**OECD States of Fragility - Review**

Institute for Economics and Peace (IEP)

09 June 2022

Contents

[**Background and purpose of this report** 1](#_Toc106362239)

[**DATA INTEGRITY CHECKS** 1](#_Toc106362240)

[**Checking country names** 1](#_Toc106362241)

[**Checking data coverage** 2](#_Toc106362242)

[**Checking NA’s** 2](#_Toc106362243)

[**DATA IMPROVEMENTS** 3](#_Toc106362244)

[**Data updates** 3](#_Toc106362245)

[**Log calculations** 3](#_Toc106362246)

[**DATA ERRORS** 3](#_Toc106362247)

[**Operating System specific errors** 3](#_Toc106362248)

[**Incorrect file links** 3](#_Toc106362249)

[**Possible syntax errors** 4](#_Toc106362250)

[**IMPACT ON FINAL RESULTS** 4](#_Toc106362251)

[**RECOMMENDATIONS** 4](#_Toc106362252)

## **Background and purpose of this report**

The OECD States of Fragility requested that IEP review the code used to produce the OECD States of Fragility Report.

IEP has generated four different tests to assess data quality.

The function ``

## **DATA INTEGRITY CHECKS**

IEP has implemented a number of tests to assess data integrity. These tests allow an individual to check whether code is functioning as expected and producing sensible results.

### **Checking country names**

The States of Fragility framework frequently converts iso3c codes to country names and vice versa. This conversion is implemented via a custom function in nearly 35 different scripts and data sources. The function is well implemented and accounts for multiple edge cases.

There are still instances in which converting from iso3c codes to country names fails. Failed conversion of iso3c codes is not always problematic if, for example it occurs for region names as opposed to country names. Unsuccessful conversion becomes problematic when it occurs for counties or country iso3c codes.

When the conversion fails, resulting values are set to be NA and a warning is produced in the terminal. NA values are frequently removed in later parts of the script with additional functions like  ***%>% drop.na()*** or ***na.omit()*** . Therefore, failed conversion can result in unnecessary data loss.

To address this issue, a check was implemented to determine the countries that could not be converted from “iso3c” codes to a country name. The function identifies the rows in which the “iso3c” codes were unable to be converted by finding the rows in which the values of “iso3c\_CONVERT” values are NA. The actual country names that could not be converted are printed to the log files and outputted to a csv file. The function takes four arguments:

* df – a dataframe (long format)
* COL1 – an unquoted column name containing what you want to convert
* COL2 – an unquoted column names containing what was converted
* SCRIPT – a string defined at the beginning of each script. This is appended to the filename to provide additional detail about which script the issue occurred in

**check\_convert\_iso(dat\_long, iso3c, iso3c\_CONVERT, "SCRIPT")**

Note that the order of the arguments iso3c and iso3c\_CONVERT is important. Reversing the columns will perform a different check.

### **Checking data coverage**

Each dataset in the States of Fragility framework contains a country, a year, a variablename and a value. For a given variable name, each country should ideally have a corresponding value for all years. However, even well curated datasets may not contain values for every single year for each country. This highlights data quality issues in the context of requiring update data each year. Therefore, it can be helpful to know which countries are missing data for the latest year.

Missing data can be problematic because necessitates a reliance on imputation and imputation may affect which countries are classified as fragile or not fragile. Imputation may be acceptable for some countries, but may not be desirable for others.

To address this issue, IEP implemented a check to compare the unique country names in the latest year versus all preceding years. This check is a quick test to determine whether data exists for the majority of countries cross the entire time series. The function takes four arguments:

* df – a dataframe (long format)
* C1 – a string name of the “iso3c” column
* C2 – a string name of the “year” column
* SCRIPT – a string defined at the beginning of each script. This is appended to the filename to provide additional detail about which script the issue occurred in

**check\_coverage(dat\_clean, "iso3c", "year", SCRIPT)**

Note: This data is also output to a csv as well as the log file

### **Checking NA’s**

It can be helpful to understand the number of NA values in each column. This can occur for a number of reasons such as unsuccessful conversion from iso code to country or vice versa, missing data or the inclusion of data that should not be there.

The function prints a notice to the terminal and save a warning to the log file indicating the number of NA values in each column.

This function counts the number of NA’s in each column. The code takes arguments:

* df – a dataframe (long format)
* SCRIPT - a string defined at the beginning of each script. This is appended to the filename to provide additional detail about which script the issue occurred in

**check\_nas(dat\_clean, SCRIPT)**

Note: the output of this function is only printed to the log file. It is not printed to a csv.

Each of the functions are flexible enough to be run at multiple stages throughout each script. The value of “SCRIPT” can be changed at any point to save a new output file within a script if desired.

## **DATA IMPROVEMENTS**

### **Data updates**

While reviewing code, IEP noted two scripts in which more recent data was available. This affected two scripts (*09-INFORM* and *21-environmental displacement*). This data has *been* updated.

### **Log calculations**

Two scripts *35 non state and one sided conflict* and *36-battle deaths* contained log transformations of the data. The initial scripts took the log of the existing value. Although this is a methodologically sound approach, IEP suggested a more nuanced approach similar to what is used in the Global Terrorism Index (GTI). The change means that the log values now reflect a score relative to the worst year on record, scaled between 0 and 10. The difference between 9 and 10 should reflect a greater change in fragility than a change from 1-2.

The code is implemented as follows:

***max\_score = max(dat\_final$value)***

***log\_base = (max\_score+1)^(1/10)***

***dat\_final$value <- log(dat\_final$value + 1, log\_base)***

***dat\_final <- as\_tibble(dat\_final)***

***dat\_final$variablename <- "non\_state\_one\_sided"***

## **DATA ERRORS**

Overall the code ran as expected and did not produce any substantial errors that would impact the calculation of the OECD Fragile States Index. The instances in which code produced overt errors are summarised below.

### **Operating System specific errors**

The data integrity check revealed a number of issues affecting different operating systems. While the existing code was able to run without errors on Mac, the same code produced errors when run on Windows. The issues seem to be caused by special or invalid characters in file names. For example:

* The initial repository contained multiple files with a timestamp that contained a colon “:”. This was caused by appending ***Sys.time()*** as part of the filename. The issue prevented cloning of the repository. Replacing the code with ***format(Sys.time(), %d, %m %y %H %M”)*** to exclude the colon and removing the previously affected files resolved the issue.
* Certain filenames contained special characters. For example, the file in ***“./data/additional data/UNSD — Methodology.csv“*** contained an em dash “—“. This caused an issue such that running the code ***source(/”src/05-Analysis”)*** produced errors but running the code by copying and pasting directly into the terminal did not. Removing the special character from the affected filename resolved the issue.

The fixes should ensure the code can be run on any computer regardless of operating system.

### **Incorrect file links**

There is potentially an incorrect link in the *05-Analysis* script in which ***list.2016.full*** links to a file on the githubIEP version of the repository while the variable ***list.2020.full*** links to the hdesaioecd repository. This runs fine on an internal IEP computer, but is likely to produce errors on external systems that do not have access to the same repository.

### **Possible syntax errors**

There were no observable syntax errors noticed throughout the script. However, there is one instance in which a conditional may not be working properly. The script *01-SFR Calculation* contains a variable ***dimensional.pca.models.*** On line 109, inside the conditional, the value of ***all.drops*** is set. The value is set through the ***<<-*** operator which assigns a value of a variable globally and is functionally different from the ***<-*** operator which assigns a value of a variable locally. Although the condition does not seem to execute, the behavior of the operator may not be doing what is intended.

## **IMPACT ON FINAL RESULTS**

The changes IEP made to the code do not appear to have impacted the calculation of the SoF index in any significant manner. A direct comparison of the file “export of final results” produced by IEP (with changes) and an earlier version of the same file produced by OECD (without changes) revealed:

* All countries that were classified as “extremely fragile” or “other “fragile” are still classified as such.
* No country that was classified as “extremely fragile” switched to being classified as “other fragile” or vice versa
* All countries that were classified as “Rest of the World” are still classified as such
* There were minimal changes in the overall rankings with most countries changing between 0-2 positions in either direction.

For this reason IEP is confident the changes that were implemented did not induce any substantial changes to the OECD states of fragility index in the scores or rankings or classifications.

## **RECOMMENDATIONS**

All changes suggested by IEP have been implemented.

The addition of unit check to test for country coverage, number of na values, checking for duplicates,

The following changes were \*not\* implemented:

* OECD originally requested the calculation of a composite measure of violence against women in script “32-sigi”. The code was amended to calculate a composite measure. However, OECD opted to use the original code containing the “attitudes” domain due to limited data availability for the calculation of the composite measure. OECD conformed with colleagues that this is conceptually sensible and IEP agreed.
* Some scripts contained duplicate data (specifically *“03-aid volatility”* and *“10-political rights”*). This was initially resolved by implementing a temporary fix that calculated the mean of the two country year pairs. OECD had opted to use the minimum of the two values in line with the standard practice of previous State of Fragility reports.