices and behaviors which cause distress to the people of Brazil.

Model and tool choice

Given that this data is represented in a time series and is not binomial the most effective models are going to be a multiple linear regression and a neural network. The multiple linear regression will show the tendencies of the variables in linear relationships. It will not show other relationships, but transformations can be done on the data if that seems appropriate to fit the data. The neural network can help to show the connections which each variable has to every other variable before reaching their conclusion (a given year).

The tools chosen for this predictive model construction is R and the Rattle analytic package. This tool provides all of the functions necessary to describe, model, evaluate, and score this dataset with the given models.

Analytic plan

Data collection

The World Governance Indicators dataset, which includes summary data for every country in the world on each of the 6 variables where available, can be downloaded from the Worldbank at:

http://info.worldbank.org/governance/wgi/index.aspx#home

Data cleaning

The dataset is stripped down to just the 6 raw scores for Brazil by year. The independent variable is time and there are 6 independent variables.

Data summary:

Summary data was produced for each of the 6 variables for Brazil. Irregularities or patterns which might become useful later on will be made note of. Meaningful descriptive statistics for this dataset which will be recorded are the minimum, maximum, mean, median, and

standard deviation. Skewness and kurtosis provide no meaningful information since this is time services data and therefore is not normally distributed. This information will be reviewed and referenced for important clues about how best to model the data. Summary visualizations will be used to provide insights into the patterns of the data.

Data modelling

Linear regression analysis as a time series: This data is provided as a time series and it is important to see whether or not any significant patterns appear in any of the variables from which predictions might be made about near future changes in those variables.

Data evaluation

To evaluate the models it will be important to use multiple tests and see which ones show the highest accuracy of the model. Unfortunately, given so few data points, it is not reasonable to employ cross validation on this dataset since taking out 20-30% of the data would significant diminish the dataset size and the produced data sets would not have enough data in them to produce cross validation models from, so we will have to depend upon other evaluation techniques. Predicted vs Observed will likely be the most effective for evaluating any time series regressions which are produced. Simple risk charts can be employed to evaluate the accuracy of decision trees and random forests which are produced.

Data deployment and scoring

Finally, and most essential for the purpose of this analysis are the conclusions which can be drawn, if any, from the models and the scoring of the models for future use. Conclusions will be summarized and further research pursuits will be suggested. Scoring engines will be produced and stored with this report, along with all of the original data and cleaned data.

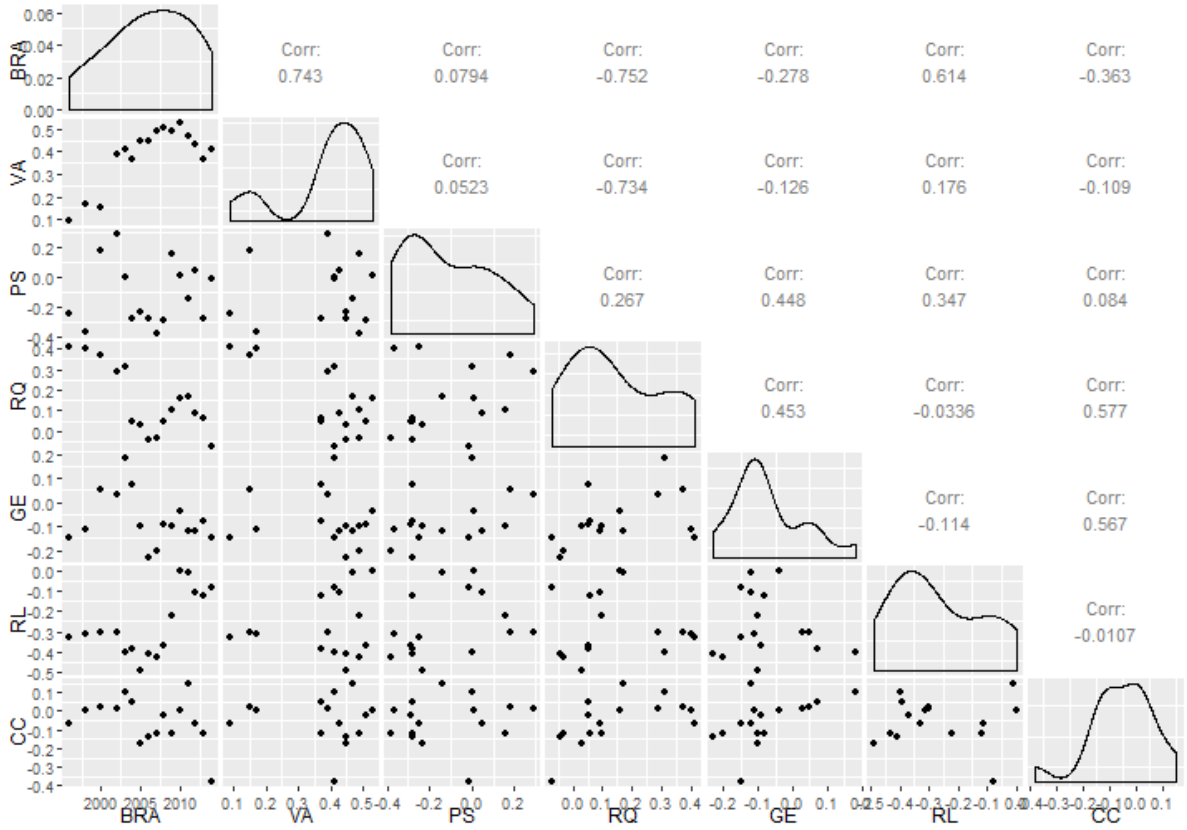
Predictive analytics

Descriptive statistics

BRA	VA	PS	RQ	GE	RL	CC
Min	0.09	-0.38	-0.07	-0.23	-0.49	-0.38
Mean	0.38625	-0.11375	0.146875	-0.0725	-0.266875	-0.055
Med	0.42	-0.185	0.095	-0.1	-0.305	-0.045
Max	0.53	0.29	0.41	0.18	0	0.15
StDv	0.13331	0.211877	0.161151	0.106927	0.156832	0.12474

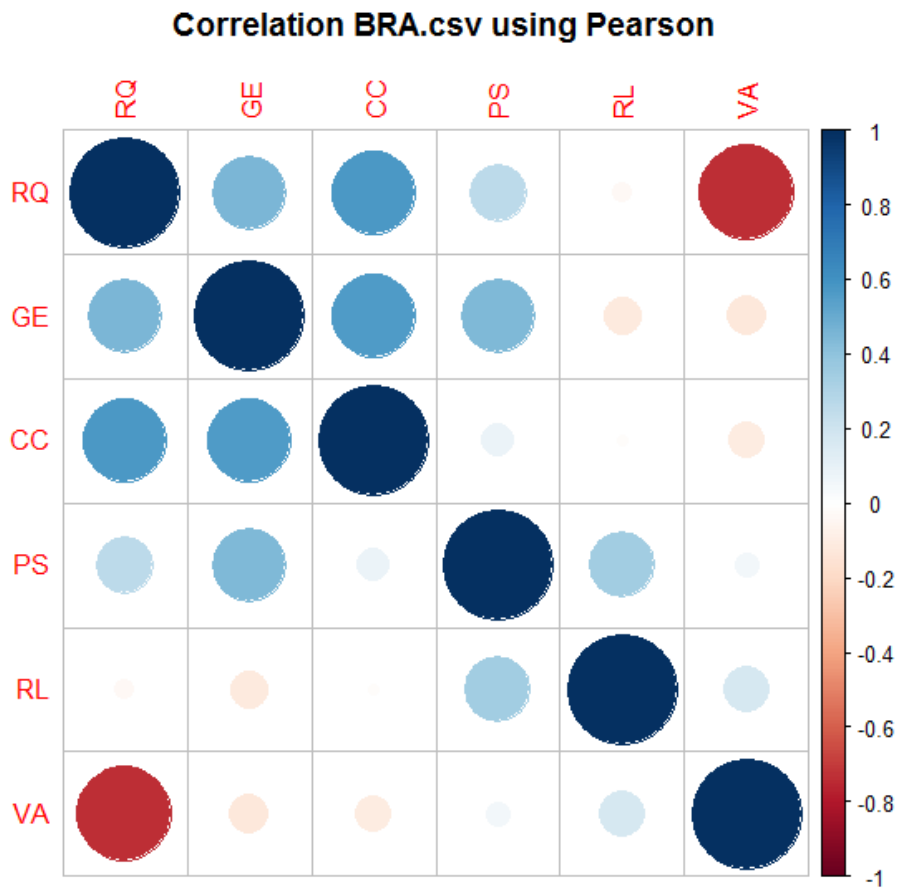
From these descriptive statistics we can see that Brazil fares poorly on most accounts, having similarly negative means and medians for all variables except voice and accountability and regulatory quality. The lowest minimum is for rule of law, with political stability and control or corruption almost as low. The only positive, and highest, minimum is voice and accountability. The lowest maximum is rule of law and the highest is voice and accountability. The lowest standard deviation is governance effectiveness, showing the least change over time, and the highest is political stability with twice as much variability as governance effectiveness, showing significantly more variability than all the other variables.

Pairs comparison:



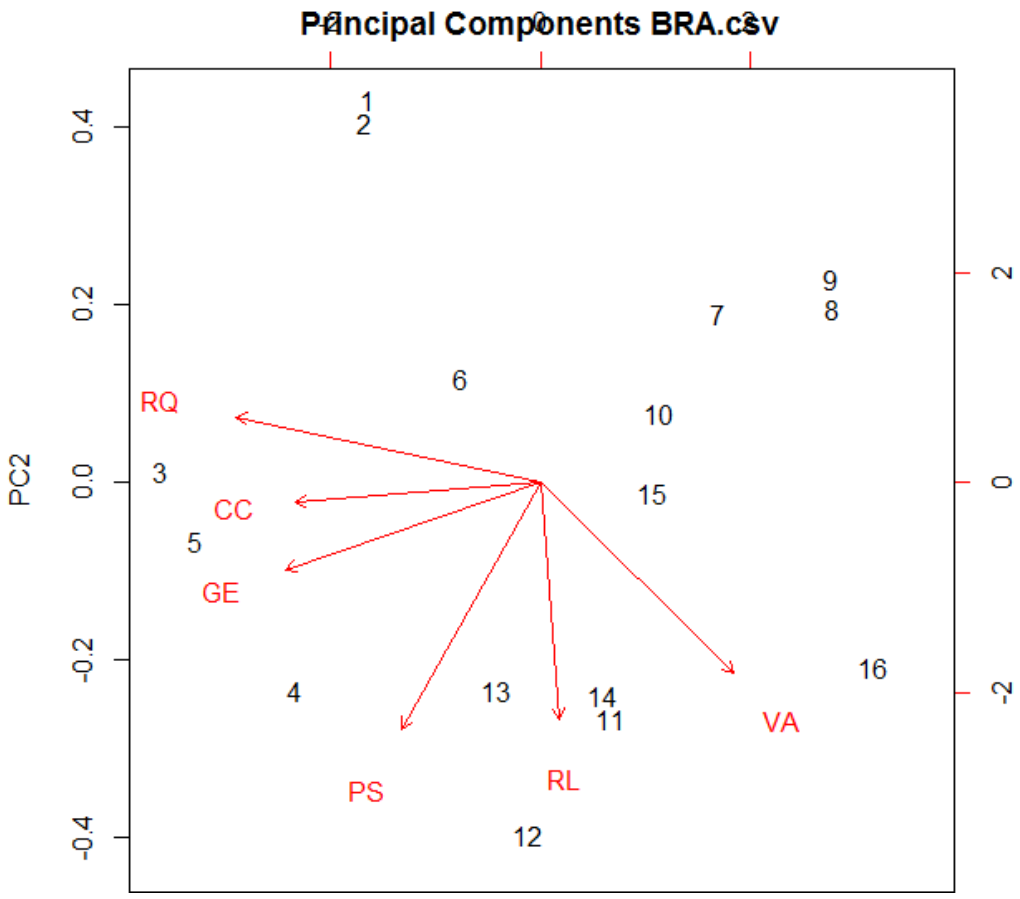
This pairs comparison shows the relationships between each of the variables. Most of the relationships show very little tendency towards any kind of pattern. The most significant column to look at is BRA which shows each variable with respect to time, and the only plot which indicates some semblance of a pattern is VA, which is likely to have a fairly strong positive linear regression. This is shown by the correlation coefficient of .743. Looking at the other correlation coefficients we can learn that RQ has a significant negative linear regression and RL might have a positive linear regression. Looking beyond time relationships we can see a negative linear relationship between VA and RQ but no other significant relationships. The relationship between VA and RQ should be investigated further though.

Correlation plot:



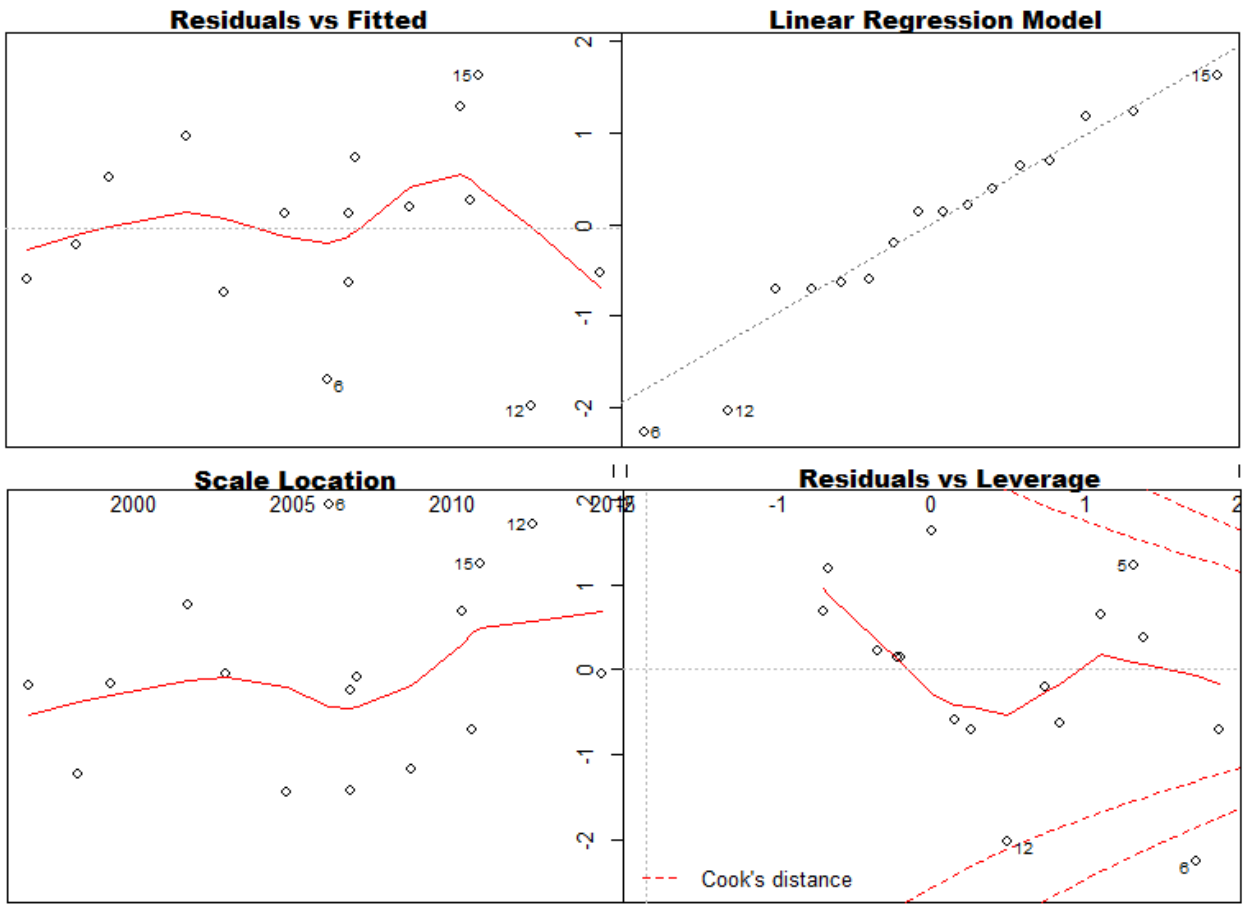
This Pearson correlation plot confirms the significance of the relationship between VA and RQ and also indicates that relationships between RQ and CC and GE and CC might be worth investigating.

Principal components



The Principle components analysis helps us to visualize these relationships further. It's quite clear that RQ and VA point in almost opposite directions while RQ, CC, and GE all point in the same direction.

Linear regression model for BRA with time as the target:



Summary of the Linear Regression model (built using lm):

lm(formula = BRA ~ ., data = crs$dataset[, c(crs$input, crs$target)])

Residuals:

Min	1Q	Median	3Q	Max
-2.4846	-0.7418	0.1936	0.7706	2.1386

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2010.7071	3.9928	503.583	0
VA	9.5720	6.6947	1.430	0.186560

PS	-0.9666	3.0160	-0.321	0.755906
RQ	-18.8258	6.9443	-2.711	0.023959
GE	6.8469	5.9299	1.155	0.277970
RL	19.9174	3.1081	6.408	0.000124
CC	-3.4112	6.2583	-0.545	0.598952

Residual standard error: 1.63 on 9 degrees of freedom

Multiple R-squared: 0.9448, Adjusted R-squared: 0.9081

F-statistic: 25.7 on 6 and 9 DF, p-value: 0.00003541

This linear regression gives us a very reliable model according to the R-squared and adjusted R-squared values which are over 90% and the p-value which is almost 0. The most significant variables in this model are RQ and RL which both have p-values of less than 5%, which is considered statistically significant. These two variables also have the most significant slopes with RQ at -18.8 and RL at 19.9, showing a definite trend for regulatory quality to decrease over time and rule of law to increase over time.

PS and CC show almost no relationship, and VA and GE have relatively weak relationships. If relationships exist for PS and CC, the do not appear to be linear so perhaps transformations should be done to tune these variables.

Creating linear regressions for PS and CC individually shows PS does not exhibit any linear relationship with a p-value of .77, but CC does, with a p-value of .11, so perhaps it's simply that PS does not have a significant impact upon the model, but CC might if we remove PS.

After removing PS from the model, the strength of the model is improved, but CC is not improve significantly. Removing CC also provides a slight improvement, but the combined effect of removing these two variables is not significant enough to merit their removal. Although no information can be gained from these two variables, their inclusion in the model might prove useful for later researchers as to why they seem to have little use and seem to be unpredictable.

Overall, this model suggestions that the government of Brazil's Rule of Law in getting better over time while Regulatory Quality degrades over time and the other variables are not predictable. This may indicate that changes happen often in the areas of Voice and Accountability, Political Stability, Governance Effectiveness, and Control of Corruption, leading to inconsistent scores in these areas. To improve these things the government of Brazil should implement policies which help them to be more consistent and ought to focus heavily on turn around the negative pattern for Regulatory Quality.

==== ANOVA ====

Analysis of Variance Table

Response: BRA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
VA	1	239.333	239.333	90.0435	0.000005527
PS	1	0.715	0.715	0.2691	0.6164433
RQ	1	56.730	56.730	21.3432	0.0012546
GE	1	3.344	3.344	1.2582	0.2910182
RL	1	108.917	108.917	40.9773	0.0001251
CC	1	0.790	0.790	0.2971	0.5989522

Residuals 9 23.922 2.658

The Analysis of Variance for this linear model further confirms our conclusions that RQ and RL are the most significant variable in this model and PS and CC are not significant.

Neural network model for BRA with time as the target:

A neural network model was also produced and will be compared the linear regression in the evaluation stage. The most effective model had 11 nodes, which produced almost 0 residuals.

Summary of the Neural Net model (built using nnet):

A 6-11-1 network with 95 weights.

Inputs: VA, PS, RQ, GE, RL, CC.

Output: BRA.

Sum of Squares Residuals: 0.0578.

Neural Network build options: skip-layer connections; linear output units.

In the following table:

b represents the bias associated with a node

h1 represents hidden layer node 1

i1 represents input node 1 (i.e., input variable 1)

o represents the output node

Weights for node h1:

b->h1 i1->h1 i2->h1 i3->h1 i4->h1 i5->h1 i6->h1

38.46 15.43 -4.14 5.33 -3.56 -10.82 -1.93

Weights for node h2:

b->h2 i1->h2 i2->h2 i3->h2 i4->h2 i5->h2 i6->h2

56.17 21.36 -6.52 8.48 -4.07 -14.64 -3.56

Weights for node h3:

b->h3 i1->h3 i2->h3 i3->h3 i4->h3 i5->h3 i6->h3

-12.76 -4.96 0.53 -2.88 1.07 1.49 1.97

Weights for node h4:

b->h4 i1->h4 i2->h4 i3->h4 i4->h4 i5->h4 i6->h4

-11.32 -4.37 0.97 -2.64 0.92 1.78 0.64

Weights for node h5:

b->h5 i1->h5 i2->h5 i3->h5 i4->h5 i5->h5 i6->h5

59.24 23.11 -6.81 8.80 -4.14 -15.60 -3.04

Weights for node h6:

b->h6 i1->h6 i2->h6 i3->h6 i4->h6 i5->h6 i6->h6

-22.08 -8.21 2.26 -2.73 1.48 5.40 1.51

Weights for node h7:

b->h7 i1->h7 i2->h7 i3->h7 i4->h7 i5->h7 i6->h7

-33.71 -12.26 4.35 -4.22 2.54 8.18 1.33

Weights for node h8:

b->h8 i1->h8 i2->h8 i3->h8 i4->h8 i5->h8 i6->h8

4.26 1.24 0.10 -3.65 3.67 -17.08 5.42

Weights for node h9:

b->h9 i1->h9 i2->h9 i3->h9 i4->h9 i5->h9 i6->h9

2.08 -5.37 0.75 -4.74 0.20 1.70 1.95

Weights for node h10:

b->h10 i1->h10 i2->h10 i3->h10 i4->h10 i5->h10 i6->h10

-66.36 -25.56 8.23 -10.07 4.41 17.12 4.16

Weights for node h11:

b->h11 i1->h11 i2->h11 i3->h11 i4->h11 i5->h11 i6->h11

16.42 6.38 -2.25 1.80 -1.02 -5.05 -1.10

Weights for node o:

b->o h1->o h2->o h3->o h4->o h5->o h6->o h7->o h8->o h9->o h10->o h11->o

508.24 187.57 269.62 309.55 244.12 332.60 235.61 243.74 319.81 269.30 262.84 183.77

i1->o i2->o i3->o i4->o i5->o i6->o

251.62 -34.41 195.67-14.22 -45.36 -91.14

(the neuralnet and RSNNS packages don't work for me in either Rstudio or R 3.2.3, so I cannot actually produce a visualization of this neural network, which is very disappointing to me because I was greatly looking forward to presenting and interpreting it)

Generally, this neural network describes the connections between every variable and every other variable and their weight on the conclusion. Overall, variables 1 (VA) and 5 (RL) have the greatest weights, and sometimes 3 (RQ) has a heavier weight (positive) than 2 (PS) and 4 (GE), which tend to have almost no weight. This shows the same results at the linear regression with the exception that VA may have more of an effect than indicated by the linear regression. In fact, the final results of the inputs on the output are that VA (positive) has the most effect overall, with RL (positive) having a little less effect and RQ actually having almost no effect along with PS, GE, and CC (all negative).

This may be indicating that the effectiveness of Brazil's government is increasing over time and that the main thrust of this growth is through Voice and Accountability and Rule of Law. If this is the case than the Brazilian government should focus its efforts on the other areas of government so as to strengthen its weakest areas while maintaining these strengths.

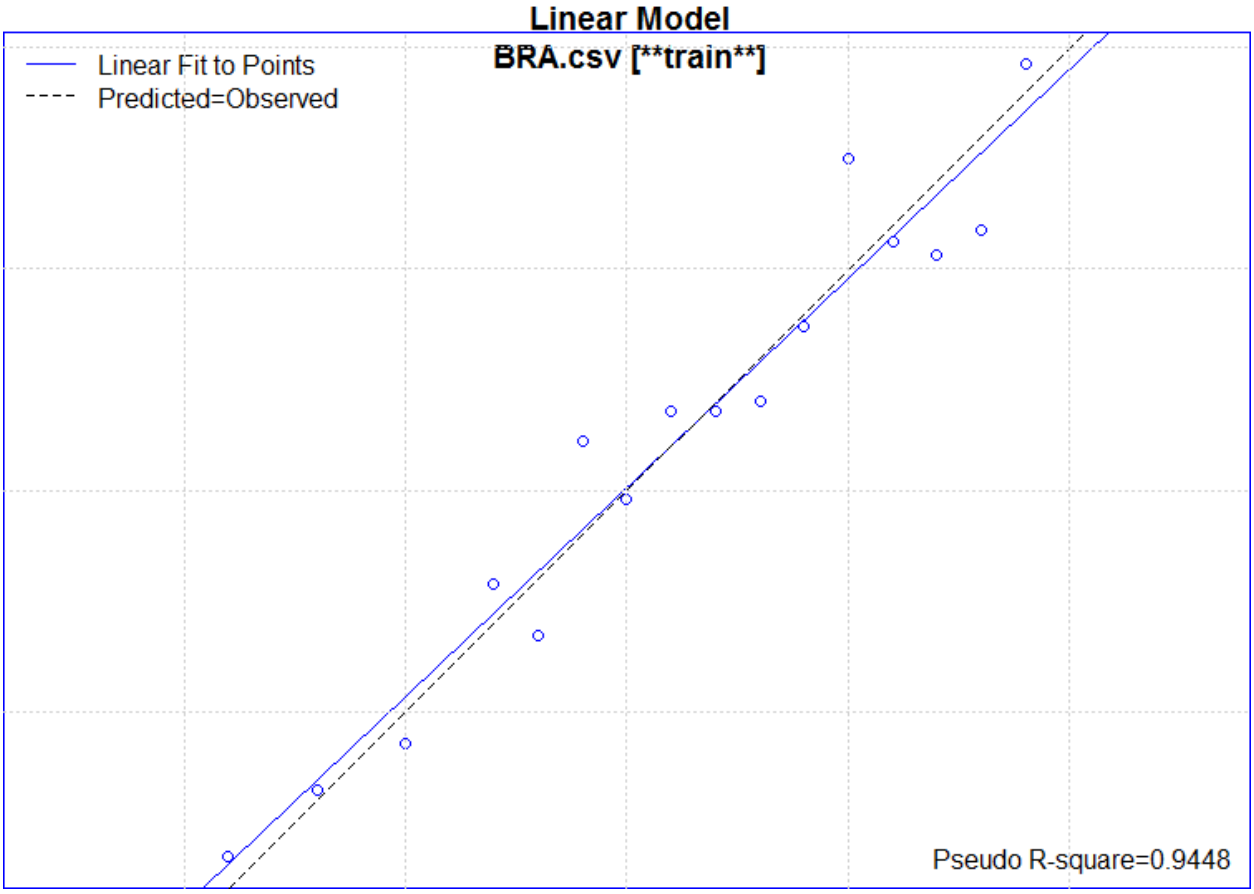
Model diagnostics and evaluation

For linear and neural net models, I have used the Predicted vs Observed function to help visualize and statistical analyse the effectiveness of the two models produced. This is the only

function available in Rattle to analyze linear regressions, but it's a highly effective method so that's not a problem. The main mechanic of this method is to examine the residuals between the model's predictions and the observed values.

It's important to note once again that the dataset was too small to split into validation, testing, and training sets, so these results represent the whole dataset.

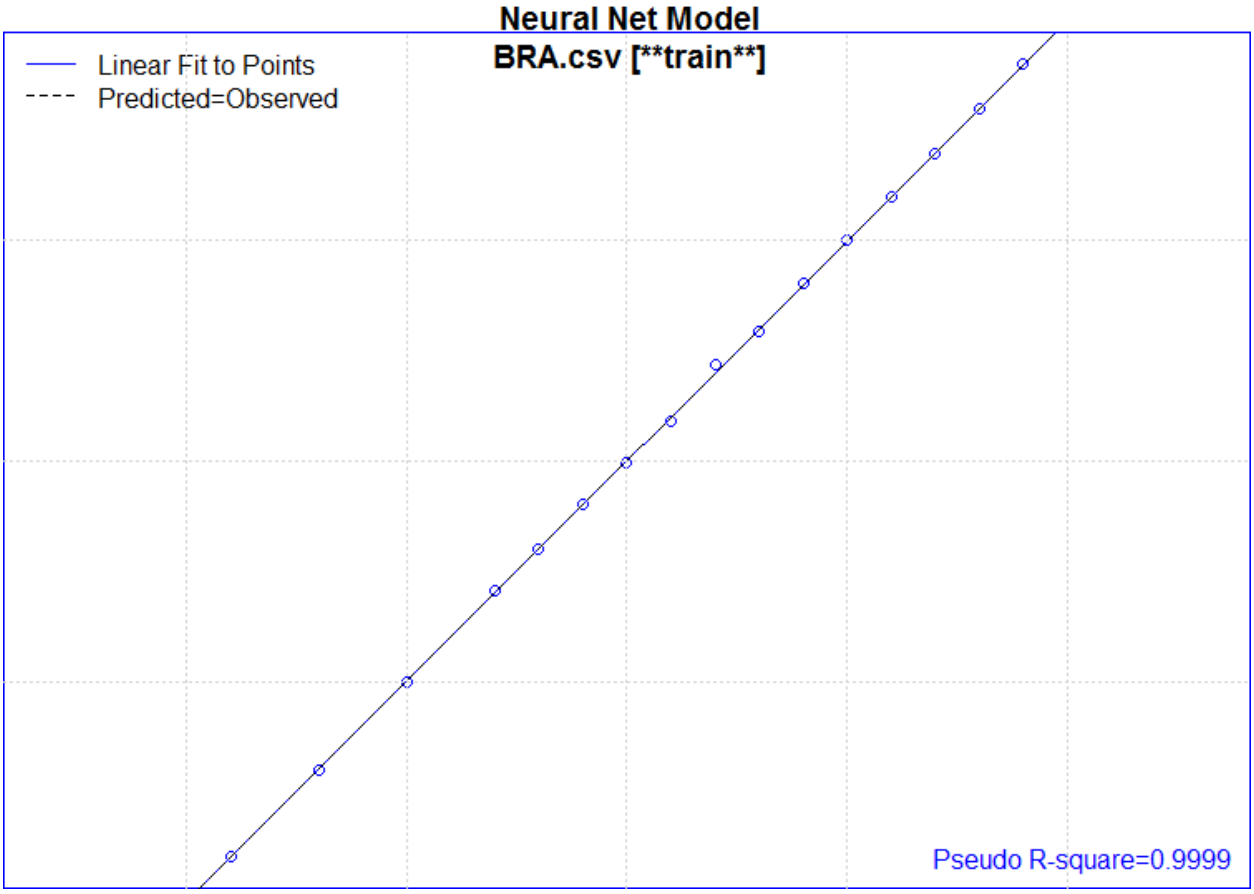
Linear regression model evaluation via Predicted vs Observed comparison:



This Predicted vs Observed evaluation shows us that the linear model is very closely aligned to the data as best can be seen by this 2D representation. The Pseudo R-squared value is over 94%, so we can expect future data, the Brazilian government continues employing similar methods and policies, is very likely to fit this model. For the purpose of this model, it is also

useful to know that if new data is significantly different from the model then we have a very accurate model to compare it against to see whether positive or negative changes have taken place.

Neural network model evaluation via Predicted vs Observed comprison:



This Predicted vs Observed comparison actually shows an almost perfect fit for the 11 node neural network. This indicates that the neural network model is actually more effective than the linear regression although the linear regression is easier to visualize and may be easier to interpret. Since this model is almost perfectly accurate we should be able to use it to very accurately tell if new data still fits the pattern or if changes have been implemented.

Scoring and summary

Finally, a scoring engine is produced for both of these models so that they can use for future data which is collected for Brazil. This model can be used to see if Brazil continues the same trends, becomes more stable, becomes less stable, or takes actions which benefit or hurt their government's effectiveness. Hopefully this current model's results will be used to encourage implementation of policies which stabilize those variables which provide little information and policies which reverse the negative trends in variables which are consistently degrading.

Works cited:

The World Bank Group. (2015). *World Governance Indicators*. The World Bank Group.

Retrieved from: http://info.worldbank.org/governance/wgi/index.aspx#home