Building the W&S Integrity Index Risk Index

WIRI Example Data

Government Transparency Institute & Water Integrity Network

December 15, 2021

Links for Today

To follow along with today's training, you will find the replication material in any of the links below:

- RStudio Cloud
- GitHub

Outline

CRI Design

Indicators

Public procurement risk indicators

We assign each public procurement contract to one of the 3 pillars using product codes specific to the nature of W&S activity defined by public procurement data systems such as the Common Procurement Vocabulary (CPV) codes.

The public procurement risk indicator is a composite score of five elementary risk indicators:

- Decision Period
- Call for Tenders
- Advertisement Period
- Procedure Type
- Single Bidding

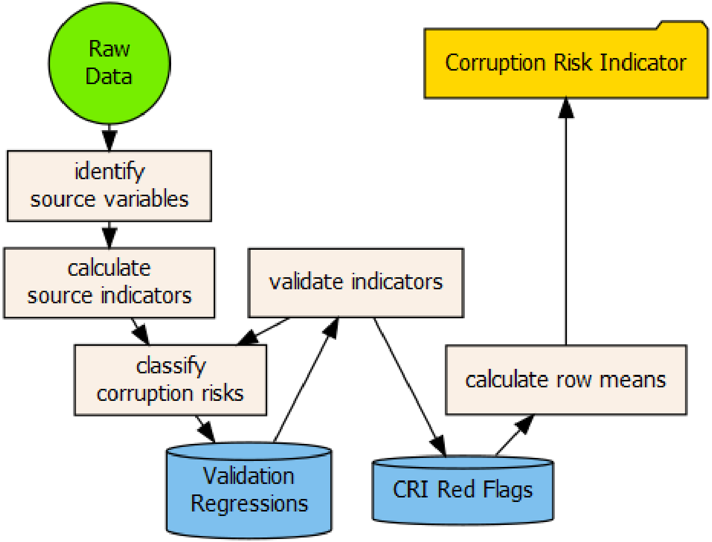
The composite score is scaled so that it falls between 0 and 100, with 100 representing the highest integrity and 0 representing the lowest integrity (lack of integrity).

Integrity Risk Indicators

INDICATOR NAME	INDICATOR DEFINITION
LENGTH OF DECISION PERIOD	100=length of decision period is unrelated to corruption risks (single bidding) 0=length of decision period OR missing decision period is related to corruption risks (single bidding)
PROCEDURE TYPE	100=open 0=non-open (accelerated, restricted, award without publication, negotiated, tender without competition)
SINGLE BIDDER CONTRACT	100=more than 1 bid received 0=1 bid received
CALL FOR TENDERS PUBLICATION	100=call for tender published in official journal 0=NO call for tender published in official journal
LENGTH OF ADVERTISEMENT PERIOD	100=length of advertisement period is unrelated to corruption risks (single bidding) 0=length of advertisement period or missing advertisement period is related to corruption risks (single bidding)

CRI Calculations

Workflow



Raw Data
identify source variables
calculate source indicators
validate indicators
Corruption Risk Indicator
classify corruption risks
calculate row means
Validation Regressions
CRI Red Flags

Main CRI Variables

Time-series WIRI

```r
source_cri_vars = c("tender_recordedbidscount",
                    "tender_publications_firstcallfortenderdate",
                    "tender_biddeadline",
                    "tender_publications_firstdcontractawarddate",
                    "tender_proceduretype",
                    "tender_isawarded")
```

CRI Source Indicators

```r
df <- df %>%
  mutate(
    #Single bidder
    singleb=ifelse(tender_recordedbidscount>1,0,1),
    singleb=as.factor(singleb),
    #Advert Period
    submp=as.duration(
      interval(tender_publications_firstcallfortenderdate,tender_biddeadline))
    %/% as.duration(days(1)),
    submp=ifelse(submp>365,NA,submp),
    submp=ifelse(submp<0,0,submp),
    submp10= ntile(submp, 10),
    #No Call for Tenders
    ncft = case_when(
      is.na(tender_biddeadline) &
        tender_isawarded==T~1,
      TRUE~0
    ))
```

CRI Source Indicators

```
df <- df %>%
  mutate(
    #Decision Period
    decp=as.duration(interval(tender_biddeadline,tender_publications_firstdcontractawarddate))
    %/% as.duration(days(1)),
    decp=ifelse(decp>365,NA,decp),
    decp=ifelse(decp<1,NA,decp),
    decp10=ntile(decp, 10),
    #Procedure Type
    proc=case_when(
      tender_proceduretype=="OPEN"~0,
      tender_proceduretype=="APPROACHING_BIDDERS"~1,
      tender_proceduretype=="DESIGN_CONTEST"~1,
      tender_proceduretype=="OTHER"~1,
      tender_proceduretype=="DPS_PURCHASE"~1,
      tender_proceduretype=="NEGOTIATED_WITH_PUBLICATION"~1,
      tender_proceduretype=="RESTRICTED"~2,
      TRUE~NA_real_)
  )
```

CRI Risk Classification

```
df <- df %>%
  mutate(
    singleb=case_when(
      singleb==0~0,
      singleb==1~100,
      is.na(singleb)~99),
    singleb=as.factor(singleb),
    ncft=case_when(
      ncft==0~0,
      ncft==1~100,
      is.na(ncft)~99),
    ncft=as.factor(ncft),
    corr_proc=case_when(
      proc==0~0,
      proc==1~50,
      proc==2~100,
      is.na(proc)~99),
    corr_proc=as.factor(corr_proc)
  )
```

CRI Risk Classification

```r
df <- df %>%
  mutate(
   corr_submp=case_when(
      submp10 %in% c(1)~0,
      submp10 %in% c(7:10)~100,
      submp10 %in% c(2:6)~50,
      is.na(submp10)~99),
    corr_submp=as.factor(corr_submp),
    corr_decp=case_when(
      decp10 %in% c(5:8)~100,
      decp10 %in% c(1:4)~50,
      decp10 %in% c(9:10)~0,
      is.na(decp10)~99),
    corr_decp=as.factor(corr_decp),
  )
```

Validation Regressions

	Dependent variable:			
	singleb			
	(1)	(2)	(3)	(4)
corr_proc50	0.4			0.8
	(0.7)			(0.9)
corr_proc100	0.7**			0.3
	(0.3)			(0.4)
corr_submp50		1.2**		1.4***
		(0.5)		(0.5)
corr_submp99		2.2*		1.2
		(1.2)		(1.5)
corr_submp100		1.5***		1.7***
		(0.5)		(0.5)
corr_decp50			1.0**	1.1**
			(0.4)	(0.4)
corr_decp99			1.8**	1.9*
			(0.8)	(1.0)
corr_decp100			1.6***	1.6***
			(0.4)	(0.4)
Constant	4.1***	4.3***	2.3	1.6
	(1.3)	(1.3)	(1.5)	(1.6)

CRI and Integrity Risk Calculations

```
main_cri_vars = c("singleb","corr_proc","corr_submp","corr_decp","ncft")

df$cri = rowMeans(df[,main_cri_vars], na.rm = T)

df$cri_integrity = 100 - df$cri
```

WIRI Design

WIRI Design

We use a data-driven approach to develop a composite Water Integrity Risk Index (WIRI) made up of a host of objective proxy indicators as well as survey-based measures of corruption experience to identify and assess integrity risks in the urban WS sector in selected settlements around the country.

We identify three main pillars of integrity in the W&S sector:

1. Public investment projects (e.g. building new pipelines or drainage),
2. Recurrent spending supporting ongoing operations (e.g. paying salaries, purchasing computers), which is addressed as operations in this work; and
3. Client-utility interactions (e.g. paying utility bills).

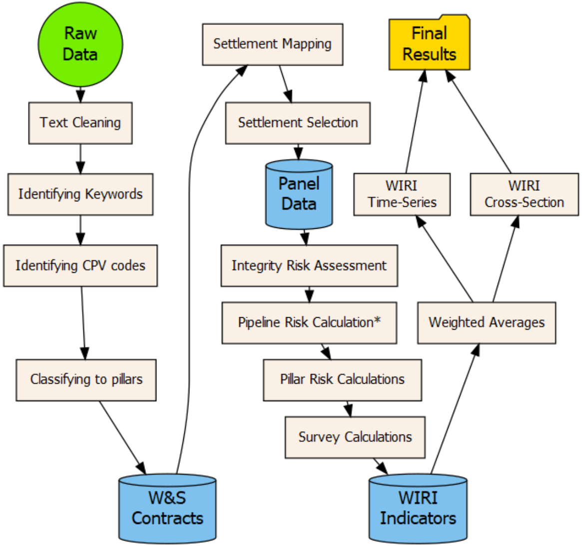
Methodology

Given that integrity is a latent variable, we must rely on proxy indicators which can, in conjunction, reveal integrity risks. We calculate the composite WIRI with the following steps:

1. We standardize each component indicator of integrity-risk so that they can be directly compared (higher values imply higher integrity).
2. Calculate the weight of each component indicator (5 in total, categorized into 3 pillars) by the amount of data points available for the timeseries in a global version of the WIRI. Fewer available data points in a component lead to a decrease its pillar weight on the index.
3. We calculate the weighted mean of each indicator to derive the composite WIRI score based on the data available.

WIRI Calculations

Workflow


Raw Data
Text Cleaning
Identifying Keywords
Identifying CPV codes
Classifying to pillars
W&S Contracts
Settlement Mapping
Settlement Selection
Panel Data
Integrity Risk Assessment
Pipeline Risk Calculation*
Pillar Risk Calculations
Survey Calculations
WIRI Indicators
Weighted Averages
WIRI Time-Series
WIRI Cross-Section
Final Results

Water Contracts Classification

The classification of WS Contracts for Kenya relies (mostly) on string matching over three contract-level variables: a) tender title, b) procuring entity name, and c) supplier name. These variables go through a first round of cleaning shown below:

```
df <- df %>%
  mutate(
    #General cleaning
    tender_title=tolower(tender_title),
    tender_title=stringi::stri_trans_general(tender_title, "Latin-ASCII"),
    buyer_name=tolower(buyer_name),
    buyer_name=stringi::stri_trans_general(buyer_name, "Latin-ASCII"),
    bidder_name=tolower(bidder_name),
    bidder_name=stringi::stri_trans_general(bidder_name, "Latin-ASCII"),
    )
```

Water Contracts Classification

After this pre-processing step, the broadest form of classification for these three variables consists of identifying references to water, pipelines, and sewage. Observations are classified as 1 if a keyword is found in the variable, and 0 otherwise.

```
df <- df %>%
  mutate(
    #Buyer Name Matching
    water_buyername=case_when(
      #broad
      grepl("water",buyer_name)~1,
      grepl("sewe+",buyer_name)~1,
      grepl("pip+",buyer_name)~1,
      TRUE~0)
    #Bidder Name Matching
    water_biddername=case_when(
      #broad
      grepl("water",bidder_name)~1,
      grepl("sewe+",bidder_name)~1,
      grepl("pip+",bidder_name)~1,
      TRUE~0))
```

Water Contracts Classification

In cases where broad string matching leads to ambiguous results, a narrower matching strategy is employed. For example, tenders that relate to bottled drinking water are discarded (classification 0).

```r
df <- df %>%
  mutate(
    #Tender Title Matching
    water_tendertitle=case_when(
      #narrow
      grepl("sanitation|sanitary|sewer",tender_title)~1,
      (grepl("water",tender_title)&
         grepl("network|construction|channel|system|testing",tender_title))~1,
      (grepl("water",tender_title)&
         grepl("district|treatment|channel|system|testing",tender_title))~1,
      (grepl("water",tender_title)&grepl("pipe+",tender_title))~1,
      (grepl("water",tender_title)&grepl("sewe+",tender_title))~1,
      (grepl("water",tender_title)&grepl("distill+",tender_title))~1,
      (grepl("water",tender_title)&grepl("gutt+",tender_title))~1,
      (grepl("exten+",tender_title)&grepl("pipe+",tender_title))~1,
      (grepl("water",tender_title)
       &grepl("rain",tender_title)
       &grepl("collect+",tender_title))~1,
      (grepl("water",tender_title)
       &grepl("supply",tender_title)
       &!grepl("drink+",tender_title))~1,
      TRUE~0))
```

Water Contracts Classification

We classify contracts based on their Common Procurement Vocabulary (CPV) codes. Depending on this classification, we determine whether they fall into one of the WIRI's pillars (first two levels). Similarly, we classify pipeline contracts based on full CPV codes and string matching.

```
#Two-level CPV codes
investment<-c(16,31,32,34,35,
              42,43,48,44,45,71)
operation<-c(03,09,15,18,19,
             22,24,30,33,37,38,
             39,41,70,50,51,60,
             63,64,65,66,72,73,
             75,77,79,80,85,90,
             92,98,55)
#Full CPV codes
pipes<-c(45232150,44162500,
         45232411,45232440,
         45232100,45232121,
         45232130,45231300,
         44134000,44163130,
         44160000,44161000,
         45231100,45231110,45231112)
```

Water Contracts Classification

```
df <- df %>%
  mutate(
    #Investment Contracts
    inv_contract=ifelse(tender_cpvs_2%in%investment,1,0),
    #Operations Contracts
    op_contract=ifelse(tender_cpvs_2%in%operation,1,0),
    #CUI Contracts
    inter_contract=water_biddername,
    #Other Contracts
    other_contract=ifelse(tender_cpvs_2%in%other,1,0),
    #Pipe Contracts
    pipe_contact=case_when(
      tender_cpvs_8%in%pipes~1,
      grepl("pipe+",tender_title)~1,
      TRUE~0))
```

WS Contracts Pruning

After the classification process, we select observations were any of the following conditions is met:

- A W&S CPV Code (full-8)
- A W&S Buyer Name
- A W&S Supplier Name
- A W&S Tender Title

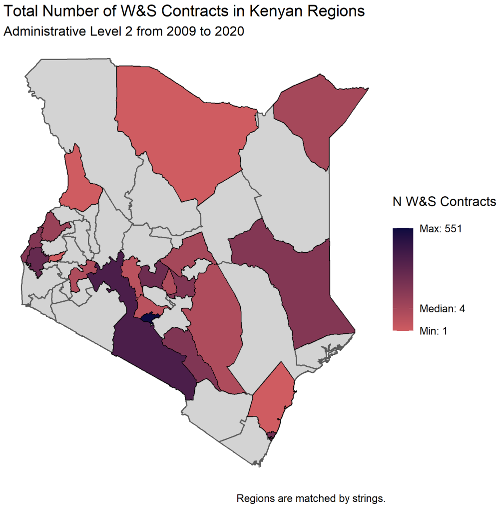
A) By Number of W&S Contracts

```
water_df <- df %>%
  filter(water_cpv==1
   | water_buyername==1
   | water_tendertitle==1
   | water_biddername==1)
```

Selecting Settlements

Top Settlements Contracts by Year

tender_year	buyer_city	N
2018	Mombasa	20
2019	Mombasa	5
2012	Nairobi	2
2014	Nairobi	1
2015	Nairobi	3
2016	Nairobi	16
2017	Nairobi	25
2018	Nairobi	303
2019	Nairobi	193
2020	Nairobi	8
2019	Nakuru	89
2017	Nyeri	1
2018	Nyeri	16
2020	Nyeri	10
2018	Siaya	12
2019	Siaya	24



Selecting Settlements

We have two options for selecting W&S Settlements: a) by slicing the top n (5) settlements buyer_city by number of observations, and b) by determining them directly based on other relevant criteria (e.g., partnerships).

A) By Number of W&S Contracts

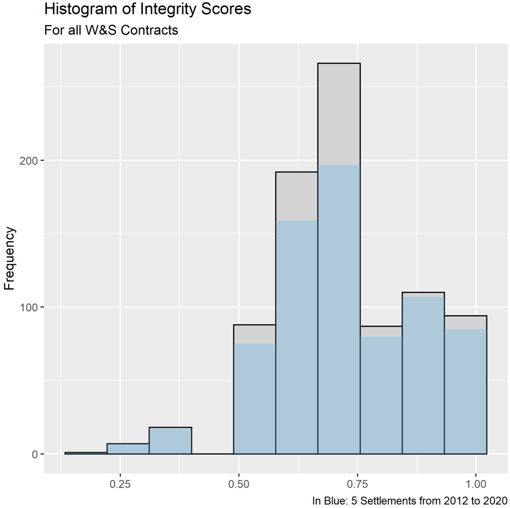
```
settlements <- df_water %>%
  summarise(n_water_contracts=n()) %>%
  filter(n_water_contracts>=10) %>%
  arrange(desc(n_water_contracts)) %>%
  head(5)

df_water <- df_water %>%
  filter(buyer_city%in%
         settlements$buyer_city)
```

B) Predetermined Settlements

```
#Example Only
water_settlements<-c("Nairobi","Busia",
                     "Nyeri","Kitui",
                     "Makueni", "Mombasa")
#Example Only
df_water <- df %>%
  filter(buyer_city%in%water_settlements) %>%
  filter(water_cpv==1 |
         water_buyername==1 |
         water_tendertitle==1)
```

Integrity Risk Indicators

Once the data set has been pruned, we identify the integrity risk indicators at the contract level.



Pillar Calculations

WIRI Components by Locality/Year (1 of 3)

settlement_name	year	count_total	count_inv	count_op	count_int	count_pipe
Mombasa	2018	20	10	3	2	1
Mombasa	2019	5	3	1		1
Nairobi	2012	2	2			1
Nairobi	2014	1	1			
Nairobi	2015	3	2	1		
Nairobi	2016	16	15	1	3	5
Nairobi	2017	25	17	5	3	
Nairobi	2018	303	129	106	19	5
Nairobi	2019	193	68	79	17	4
Nairobi	2020	8		8	1	
Nakuru	2019	89	89			
Nyeri	2017	1	1			
Nyeri	2018	16	15	1		5
Nyeri	2020	10	10			10
Siaya	2018	12	11	1		1
Siaya	2019	24	23			9

Pillar Calculations

WIRI Components by Locality/Year (2 of 3)

settlement_name	year	avg_int_all	avg_int_inv	avg_int_op	avg_int_inter	avg_int_pipe
Mombasa	2018	0.67	0.66	0.67	0.70	0.60
Mombasa	2019	0.68	0.70	0.60		0.60
Nairobi	2012	0.50	0.50			0.60
Nairobi	2014	0.30	0.30			
Nairobi	2015	0.73	0.80	0.60		
Nairobi	2016	0.69	0.72	0.20	0.90	0.86
Nairobi	2017	0.66	0.66	0.66	0.73	
Nairobi	2018	0.71	0.70	0.68	0.63	0.75
Nairobi	2019	0.67	0.68	0.65	0.64	0.75
Nairobi	2020	0.72		0.72	0.40	
Nakuru	2019	0.98	0.98			
Nyeri	2017	0.50	0.50			
Nyeri	2018	0.61	0.61	0.70		0.70
Nyeri	2020	0.70	0.70			0.70
Siaya	2018	0.63	0.63	0.60		0.70
Siaya	2019	0.65	0.66			0.68

Pillar Calculations

WIRI Components by Locality/Year (3 of 3)

settlement_name	year	contract_value_total	contract_value_inv	contract_value_op	contract_value_int	contract_value_pipe
Mombasa	2018	90, 994, 050	79, 423, 323	2, 914, 101	16, 407, 187	3, 475, 650
Mombasa	2019	7, 766, 548	7, 406, 548	130, 000		638, 000
Nairobi	2012	186, 444, 800	186, 444, 800			138, 789, 800
Nairobi	2014	233, 703, 000	233, 703, 000			
Nairobi	2015	13, 994, 105, 570	13, 984, 934, 486	9, 171, 084		
Nairobi	2016	7, 555, 421, 871	7, 554, 952, 831	469, 040	425, 100, 219	6, 198, 344, 114
Nairobi	2017	3, 053, 727, 631	747, 458, 445	16, 266, 291	154, 513, 176	
Nairobi	2018	14, 443, 250, 196	12, 498, 724, 106	494, 072, 921	1, 319, 798, 444	4, 073, 244
Nairobi	2019	2, 557, 753, 539	1, 449, 299, 085	776, 920, 917	109, 263, 543	252, 801, 098
Nairobi	2020	8, 224, 940		8, 224, 940	195, 000	
Nakuru	2019	7, 773, 656, 575	7, 773, 656, 575			
Nyeri	2017	58, 271, 400	58, 271, 400			
Nyeri	2018	5, 833, 030, 687	5, 821, 167, 981	11, 862, 706		22, 252, 576
Nyeri	2020	32, 948, 938	32, 948, 938			32, 948, 938
Siaya	2018	37, 730, 236	36, 855, 236	875, 000		787, 485
Siaya	2019	86, 884, 333	85, 683, 833			69, 900, 670

Pillar Calculations

Calculate Integrity Scores 100-ratio of bribery admissions All contract values are adjusted by PPP.

```
df<-df %>%
  mutate(total_pipe_valueinUSD = contract_value_pipe / bf_wb_ppp,
         contract_value_total_IUSD = contract_value_total / bf_wb_ppp,
         contract_value_inv_total_IUSD = contract_value_inv / bf_wb_ppp,
         contract_value_op_total_IUSD = contract_value_op / bf_wb_ppp,
         contract_value_int_total_IUSD = contract_value_int / bf_wb_ppp)
```

If pipe length data is available, a pipeline investment integrity score $\boxed{pipe_{i}nt}$ can be calculated. For the Kenya, this is a missing value.

Survey Data

All contract values are adjusted by PPP.

```
df <- left_join(df, survey_data)

# Add WIN survey data (if available)
# df <- left_join(df, win_survey)

df <- df %>%
  mutate(
         cui_bribery=(bribes/n)*100,
         # If WIN Survey is available:
         # cui_bribery=(cui_bribery+win_score)/2,
         cui_bribery_int = (100-cui_bribery)/100,
         )
```

If pipe length data is available, a pipeline investment integrity score $\boxed{\text{win}_s urvey}$ can be calculated.

Pillar Weights

WIRI Pillar Weights

column_name	missing_x	number_of_rows	rate	weight
avg_cri_inv_int_100	10	96	0.90	0.35
pipe_int	63	96	0.34	0.13
avg_cri_op_int_100	11	96	0.89	0.35
avg_cri_inter_int_100	58	96	0.40	0.15
cui_survey_int	92	96	0.04	0.02

Note: These weights are taken from the UNESCO Working Paper.

Cross-Sectional WIRI

The Investments Pillar is calculated as the average integrity score for W&S investment contracts and pipeline investments (when available).

The Investments Pillar

```
inv_int = df %>%
  group_by(settlement_name) %>%
  summarise(avg_int_inv_int_100 =
  mean(avg_int_inv_int_100, na.rm=TRUE),
            pipe_int=mean(pipe_int, na.rm=TRUE))

inv_int$wiri_inv <-rowMeans(inv_int [,c("avg_int_inv_int_100", "pipe_int")], na.rm=TRUE)
```

Cross-Sectional WIRI

The Investments Pillar is calculated as the average integrity score for W&S investment contracts and pipeline investments (when available) by settlement.

The Interactions Pillar

```
# Survey Data
df$cui_survey_int <- df$cui_bribery_int

# Procurement Data
cui_int <- df %>%
  group_by(settlement_name) %>%
  summarise(avg_int_inter_int_100=mean(avg_int_inter_int_100, na.rm=TRUE),
            cui_survey_int=mean(cui_survey_int , na.rm=TRUE))

cui_int$wiri_cui <-rowMeans(cui_int [,c("avg_int_inter_int_100",
                                        "cui_survey_int")], na.rm=TRUE)
```

Cross-Sectional WIRI

The Operations Pillar is calculated as the average integrity score for W&S operations (maintenance) contracts by settlement.

The Operations Pillar

```
ops_int <- df %>%
  group_by(settlement_name) %>%
  summarise(avg_int_op_int_100=mean(avg_int_op_int_100, na.rm=TRUE))

ops_int$wiri_ops<-rowMeans(ops_int[,c("avg_int_op_int_100")], na.rm=TRUE)
```

Cross-Sectional WIRI

The Investments Pillar is calculated as the average integrity score for W&S investment contracts and pipeline investments (when available) by settlement.

Cross-sectional WIRI

```
WIRI<-left_join(cui_int, inv_int)
WIRI<-left_join(WIRI,ops_int)

# Missing values are penalized as 0
WIRI<-WIRI %>%
  mutate(wiri_inv=ifelse(is.na(wiri_inv),0,wiri_inv),
         wiri_ops=ifelse(is.na(wiri_ops),0,wiri_ops),
         wiri_cui=ifelse(is.na(wiri_ops),0,wiri_cui)
         )

# Each Pillar is Weighted
WIRI$WIRI <- rowWeightedMeans(as.matrix(WIRI[,c("wiri_inv", "wiri_ops", "wiri_cui")]),
                              pillar_weights, na.rm = T)
```

Time Series WIRI

Time-series WIRI

```
WIRI_ts <- df

#Investment
WIRI_ts$wiri_inv_ts <-rowMeans(WIRI_ts [,c("avg_int_inv_int_100", "pipe_int")],
                               na.rm=TRUE)

#Operations
WIRI_ts$wiri_ops_ts <-rowMeans(WIRI_ts[,c("avg_int_op_int_100")],
                               na.rm=TRUE)

#Interactions
# For One Survey
WIRI_ts$cui_survey_int <- WIRI_ts$cui_bribery_int
# If more than one survey:
# WIRI_ts£cui_survey_int <-rowMeans(WIRI_ts [,c("cui_afrobarometer_bribery_int",
#                                               "cui_gcb_bribery_int")], na.rm=TRUE)
WIRI_ts$wiri_cui_ts <-rowMeans(WIRI_ts [,c("avg_int_inter_int_100", "cui_survey_int")],
                               na.rm=TRUE)
```

Time Series WIRI

Time-series WIRI

```
# Missing values are penalized as 0
WIRI_ts<-WIRI_ts %>%
  mutate(wiri_inv_ts=ifelse(is.na(wiri_inv_ts),0,wiri_inv_ts),
         wiri_ops_ts=ifelse(is.na(wiri_ops_ts),0,wiri_ops_ts),
         wiri_cui_ts=ifelse(is.na(wiri_cui_ts),0,wiri_cui_ts)
         )

# Each Pillar is Weighted
WIRI_ts$WIRI_ts <-rowWeightedMeans(
  as.matrix(WIRI_ts[,c("wiri_inv_ts", "wiri_ops_ts", "wiri_cui_ts")]),
  pillar_weights, na.rm = T)
```

Example Results

Time-Series WIRI