IIMB

तेजस्वि नावधीतमस्तु

Statistical Analysis of Corruption Perception Index across countries

AMDA Project – Summary Report

(Under the guidance of Prof Malay Bhattacharya)

Group 3

Anit Suri 1511007

Avishek Biswas 1511013

Diwakar Jain 1511019

Gaurav Bharti 1511245

Contents

1. MOTIVATION

Corruption is one of the most burning issues in recent times. While countries like Norway, Sweden, and Denmark etc. rank favorably in terms of corruption perception, the same doesn't apply in countries like India, Bangladesh, Venezuela, and Zimbabwe etc. India was ranked an unimpressive 76th out of 168 countries in the latest Corruption Perception Index published by Transparency Internationali. Governments have been for long trying to grapple with the menace of corruption and have been struggling to find out controllable levers that reduce the level of corruption but the solution is still far from being achieved. While intuitively people can think of certain specific factors that seem to shape the perception of corruption in the countries, however, it makes sense to delve deeply into the subject and carry out a mathematical analysis that reveals the underlying factors leading to the same which will help in unveiling the obscurity around the causes of corruption and hopefully provide a solution to this problem.

2. OBJECTIVE

- ➔ Identifying what are the significant variables impacting the corruption perception of the country
- ➔ Assessing the degree to which (relative importance) these variables affect the corruption perceptions
- ➔ Interpreting the variables in terms of the underlying unobservable factors
- ➔ Developing a model to classify any country with a given set of variables into a distinct corruption perception bucket
- ➔ Understanding which of the significant variables obtained are controllable in nature and how the underlying factors (contributing to those variables) can be modified through appropriate steps to change the corruption perception of the nation

3. DATA COLLECTION & CLEANING

DATA COLLECTION

- ➔ We used the latest Corruption Perception Index (CPI) published by Transparency International as a dummy indicator for the corruption in a country.
- ➔ For the predictor variables, we decided to gather information on economic, political, social and education-related variables in order to capture a wide variety of information which can have impact on corruption. These data were collected from reliable sources like World Bank, UNICEF and Freedom House

DATA CLEANING

- ➔ Post collection of data, we checked for data quality (missing variable values). It was noted that for some of the variables, data was not available for some countries.

- ➔ For the variables with a lot of missing values, we removed the variable from the data set.
- ➔ After this, we checked for countries which don’t have sufficient number of variables populated with values. We removed those countries from our dataset.
- ➔ Finally, we checked for outliers in the data. We decided to consider all the values beyond 3σ as outliers for each of the variables and decided to remove the corresponding rows from our dataset.
- ➔ After this, we were left with 142 data points.

4. EXPLORATORY ANALYSIS

In order to build a robust regression model, it was first of all important to understand the relationship between CPI and some of the key variables considered. This was required so as to logically include/not include certain variables in the process as the explanatory power of some of the variables could have been captured in other variables too. The variables considered being included in the regression model and the intuitive rationale for the same is given below.

VARIABLES	**RATIONALE**	**NATURE OF VARIABLE**	**VARIABLE NAME IN THE MODEL**
CPI	Indication of countries’ corruption perception, higher value indicates better corruption perception	-	*CPI10*
GINI	Measure of social inequality, more the GINI index for a country, higher should be the CPI	Social	*GINI*
Literacy Rate	More literate the population, better is the corruption perception (higher CPI)	Social	*Lit_rate*
Unemployment Rate (%)	Higher the unemployment rate, more dissatisfied are people and more are the chances that they will indulge in corrupt practices to fulfill their needs and wants	Economic	*Un_rate*
Percentage Below Poverty Line	Indication of the living standard of people. Poorer the people, more is the tendency to indulge in corruption related acts and higher should be the CPI	Economic	*Per_BPL*
Urban Population (% of total population)	Measure of living standard, awareness & opportunities available to people. If the % of urban population is high, higher should be the CPI	Social	*Urb_pop*
Constitutional	Govt. form in a country shapes the corruption	Political	*Is_Republic*

Form	perception. If govt. is republic, CPI should be low; if it is constitutional monarchy then CPI should be high		
Government Expenditure	Higher the government expenditure on social schemes, lower should be the CPI	Political	*Gov_exp*
GDP per capita	This is a measure of the disposable income in the hands of the public, more is the GDP per capita lesser should be the corruption perception	Economic	*GDP_CK*
Regulatory Quality	Measure of the ability of govt. to formulate and implement policies for private sector development. Better regulatory quality should lead to higher CPI	Political	*Reg_qlt*
% of govt. expense on education	Indicates the level of education in a country; higher education expense should lead to higher CPI	Political	*Exp_edu*
Political stability Estimate	More stable the political climate, higher should be the CPI as govt. is perceived to be stronger and stable	Political	*Pol_stab*
Press Freedom Index	Indication of the ease with which information is available to the public. Better the information access, higher should be the CPI	Social	*PFI*
GDP	Indication of the purchasing power & production capacity of the economy; Higher GDP should lead to higher CPI	Economic	*GDP*
Mobile subscriptions (per 100 users)	Indication of the ease of information access as well as the living standard of people; higher value should lead to higher CPI	Social	*Mob_sub*

5. MODEL BUILDING

5.1 CORRELATION COEFFICIENTS

Prior to initiation of the model building process a broad level analysis was carried out for all the above mentioned variables by computing their correlation coefficients with the Corruption Perception Index (CPI). This was done to ensure that the intuitive reason for incorporating the above variables was justified. Please refer to Table 1 in Detailed Analysis File for correlation coefficient values.

5.2 ANALYSIS

For the analysis of the impact of the selected variables on the Corruption Perception Index (CPI), first of all a basic linear regression model was developed.

The process included formulation of a few models through appropriate transformations, checking the model fit through relevant diagnostic measures. After these steps, the final model was arrived at. Please refer to Detailed Analysis file (Section – **Model Building**) for all the above steps.

5.3 FINAL REGRESSION MODEL

The final regression equation obtained is-

$$CPI10 = 6.312 + 0.74 * Pol_{stab} + 0.742 * \ln\left(Urb_{pop}\right) - 1.225 * \ln(PFI) - 0.018 * per\_BPL + 0.048 * \text{Gov\_exp} - 0.038 * Un\_rate + error$$

From the Table 8, we can see that the variables Pol_stab and ln(PFI) are the ones having the highest impact on the corruption perception of any country. The rest of the variables have an almost equal impact on CPI.

DIAGNOSTICS

- ➔ F statistic value is 68.153 and shows that overall model is significant (Table 9)
- ➔ Adjusted R-square is 0.772 and shows sufficient explanatory power (Table 10)
- ➔ Durbin-Watson test stat. of 2.102 shows absence of error auto-correlation (Table 10)
- ➔ Error normality is satisfied (Figure 1)
- ➔ White's test shows p-value of 48.9% indicating absence of heteroscedasticity (Table 11)

The table shown below lists the various significant variables in decreasing order of the impact (absolute impact) on the CPI.

SUMMARY TABLE

As per our regression model, below are the major variables and their corresponding impact on the corruption perception of a country-

VARIABLES	NATURE OF IMPACT[a]
Pol_stab	Higher political stability leads to improved corruption perception
ln(PFI)	Greater freedom of press leads to improved corruption perception
ln(Urb_pop)	Higher proportion of urban population leads to improved corruption perception
Per_BPL	Higher % of people below poverty line leads to poor corruption perception

Gov_exp	Higher government expenditure leads to improved corruption perception
Un_rate	Higher unemployment rate leads to poor corruption perception

a: Keeping all other variables constant

INFERENCE

- Higher political stability helps in improving the corruption situation and perception in the country. Political stability provides time and opportunity to the govt. for focusing on policies and implementing the same which can help in reducing corruption in the country.
- Freedom of press is a critical factor in improving the corruption perception of a country. As the freedom of press increases, this leads to an increased ability of the press to spread information among the public leading to rising awareness. As awareness increases, people are less likely to indulge in corruption acts which would improve the corruption perception of the country.
- Urban population proportion is a reflection of the availability of resources and opportunities which in turn leads to higher satisfaction in life (higher standard of living). The more people are satisfied, the lesser are their chances to get involved in corruption-related acts.
- The percentage of population living below poverty line is an indication of the quality of life and the disposable income in the hands of the public. Thus, as the number people living below poverty line increases, the tendency to indulge in corruption increases and hence the corruption perception of the country deteriorates.
- Higher government expenditure is an important factor affecting corruption perception. The more is the government expenditure the better is the reinforcement of people’s faith in their government leading to an improvement in corruption perception.
- A higher unemployment rate is indicative of poorer economic condition of the people due to which they are not able to satisfy their daily needs and wants. Thus, there is a tendency to take part in corrupt activities. Moreover, unemployment rate is also indicative of the lower education level in the country. Lower education leads to lower awareness which is one of the reasons behind corruption and hence leads to poorer corruption perception.

5.4 MODEL VALIDATION

The regression model was built considering data from 120 countries. The remaining set of 23 countries was chosen as the test data which was used to validate the regression model formulated. The MAPE value (Mean Absolute Percentage Error) was computed on the test data which came out to 9.82% (Refer to data set file – **MAPE_Test** already uploaded). Consequently, we can infer that the model is able to track the CPI fairly accurately.

6. FACTOR ANALYSIS

Based on the significant variables identified from Multiple Linear Regression model developed earlier, a Factor Analysis was performed on the dataset.

DIAGNOSTICS

- ➔ KMO value is 0.710 which is greater than the acceptable value of 0.6. (Table 12)
- ➔ Bartlett's Test statistic is 199.015 showing that performing a makes sense as the variables are correlated (Table 12)
- ➔ Residual correlation matrix shows that approximately 75% of the number of off-diagonal elements are less than 0.1 showing a proper capture of variable variance (Table 13)
- ➔ All the diagonal elements of anti-image matrix (Table 14) have values greater than 0.5 which suggests adequacy of samples (meaning sample size sufficient to run factor analysis)
- ➔ Communality table (Table 15) shows that at least 70% of the variance has been captured for all the variables through the 3 factors extracted
- ➔ Table 16 shows that in cumulative terms, 77% of the total variance of the variables is explained through the 3 factors
- ➔ Scree plot (Figure 2) shows sufficient explanatory power captured through 3 factors (elbow at component number 3)

COMPONENT EXTRACTION

- ➔ The original (un-rotated) component matrix (Table 17) shows almost equal loadings on the factors for some of the variables leading to unclear interpretation of the factors
- ➔ VARIMAX rotation was carried out leading to higher loadings for all the variables on one of the factors (Table 18)

INTERPRETATION

- ➔ *The variables Pol_stab (Political Stability), Gov_exp (Government Expenditure) and ln(PFI) [Press Freedom Index] have a significant proportion of loading on Factor 1.*

  These can be considered to be part of a factor which we will name as "**Political and Democratic Climate**". This factor encompasses stability of government, acceptance of government policies by people, level of freedom available to people & institutions in the nation and the degree to which democracy is present in the nation.

- ➔ *The variables per_BPL (% of people below poverty line) and ln(Urb_pop) [Urban Population%] have a significant proportion of loading on Factor 2.*

These can be considered to be part of a factor which we will name as "**Living Standard**". This factor encompasses the economic condition of people (from %BPL), level of information & awareness among people (we are using % of urban population as a dummy for this).

➔ *The variable Un_Rate (Unemployment Rate) is mostly captured by Factor 3.*

We will consider a 3rd factor for this namely "**Skill development and availability of opportunity**". Unemployment rate gives an idea about the skill level of general population, level of education and the employment opportunities available in the country.

CONCLUSION

From factor analysis, it can be clearly seen that the 3 main factors impacting corruption perception of the countries span across political, economic and social spheres.

Factor 1 (**Political and Democratic Climate)** is political in nature and lends itself as a controllable lever to improve the corruption perception of the country. By maintaining stability in the political scenario and respecting the democratic rights of citizens, governments can send a strong signal regarding political inclusiveness thereby garnering public faith and shaping an improved corruption perception.

Factor 2 (**Living Standard**) is economic in nature. Governments can improve the corruption perception of their countries by guaranteeing a satisfied lifestyle through efforts directed at poverty reduction and increasing the living standard of people through urbanization.

Factor 3 (**Skill development and availability of opportunity**) is social in nature. Through focused efforts aimed at skill development and increasing education level, governments can increase the employment opportunities for the public which can help in reducing their tendency to indulge in corruption related activities.

7. CLUSTERING

7.1 INITIAL ANALYSIS – DETERMINATION OF NUMBER OF CLUSTERS

For clustering, Hierarchical method of clustering (Agglomerative) was used because of the number of data nodes was not significantly large for using non-hierarchical clustering techniques like k-means to be used. Initially, clustering was run to determine the number of clusters through Dendrogram plot. For calculating similarity, Ward's linkage and Squared Euclidean Distance were used.

As can be inferred from the Dendrogram (Figure 3), the optimal number of clusters in this case would be **3**.

7.2 CLUSTERING

The clustering analysis was again run using **Ward's linkage** and **squared Euclidean distance** with cut-off set at **3 clusters**. We used 120 data nodes for clustering. The main objective of clustering was to use these clusters in the Discriminant Analysis. For that it was necessary to have good separations between the clusters. Therefore, we used the dependent variable – Corruption Perception Index (*CPI*) to create clusters. Below are the results of the clustering analysis.

For agglomeration, Ward's Linkage was used and the similarity calculation was done using Squared Euclidean Distance. Ward's linkage maximizes the significance of the difference between clusters and therefore has the highest statistical power among all methods. But since, it is prone to outliers we treated the data first to remove or modify outliers.

7.3 CLUSTER STATISTICS

Below are the statistics of each cluster. As can be seen from the table the separation between each cluster has been very good. There is no overlapping in the range of each cluster and the standard deviation for each cluster is also less. The range is representative of the CPI of the country. Therefore, Cluster 1 represents the countries which have a lower CPI, Cluster 2 represents the countries having medium CPI and Cluster 3 represents countries with High CPI.

Hence, we get three clusters, **Low CPI countries, Medium CPI countries** and **High CPI countries** with 72, 32 and 16 nodes in them respectively.

	Cluster 1	Cluster 2	Cluster 3
Min	12	44	74
Max	42	71	91
Mean	30	56	83
SD	7	9	5
Count	72	32	16

8. DISCRIMINANT ANALYSIS

8.1 TEST OF EQUALITY OF GROUP MEANS

Initial: Initially we used all the variables that we had obtained in the Regression model for discriminant analysis. However, on the basis of level of significance we rejected the variable *Un_Rate* (Unemployment Rate) (Table 20).

Final: After removing the variable *Un_Rate* we see that all the five variables are significant and can be used for calculating the discriminant score (Table 21).

8.2 SUMMARY OF CANONICAL DISCRIMINANT FUNCTIONS

From the eigenvalues it is clear that the first discriminant function is capturing the highest variance and the second discriminant function is not as powerful as the first one (Table 23). The reason for this is that the number of nodes in one of the clusters is a significant percentage (60%) of the total number of data nodes. The first discriminant function discriminates the first cluster from the other two clusters.

Again, the Wilk's Lambda for the first function is closer to zero as compared to the second one because of higher number of data nodes in the first cluster. In the other two clusters, the number of data nodes is very small and therefore, the within group variance tends to be higher (Table 24).

8.3 DISCRIMINANT FUNCTION OUTPUT

The equations for the discriminant functions can be obtained using the standardized canonical discriminant function coefficients (Table 25). These can be used to calculate the discriminant scores. In this case, the scores can be calculated using the below equations,

$$\begin{aligned} Score_1 = \; & -0.213 * perBPL + 0.288 * GovExp + 0.341 * \ln(UrbPop) + 0.486 * PolStab \\ & - 0.481 * \ln(PFI) \end{aligned}$$

$$\begin{aligned} Score_2 = \; & -0.233 * perBPL + 0.567 * GovExp - 0.045 * \ln(UrbPop) + 0.535 * PolStab \\ & + 0.966 * \ln(PFI) \end{aligned}$$

8.4 PREDICTED CLASSIFICATIONS

After obtaining the discriminant functions, discriminant score is calculated for each node and it is classified using that score into different clusters. It can be seen from the classification table (Table 29) that the predictions of the discriminant function show 80.0% accuracy while classifying the countries according to the given set of variables. Using these functions, we can check in which cluster/class a particular country would lie. Now, by controlling the controllable levers like Government Expenditure (*Gov_Exp*) and Press Freedom Index (*ln(PFI)*), governments can change the corruption perception of the country as now they can analyze how much of change in those levers is required.

9. CONCLUSION & RECOMMENDATIONS

- ➔ The major variables (obtained through regression) affecting the corruption perception of countries are – **political stability**, **freedom of press**, **percentage of population below poverty line**, **urban population as a proportion of total population**, **government expenditure** and **unemployment rate**.

➔ Based on the above, it can be inferred that corruption perception is a result of elements that span across political, economic and social dimensions.

➔ The major underlying unobservable factors (obtained through factor analysis) shaping up the above variables leading to corruption perception are – **political & democratic climate** (*political* in nature), **living standard** (*economic* in nature) and **skill development & availability of opportunity** (*social* in nature).

➔ From the above analysis and discussion, it can be inferred that there are certain levers that are under the control of governments which can be altered to achieve the desired level of corruption perception.

Finally, it is recommended that governments across the world should focus on the following in order to improve their countries' corruption perception –

1. Try to maintain a stable political environment for the smooth functioning of the country. This will increase public faith and acceptance of government policies thereby reducing the tendency to indulge in corrupt practices.

2. Provide sufficient freedom to the press and media. This will enable the media to spread information and awareness among people without any prejudice in case anybody tries to indulge in corrupt acts. This will discourage people from engaging in corrupt acts.

3. Create sufficient employment opportunities through public spending and skill development. This will help reduce unemployment rates and enable public to meet their needs without depending on corruption activities.

[i] http://www.transparency.org/cpi2015?gclid=CJyF4KTesdACFdSJaAodoqQFhg