GLOBAL CORRUPTION INDEX 2021

BY GRP

Technical Methodology

Global Corruption Index 2021
Technical Methodology

The Global Corruption Index (GCI) is a composite index offering a classification of exogenous risks by country with regards to anti-corruption and anti-white collar crimes. The GCI was built in the framework of a more general risk mitigation plan, compliant with international recommendations and current binding legal requirements, such as the US FCPA, the UK Bribery act and the French law “Sapin II”.

This technical methodology is meant to provide all necessary information for understanding the variables used and their respective impact in rankings.

1- Indicators and Data Overview

The GCI covers 196 countries & territories and is composed of 43 variables constructed based on datasets that are exclusively borrowed from internationally recognized entities. Among these variables, 30 are for the calculation of a sub-index, the Corruption Index (CI), while the remaining 13 allow for a White Collar Crimes (WCC) sub-index calculation. As can be seen from tables 1 and 2 below, the Corruption Index is divided into 4 indicators: (1) Ratification Status of Conventions, (2) Corruption Perception, (3) Corruption Experience and (4) Country Characteristics.

Table 1 Levels of aggregation

<table>
<tr><td>Index</td><td colspan="43">Global Corruption Index</td></tr>
<tr><td>Weight</td><td colspan="30">0,7</td><td colspan="13">0,3</td></tr>
<tr><td>Sub-index</td><td colspan="30">Corruption Index</td><td colspan="13">WCC Index</td></tr>
<tr><td>Weight</td><td colspan="2">0,15</td><td colspan="28">0,85</td><td>0,4</td><td colspan="12">0,6</td></tr>
<tr><td>Indicator</td><td colspan="2"></td><td colspan="28">Corruption Evaluation</td><td></td><td colspan="12">WCC Evaluation</td></tr>
<tr><td>Weight</td><td colspan="2"></td><td colspan="3">0,3</td><td colspan="3">0,2</td><td colspan="22">0,5</td><td></td><td>0,5</td><td colspan="11">0,5</td></tr>
<tr><td>Indicator</td><td colspan="2"></td><td colspan="3"></td><td colspan="3"></td><td colspan="22">Country Characteristics</td><td></td><td></td><td colspan="11"></td></tr>
<tr><td>Weight</td><td colspan="2"></td><td colspan="3"></td><td colspan="3"></td><td colspan="4">0,2</td><td colspan="9">0,3</td><td colspan="7">0,3</td><td colspan="2">0,2</td><td></td><td></td><td>^(1/3)</td><td colspan="10">^3</td></tr>
<tr><td>Indicator</td><td colspan="2">Ratification status of Conventions</td><td colspan="3">Corruption Perception</td><td colspan="3">Corruption Experience</td><td colspan="4">Citizen's Voice and Transparency</td><td colspan="9">Government Functoning and Effectiveness</td><td colspan="7">Legal Context</td><td colspan="2">Political Context</td><td>Basel AML Index</td><td>Anti-Money Laundering</td><td>Export of Insurance and Financial Services</td><td colspan="10">WCC Standards</td></tr>
<tr><td>Weight</td><td>0,9</td><td>0,1</td><td>0,5</td><td>0,25</td><td>0,25</td><td>0,35</td><td>0,35</td><td>0,3</td><td>0,31</td><td>0,23</td><td>0,23</td><td>0,23</td><td>0,2</td><td>0,07</td><td>0,09</td><td>0,09</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,125</td><td>0,125</td><td>0,4</td><td>0,08</td><td>0,08</td><td>0,08</td><td>0,08</td><td>0,08</td><td>0,2</td><td>0,7</td><td>0,3</td><td>1</td><td>1</td><td>1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td><td>0,1</td></tr>
<tr><td>Variable Ref.</td><td>V1</td><td>V2</td><td>V3</td><td>V4</td><td>V5</td><td>V6</td><td>V7</td><td>V8</td><td>V9</td><td>V10</td><td>V11</td><td>V12</td><td>V13</td><td>V14</td><td>V15</td><td>V16</td><td>V17</td><td>V18</td><td>V19</td><td>V20</td><td>V21</td><td>V22</td><td>V23</td><td>V24</td><td>V25</td><td>V26</td><td>V27</td><td>V28</td><td>V29</td><td>V30</td><td>V31</td><td>V32</td><td>V33</td><td>V34</td><td>V35</td><td>V36</td><td>V37</td><td>V38</td><td>V39</td><td>V40</td><td>V41</td><td>V42</td><td>V43</td></tr>
</table>

The following table lists the data variables used to calculate the GCI, together with their respective variable reference code.

Table 2 Data Overview

Ref.	Variable
V1	Ratification status of the UN Convention against Corruption
V2	Ratification status of the OECD Anti-Bribery Convention
V3	Absence of Corruption
V4	Corruption Perception Index
V5	Control of Corruption
V6	Bribery Incidence
V7	Bribery Depth
V8	Bribery Rate
V9	Open Government
V10	Voice and Accountability
V11	Public Participation in the Budget Process
V12	Provision of Budget Information
V13	Rule of Law
V14	Regulatory Quality
V15	Regulatory Enforcement
V16	Law Enforcement
V17	Government Effectiveness
V18	Political Leadership and Governance
V19	Constraints on Government Powers
V20	Budget Oversight
V21	Government Transparency
V22	Weakest Dimension of Justice
V23	Civil Justice is Free of Corruption
V24	Criminal System is Free of Corruption
V25	Criminal System is Impartial
V26	Civil Justice is Free of Improper Government Influence
V27	Criminal System is Free of Improper Government Influence
V28	Judicial System and Detention
V29	Democracy Score
V30	Political Stability
V31	Basel AML Index
V32	Anti-Money Laundering
V33	Size of Insurance and Financial Services
V34	Banking Secrecy
V35	Know your Customer / Partner
V36	Reporting
V37	Ownership Transparency
V38	Corporate Transparency
V39	Appropriate Gathering and Use of Information
V40	Appropriate monitoring
V41	Coordinated Actions to Combat Money Laundering
V42	International Cooperation
V43	Sanctions and Debarments

A number of criteria were considered during the selection process:

- All data variables are linked to the measure of corruption and / or white collar crimes, either directly or indirectly. The direct measures of the Corruption Index consist of three surveys of perception and three surveys of reported experience, while the remaining variables are indirect measures meant to capture prevention mechanisms, related effects, causal effects and consequential effects. The objective of the latter group of measures is to unearth the latent information z (corruption) common to all selected variables

- In order to ensure cross-country comparability, no country specific information is considered. Such data would generate valuations relying on different bases / concepts, which is unsuitable for rankings

- Data sources with limited coverage are set aside. The lowest coverage rate considered is that of the variable V8, *Bribery Rate* (54% of country coverage)[1]

- Although some variables have wider coverage than others, none of them are limited to a specific cluster of countries. This decision is meant to guarantee that each variable offers a data set with scattered points across the full spectrum of possibilities

- During the selection process, preference is given to quantitative type of data. Qualitative information is also considered if and only if the transitivity axiom is ultimately satisfied

[1] The coverage rate considered is that obtained after the imputation processes based on real values, that is the linear extrapolation from the five available years and / or the LOCF, if applicable.

2- Missing Data

Several methods exist to deal with missing values, which can be grouped into two types of treatments: deletion - such as listwise deletion (complete-case analysis) and simple case deletion - or imputation.
Considering that most of our missing data is either of type MAR (Missing at Random) or MNAR (Missing not at Random), deletion is hardly appropriate and would lead to biased estimates. As the aim is to estimate corruption and white collar crimes, the lack of transparency for example is an important information. In this consideration, skipping countries for them to have a lack of due reporting would be counterproductive.

The processing of missing data is thus handled on a case-by-case basis depending on the structure of the datasets.

2.1- Imputation

First, in the case of time series datasets with visible trends, we proceed with a linear extrapolation from the five last available years. This method allows to estimating parameters based on real past values.

The second approach used is the method of the Last Observation Carried Forward (LOCF), which is a common statistical approach for time series data that consists in imputing the last available observation. Similar to the first method, only the last five available years are considered.

The two above-mentioned methods are selected to be based on known values that are specific to the countries, and consequently true at a point in time. In most cases, such methods can't be applied because no current nor past value is available. In these cases, we consider a third imputation method: Predictive Mean Matching (PMM) with multiple imputation.

Single imputation provides only one parameter estimate for each missing value and omits possible alternatives. It therefore tends to underestimate the standard errors and consequently overestimate the validity of the estimated scoring. As opposed to single imputation, multiple imputation provides *n* different possibilities for each missing value. These *n* possibilities allow for two desirable outputs:

- First, each imputed value results from the pooling of the *n* parameter estimates, thus providing a better approximation of the true value
- Second and more importantly, multiple imputation allows for measures of uncertainty, by sampling *n* times from the posterior predictive distribution.

As previously mentioned, the selected method of multiple imputations is that of Predictive Mean Matching (PMM). This approach allows us to preserve the distributions in the data and ensures that imputed values are plausible as it fills in values from real observations (Vink et al., 2014[2]). PMM provides a random value from a donor, based on the closeness of the regression-predicted values of the donor $\hat{\beta}$, with that of the recipient β^*. This implies that linear regressions are not used to generate imputed values but rather to determine the donor (Schenker, N. & Taylor, J.M.G., 1996[3]).

The process by which PMM is performed is as follows (Vink et al., 2014[4]):

1. First, an Ordinary Least Squares (OLS) linear regression of γ given the selected predictors χ is performed to obtain the parameter estimates $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{\varepsilon}$, respectively the regression coefficient, the variance and the random error
2. In a second step, random draws of σ^{2*} and β^* are performed based on the posterior predictive distributions to provide new sets of coefficients. These draws allow for the calculation of $\hat{\gamma}_{missing}$
3. Predicted values are then generated by calculating $\hat{\gamma}$ for both cases with values (potential donors) and missing values (recipients), using the parameter estimates $\hat{\beta}$ and β^* respectively
4. The closeness of predicted values between donors and recipients is evaluated, so as to identify the three cases which minimizes $|\hat{\gamma}_{observed} - \hat{\gamma}_{missing}|$
5. Missing values are substituted from a random donor among those that satisfy the minimization criteria of the previous step.

---

[2] Vink, G., Frank, L. E., Pannekoek, J., and van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica. 68(1). 61-90

[3] Schenker, N., & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. Computational Statistics & Data Analysis, 22(4), 425–446

[4] Vink, G., Frank, L. E., Pannekoek, J., and van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica. 68(1). 61-90

6. Considering this index uses PMM for multiple imputation, the process starting from the random draws of σ^{2*} and β^{*} to the final imputation is repeated *n* times, in order to provide *n* complete datasets with *n* possible values for each missing case

2.2- Case Deletion

For some variables, no PMM imputation was performed and only true values were considered in the analysis. This is due to the structure of the data and the absence of correlation with other variables. In the case of a missing value, the algorithm proportionally redistributes the according weight to variables measuring the same indicator.

3- Standardization

Aside from binary variables, all datasets were tested for skewness, then transformed and recoded if necessary. The mean and standard deviation is calculated and all variables are then standardized, to allow for a proper aggregation in the global scoring. Several normalization methods exist. The one used here is that of z-scores, which converts datasets to a common scale with a mean of zero and a standard deviation of one, as follows:

$$I_{i,c} = \frac{X_{i,c} - X_{i,\bar{c}}}{\sigma_{\bar{c}}}$$

with:
ι *= variable*
c = country
$\bar{c}$ *= reference country*
σ *= standard deviation*

4- Aggregation

The aggregation process converts all data points to a scale of 0-100, where 0 represents the lowest risk of corruption and white collar crimes, and 100 corresponds to the highest risk of corruption and white collar crimes. Each country's global score is then calculated following the weights previously presented.

5- Measure of Uncertainty

Based on the *n* datasets obtained from the multiple imputation process, a standard error and a 90 percent confidence interval are calculated for each dataset to reflect the variance around the different scores.

Global Risk Profile Sàrl
91 Rue de Lyon
1203 Geneva
Switzerland

Global Risk Profile Sàrl
Email.info@globalriskprofile.com
CH +41 (0) 22 794 39 11
FR +33 (0) 1 73 19 11 72
Web. www.globalriskprofile.com

Global Risk Profile France
27 Avenue de l'Opéra
75001 Paris
France

Index Research & Development
Sonia THURNHERR
Lead Data Scientist