Configuration Manual

MSc Research Project
Fintech

Anthony Mobolaji Falola
Student ID: 20242727

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Falola Anthony Mobolaji
Student ID:	20242727
Programme:	M.Sc. Fintech **Year:** 2021/2022
Module:	M.Sc. Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	August 15, 2022
Project Title:	Investigating the Impact of Socioeconomic Factors on Financial Inclusion in Sub-Saharan Africa
Word Count:	1,671 **Page Count:** 10

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Anthony Mobolaji Falola
20242727

1 An Introduction

This is a user configuration manual containing technical details, specifications and procedures necessary to reproduce the research analysis titled: "Investigating the Impact of Socioeconomic factors on Financial Inclusion in Sub-Saharan Africa"

2 System Requirements

2.1 Hardware

- MacBook Air (M1, 2020) 13.3-inch laptop
- Mac OS Monterey version 12.4
- Memory – 8GB unified memory
- Storage - 256 GB SSD

2.2 Software

- R programming language and R Studio - version 4.2.1 - 'funny looking kid' – for data and statistical analysis
- Microsoft Word for Mac – used for report writing
- Microsoft Excel for Mac – used for primary data inspection and analysis of tables and data

3 Data

The International Monetary Fund (IMF) Financial Access Survey dataset can be downloaded from the IMF database[1] while the World Bank World Development Indicators[2] is the source for the second dataset. The downloaded CSV files for both datasets is saved in the downloads folder. The former is named the 'FAS' while the latter is named 'WDI'.

[1] https://data.imf.org/?sk=E5DCAB7E-A5CA-4892-A6EA-598B5463A34C&sId=1460043522778
[2] https://databank.worldbank.org/source/world-development-indicators

4 Analysis

4.1 Installation of Required Packages for Analysis

Package Name	Version	Use
Amelia	1.8.0	Missing data treatment and visualisation
BiocManager	1.30.18	To access the Bioconductor project repository for ppca
car	3.1-0	Used to carry out multicollinearity test with the vif() function
dplyr	1.0.9	used for effective data manipulation
lmtest	0.9-40	Used to test regression models
missMDA	1.18	Used to impute missing values for PCA
pcaMethods	1.88.0	Used for ppca imputation
plm	2.6-1	panel data analysis
psych	2.2.5	Used for validation test like Bartletts and KMO test
rrcov	1.7-0	Necessary to carry out robust PCA
stargazer	5.2.3	Used to summarize regression coefficients
stringr	1.4.0	Used to manipulate data strings

4.2 Data Pre-processing and Transformation

Prior to getting a final dataset for the analysis, it is important to import the datasets and pre-process before they are merged.

4.2.1 IMF FAS Dataset

- Import the dataset and save as FAS

```
FAS = read.csv("FAS Excel.csv", header = TRUE, na.strings = c(""))
```

- Select the indicators as proposed by the G20 on financial inclusion

```
FASG20 = FAS %>% select(Economy,
                        Year,
                        Number.of.commercial.bank.branches.per.100.000.adults,
                        Number.of.ATMs.per.100.000.adults,
                        Number.of.registered.mobile.money.agent.outlets.per.100.000.adults,
                        Number.of.mobile.money.transactions..during.the.reference.year..per.1.000.ad
                        ults,
                        Number.of.deposit.accounts.with.commercial.banks.per.1.000.adults,
                        Number.of.life.insurance.policies.per.1.000.adults,
                        Number.of.Non.life.insurance.policies.per.1.000.adults,
                        Number.of.SME.deposit.accounts.with.commercial.banks....of.non.financial.c
                        orp.,
                        Loan.accounts.with.commercial.banks..o.w.SME
                        Number.of.loan.accounts.with.commercial.banks.per.1.000.adults)
```

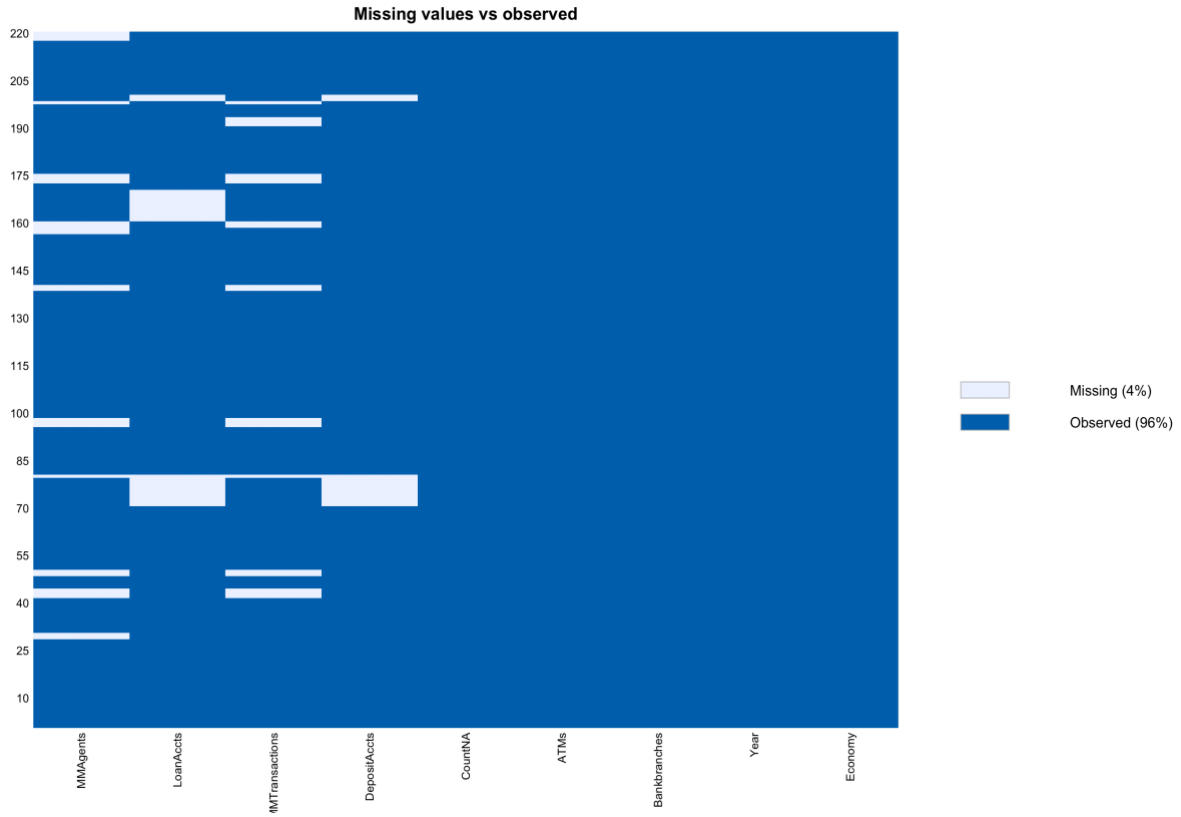
- Remove the countries and columns with missing data notoriety. This excludes countries such as Benin, Cabo Verde, and Burkina Faso
- Change mobile money values to zero for countries with record of no mobile money in some of the years observed such as Chad in 2013 and Equatorial Guinea in 2016
- Change the value of the all the financial inclusion indicators to numeric
- Change index number by the removal of the unwanted rows, the row index numbering should be re-ordered in an ascending manner.
- Rename columns to prevent lengthy codes and errors in spelling

```
colnames(FASG20Q) = c("Economy", "Year", "Bankbranches", "ATMs", "MMAgents", "MMTransactions", "DepositAccts", "LoanAccts", "CountNA")
```

Missing data treatment

- Examine the dataset for pattern of missingness
- Use Last Observation Carried Forward (LOCF) and mean substitution to replace values with previous or subsequent values present within the same entity.
- Use missmap() to check missing data chart

```
missmap(FASG20Q, main = "Missing values vs observed")
```



PPCA Imputation
Since the outstanding missing data is missing at random, we use probabilistic principal component analysis technique which is effective for robust principal component analysis with a significantly better performance than other correctional methods(Li Qu *et al.,* 2009).

```
NormalIMFFAS = scale(IMFFAS[3:8])
estim_ncpPCA(NormalIMFFAS,method.cv = "kfold",scale = TRUE)
PPCAData = pca(NormalIMFFAS, nPcs = 3, method = "ppca")
ImputedData = completeObs(PPCAData)
```

- Where nPcs = 3, is the number of principal components derived from the estimation.

```
CompleteFAS = cbind(IMFFAS[, 1:2],ImputedData)
```

- Create complete FAS dataset by binding new values provided by the PPCA imputation with the Year and Economy column

4.2.2 WDI Dataset

Import the dataset and save as WDI

```
WDI = read.csv("WDIData.csv", header = TRUE, na.strings = c(""))
```

- Remove empty and redundant rows and columns – such as country and year code
- Reorder rows and columns and their index numbers
- Change column names

```
colnames(WDI) = c("Economy", "Year", "UrbanPopulation", "ElectricityAccess", "InternetUsage", "Unemployment", "FemalePopulation", "GDPpercapita", "WorkingAge", "PrimarySchool")
```

- Change '..' to NA as missing values in the WDI dataset is being represented by two dots instead of NA
- Check data for missing values using missmap()
- Replace unemployment number in Seychelles with numbers from the national estimate from the WDI database

4.2.3 Robust Principal Component Analysis

4.2.3.1 Suitability of Data for rPCA

- Check data correlation using the 'cor' function
- Carry out Bartlett test using the 'cortest.bartlett' function available in the psych package.
- Perform KMO test using the 'kmo' function

```
CorCompleteFAS = cor(CompleteFAS[,3:8])
library('psych')
cortest.bartlett(CorCompleteFAS,n = nrow(CompleteFAS))
KMO(CorCompleteFAS)
```

Output

```
$chisq
[1] 2037.395

$p.value
[1] 0

$df
[1] 15

> #KMO test
> KMO(CorCompleteFAS)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = CorCompleteFAS)
Overall MSA =  0.83
MSA for each item =
  Bankbranches           ATMs       MMAgents MMTransactions    DepositAccts      LoanAccts
          0.91           0.84           0.59           0.55           0.85           0.87
> |
```

- Overall MSA is 0.83, greater than the minimum required

4.2.3.2 Creating the Indices

Robust PCA is used to estimate the indices for the usage and access dimension, before a second stage indexing to create the Index of Financial Inclusion (IFI)

Access

```
pcaAccess <- PcaCov(CompleteFAS[, c(3:5)], scale = TRUE,center = TRUE)
```

- The normalised scores and eigenvalues generated from the above is substituted in the equation for *AccessIndex* below

$$AccessIndex_{i,t} = \frac{\sum_{j,k=1}^{t} \lambda_j P_{k,i,t}^{a}}{\sum_{j,k=1}^{t} \lambda_j^{a}}$$

- Where $P_{ki,t}$ is the scores from the rPCA for each of the lines

	PC1	PC2	PC3
1	4.27804238	-2.7243705481	2.704547134
2	4.22898347	-2.7448605221	2.612681868
3	4.58076241	-2.7311660631	2.779646429
4	4.67638702	-3.0747417105	2.963092642
5	5.21134602	-2.9368457139	3.686777373
6	5.20716948	-2.9953861091	3.600416266
7	5.27260136	-3.0293585341	4.042421467

- Eigenvalues, λ_j is the variance of the access dimension as shown below

```
> print(variance_access)
[1] 2.2423970 0.5860797 0.1715234
```

Usage

- The same process is repeated for the usage index using the indicators that represent the usage dimension

```
pcaUsage <- PcaCov(CompleteFAS[, c(6:8)], scale = TRUE,center = TRUE)
```

- The variance and scores computed is used for the usage formula to produce the usage index, UI

```
UsageIndex = t(t(NormUsage)*variance_usage)
UI = rowSums(UsageIndex)/sum(variance_usage)
```

Where NormUsage is the min-max normalisation of the pcaUsage$scores.

The UsageIndex represents the index for the usage dimension

```
> head(UI)
        1         2         3         4         5         6
0.4122731 0.4163921 0.4070241 0.4057363 0.3987546 0.3888258
```

Index of Financial Inclusion

The *AccessIndex* and *UsageIndex* is added to the CompleteFAS dataset to form the 9th and 10th column which is then used for the second stage rPCA

```
pcaIFI <- PcaCov(CompleteFAS[, c(9:10)], scale = TRUE,center = TRUE)
```

The values are substituted in the formula below

$$IFI_{i,t} = \frac{\sum_{j=1}^{2} \lambda_j (\psi_{j1} AccessIndex_{i,t} + \psi_{j2} UsageIndex_{i,t})}{\sum_{j=1}^{2} \lambda_j}$$

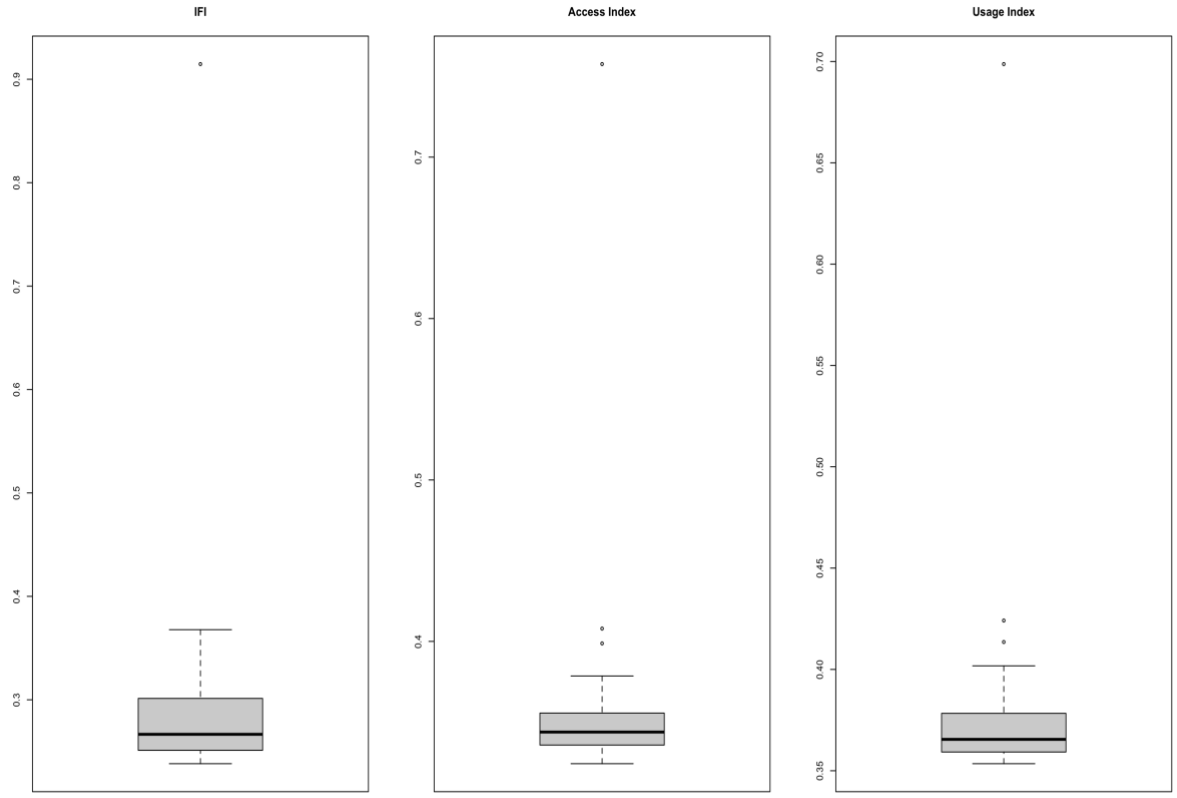
Where the variance is

```
> print(variance_IFI)
[1] 1.92595634 0.07404366
```

The scores based on the principal component analysis of both dimensions is

```
> head(IFIscores)
       PC1       PC2
1 6.205870 -3.534594
2 6.562396 -3.935589
3 5.834176 -2.940930
4 5.722938 -2.815319
5 5.307600 -1.946647
6 4.385149 -1.043073
```

- A boxplot of the mean values of each of the indices for each country was used to investigate the presence of outliers

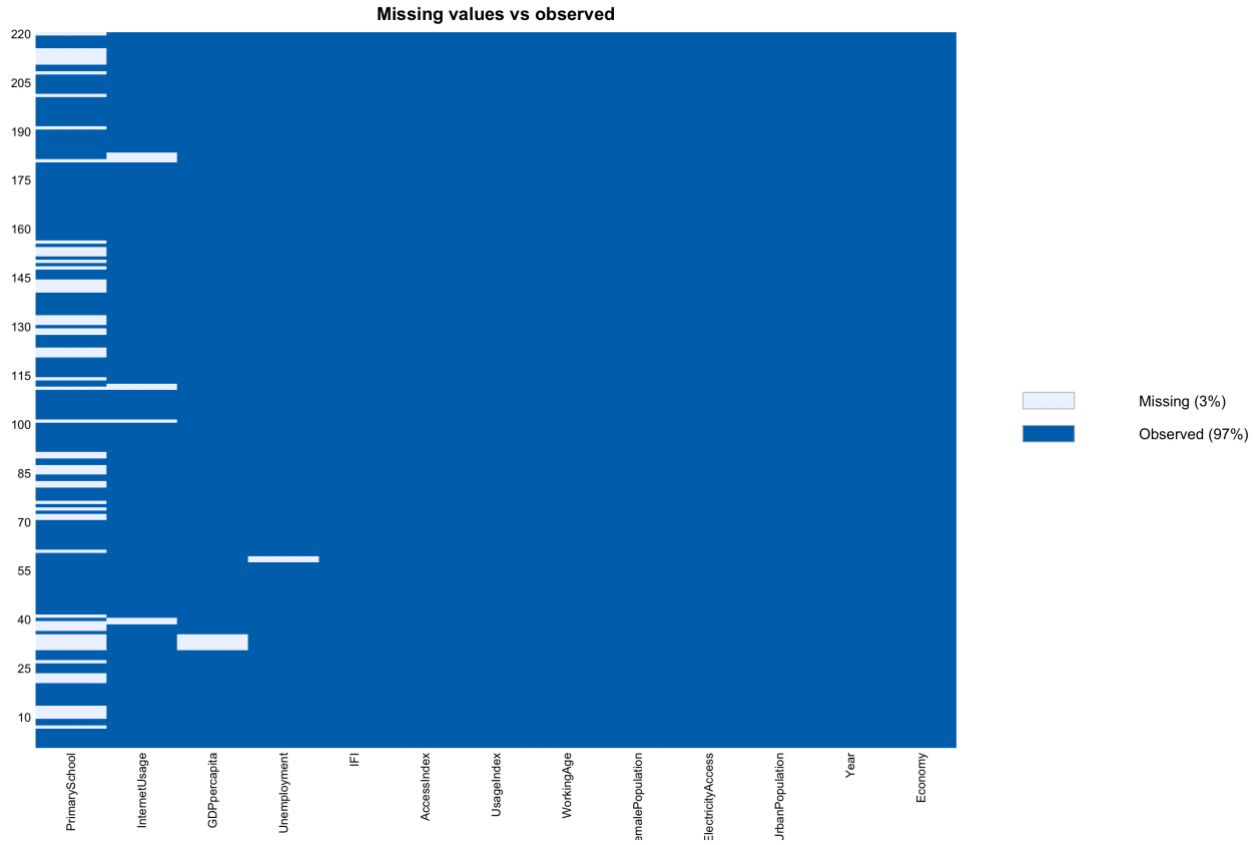


4.2.4 Merged Dataset

The merged dataset is created by combining the WDI dataset with the indices produced from the FAS dataset

```
FullDataset = cbind(WDI, Indices[, 3:5])
```

- The missing data analysis is visualised to inspect for patterns in missingness



- Use mean substitution to replace the missing values in unemployment rate in Seychelles
- Use linear regression to impute missing values in internet usage data for Eswatini
- Use LOCF for missing data imputation for internet usage data in South Sudan 2011 and 2012 and Madagascar in 2019 and 2020
- GDP per capita data for South Sudan was sourced from UN Database[3]
- The data structure for all the variables except the year and economy was changed to numeric
- All percentages were changed to decimals by dividing through with 100
- Lastly, the merged dataset was inspected for correlation

```
CorFullDataset = cor(FullDataset[,3:12])
```

[3]http://data.un.org/Data.aspx?q=sudan&d=SNAAMA&f=grID%3A101%3BcurrID%3AUSD%3BpcFlag%3A1%3BcrID%3A728%2C729%2C736

5 Data Mining

5.1 Panel Data Analysis

The data variables that make the dataset a panel dataset is assigned as follows

```
MyData = pdata.frame(FullDataset, index = c("Economy" , "Year"))
```

5.2 Fixed Effects Estimation

Using the plm library, fixed effects estimates is carried out with all the independent variables on each of the dependent variable

```
Model1 = plm(log(AccessIndex)~UrbanPopulation + ElectricityAccess + InternetUsage + Unemployment + FemalePopulation + log(GDPpercapita) + WorkingAge,
        data = MyData,
        model = "within")
```

The results are analysed using the summary() and stargazer() function

5.3 Hausman Test

This is done by comparing the results of a random effects estimation and a fixed effect estimation with the same parameters using the phtest() function

```
Model4 = plm(log(IFI)~UrbanPopulation + ElectricityAccess + InternetUsage + Unemployment + FemalePopulation + log(GDPpercapita) + WorkingAge,
        data = MyData,
        model = "random")

phtest(Model4, Model3)
```

5.4 VIF Test

This is done by using the vif() function from the library car on the regression model

	IFI	
	VIF	**1/VIF**
UrbanPopulation	1.922	0.520
ElectricityAccess	5.899	0.170
InternetUsage	4.678	0.214
Unemployment	1.834	0.545
FemalePopulation	1.458	0.686
log(GDPpercapita)	4.407	0.227
WorkingAge	4.155	0.241

5.5 Breusch-Pagan Test

To test for heteroscedasticity, the bptest() function from the lmtest library is used.

```
bptest(Model1)
```

- the homoscedasticity null hypothesis is not rejected

5.6 Robust-Covariance Matrix

This is used to control heteroscedasticity in the model

```
ControlIFI = coeftest(Model3,vcovHC(Model3,method="white2"))
```

5.7 Excluding Outliers

The inclusion results for South Africa are an outlier compared to other economies, hence this is removed from the dataset while the model is tested on the new dataset to compare the result.

References

Li Qu *et al.* (2009) "PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach," *IEEE Transactions on Intelligent Transportation Systems*, 10(3), pp. 512–522. Available at: https://doi.org/10.1109/TITS.2009.2026312.