

NER_2019_RIMA

Background and Introduction

Niger 2019 Resilience Index Measurement and Analysis (RIMA) captures the food security and resilience capacity of households in Niger in 2019. The purpose of this study is to explain the role of key variables in determining the current level of resilience of the population in Niger and to provide decision-makers with actionable information to inform response planning. This analysis is based on 2011 household data obtained from the National Survey of Household Living Conditions and Agriculture (or ECVMA, according to the French acronym), as part of the Living Standards Measurement Study - Integrated Surveys on Agriculture (LSMS-ISA) in collaboration with the Niger National Institute of Statistics (NIS). Data are representative at the national- and urban/rural level.

Data quality check and Anonymization procedure

First we read in the dataset and perform basic data quality checks as described in the **SDC protocol**.

In most cases, the following data quality checks was conducted:

1. All personal identifying information (i.e. direct identifiers) have been removed.
2. The dataset does not contain any sensitive information which if disclosed would cause harm to the data subjects.
3. The variables names are labeled correctly.
4. Missing (or suppressed) values are coded as “NA”.
5. Reasonable values were confirmed for all numeric variables during cleaning.
6. There are no duplicates.
7. No dataset contain columns of all missing variables.

Read in the dataset

```
# load data
dd <- "/Users/shang/Food and Agriculture Organization/Anidi, Oluwakayode (OCS) - FAM/ESA_RIMA/Niger/NER/
data.files <- list.files(dd)

for(i in 1:length(data.files)){
  assign(data.files[i],
        read_dta(file = paste0(dd, "/", data.files[i]))
  )
}
```

The community_variables.dta dataset is the same as the household_ag_production.dta dataset. so community_variables.dta is removed.

```
remove(community_variables.dta)
```

There are sensitive information including latitude, longitude and altitude of the household. These variables are removed too.

```
household_characteristics.dta <- household_characteristics.dta %>%  
  select(-latitude, -longitude, -altitude)  
  
household_essential_practices.dta <- household_essential_practices.dta %>%  
  select(-altitude)
```

Then we join all the datasets together

```
# join the dataset together  
dfs <- Filter(function(x) is(x, "data.frame"), mget(ls()))  
  
df_all <- reduce(dfs, left_join)  
  
## Joining, by = c("village_code", "cluster", "hhid")  
## Joining, by = c("village_code", "cluster", "hhid")  
  
## Joining, by = c("village_code", "cluster", "hhid", "region_code")  
  
## Joining, by = c("village_code", "cluster", "hhid")  
## Joining, by = c("village_code", "cluster", "hhid")  
  
## Joining, by = c("village_code", "cluster", "hhid", "other")  
  
## Joining, by = c("village_code", "cluster", "hhid")  
## Joining, by = c("village_code", "cluster", "hhid")  
  
## Joining, by = c("village_code", "cluster", "hhid", "food_for_work")
```

Then the empty columns and rows are removed.

```
df_all <- df_all %>%  
  remove_empty(which = c("rows", "cols"))
```

```
d_org <- df_all
```

```
d <- d_org
```

Measure the risk in the dataset and apply disclosure limitation methods

We consider two types of disclosure scenarios: 1) matching to other publicly available datasets and 2) spontaneous recognition.

To do this, we create the sdc object. The identified key variables which might be in public datasets are

```
keyvars <- c("region_code",
"department_code",
"commune_code",
"village_code",
"cluster",
"hh_main_occupation",
"hh_educ",
"agehead",
"hhsizes")
cbind(keyvars,var_label(d[,keyvars]))
```

```
##                keyvars
## region_code    "region_code"    NULL
## department_code "department_code" NULL
## commune_code   "commune_code"    NULL
## village_code   "village_code"    "village code"
## cluster        "cluster"         "cluster"
## hh_main_occupation "hh_main_occupation" "Household head main occupation"
## hh_educ         "hh_educ"         "Household head years of education"
## agehead        "agehead"         "Age of household head"
## hhsizes        "hhsizes"         "(max) hhsizes"
```

The variables that may cause spontaneous recognition are in pramvars listed below

```
pramvars <- c("television",
"car",
"improved_toilet",
"roof",
"wall")

d[pramvars] <- lapply(d[pramvars], as.factor)

cbind(pramvars,var_label(d[,pramvars]))
```

```
##                pramvars
## television      "television"    NULL
## car             "car"           NULL
## improved_toilet "improved_toilet" NULL
## roof            "roof"          NULL
## wall            "wall"          NULL
```

The sensitive numerical variables that may lead to spontaneous recognition are numvars and are listed below

```
numvars = c("monthly_income", "emergency_food_aid")
cbind(numvars,var_label(d[,numvars]))
```

```
##                numvars
## monthly_income  "monthly_income"
## emergency_food_aid "emergency_food_aid"
##
## monthly_income  "Household total income (monthly, USD)"
## emergency_food_aid "Source of formal transfer, cash+in kind usd"
```

We then measure the initial disclosure risk. The number of observations violating 2-anonymity is 2107 (91.609%).

```
#d[keyvars] <- lapply(d[keyvars], as.factor)
#d[pramvars] <- lapply(d[pramvars], as.factor)

#sdcInitial<- createSdcObj(dat=d, keyVars=keyvars)

sdcInitial<- createSdcObj(dat=d, keyVars=keyvars,pramVars = pramvars,numVars = numvars)
sdcInitial
```

```
## The input dataset consists of 2300 rows and 224 variables.
## --> Categorical key variables: region_code, department_code, commune_code, village_code, cluster, hh_main_occupation, hh_educ, agehead, hhsized
## --> Numerical key variables: monthly_income, emergency_food_aid
## -----
```

```
## Information on categorical key variables:
##
## Reported is the number, mean size and size of the smallest category >0 for recoded variables.
## In parenthesis, the same statistics are shown for the unmodified data.
## Note: NA (missings) are counted as separate categories!
```

##	Key Variable	Number of categories	Mean size
##	region_code	2 (2)	1150.000 (1150.000)
##	department_code	4 (4)	575.000 (575.000)
##	commune_code	4 (4)	575.000 (575.000)
##	village_code	92 (92)	25.000 (25.000)
##	cluster	94 (94)	24.468 (24.468)
##	hh_main_occupation	13 (13)	191.583 (191.583)
##	hh_educ	13 (13)	176.923 (176.923)
##	agehead	73 (73)	31.507 (31.507)
##	hhsized	20 (20)	115.000 (115.000)
##	Size of smallest (>0)		
##		1148 (1148)	
##		551 (551)	
##		551 (551)	
##		3 (3)	
##		3 (3)	
##		1 (1)	
##		1 (1)	
##		1 (1)	
##		1 (1)	

```
## -----
```

```
## Infos on 2/3-Anonymity:
##
## Number of observations violating
## - 2-anonymity: 2107 (91.609%)
## - 3-anonymity: 2259 (98.217%)
## - 5-anonymity: 2300 (100.000%)
##
## -----
```

```
## Numerical key variables: monthly_income, emergency_food_aid
##
## Disclosure risk is currently between [0.00%; 100.00%]
##
## Current Information Loss:
##   - IL1: 0.00
##   - Difference of Eigenvalues: 0.000%
## -----
```

Recode

We recode the age of household head

```
sdcInitial@manipKeyVars$agehead[sdcInitial@manipKeyVars$agehead >= 0 &
sdcInitial@manipKeyVars$agehead <= 18] <- 18

sdcInitial@manipKeyVars$agehead[sdcInitial@manipKeyVars$agehead > 18 &
sdcInitial@manipKeyVars$agehead <= 20] <- 20

sdcInitial@manipKeyVars$agehead[sdcInitial@manipKeyVars$agehead > 20 &
sdcInitial@manipKeyVars$agehead <= 30] <- 30

sdcInitial@manipKeyVars$agehead[sdcInitial@manipKeyVars$agehead > 30 &
sdcInitial@manipKeyVars$agehead <= 40] <- 40

sdcInitial@manipKeyVars$agehead[sdcInitial@manipKeyVars$agehead > 40 &
sdcInitial@manipKeyVars$agehead <= 50] <- 50

sdcInitial@manipKeyVars$agehead[sdcInitial@manipKeyVars$agehead > 50 &
sdcInitial@manipKeyVars$agehead <= 65] <- 60

sdcInitial <- topBotCoding(obj = sdcInitial, value = 65,
replacement = 65, kind = 'top', column = 'agehead')
```

Recode the education level of household head

```
sdcInitial@manipKeyVars$hh_educ[sdcInitial@manipKeyVars$hh_educ >= 0 &
sdcInitial@manipKeyVars$hh_educ <= 6] <- 6

sdcInitial@manipKeyVars$hh_educ[sdcInitial@manipKeyVars$hh_educ > 6 &
sdcInitial@manipKeyVars$hh_educ <= 9] <- 9

sdcInitial@manipKeyVars$hh_educ[sdcInitial@manipKeyVars$hh_educ > 9] <- 10
```

Recode is applied to household size.

```
sdcInitial@manipKeyVars$hhsize[sdcInitial@manipKeyVars$hhsize >= 0 &
sdcInitial@manipKeyVars$hhsize <= 5] <- 5

sdcInitial@manipKeyVars$hhsize[sdcInitial@manipKeyVars$hhsize >= 6 &
sdcInitial@manipKeyVars$hhsize <= 10] <- 10
```

```

sdcInitial@manipKeyVars$hhsz[sdcInitial@manipKeyVars$hhsz >=11 &
sdcInitial@manipKeyVars$hhsz <= 15] <- 15

sdcInitial@manipKeyVars$hhsz[sdcInitial@manipKeyVars$hhsz >=16] <- 16

```

```
#Pram
```

```

set.seed(seed)
sdcInitial <- pram(sdcInitial, strata_variables = "region_code", pd = 0.8)
sdcInitial

```

```
## The input dataset consists of 2300 rows and 224 variables.
```

```
## --> Categorical key variables: region_code, department_code, commune_code, village_code, cluster, l
```

```
## --> Numerical key variables: monthly_income, emergency_food_aid
```

```
## -----
```

```
## Information on categorical key variables:
```

```
##
```

```
## Reported is the number, mean size and size of the smallest category >0 for recoded variables.
```

```
## In parenthesis, the same statistics are shown for the unmodified data.
```

```
## Note: NA (missings) are counted as separate categories!
```

##	Key Variable	Number of categories	Mean size
##	region_code	2 (2)	1150.000 (1150.000)
##	department_code	4 (4)	575.000 (575.000)
##	commune_code	4 (4)	575.000 (575.000)
##	village_code	92 (92)	25.000 (25.000)
##	cluster	94 (94)	24.468 (24.468)
##	hh_main_occupation	13 (13)	191.583 (191.583)
##	hh_educ	3 (13)	766.667 (176.923)
##	agehead	7 (73)	328.571 (31.507)
##	hhsz	4 (20)	575.000 (115.000)

```
## Size of smallest (>0)
```

```
## 1148 (1148)
```

```
## 551 (551)
```

```
## 551 (551)
```

```
## 3 (3)
```

```
## 3 (3)
```

```
## 1 (1)
```

```
## 5 (1)
```

```
## 11 (1)
```

```
## 20 (1)
```

```
## -----
```

```
## Infos on 2/3-Anonymity:
```

```
##
```

```
## Number of observations violating
```

```
## - 2-anonymity: 790 (34.348%) | in original data: 2107 (91.609%)
```

```
## - 3-anonymity: 1344 (58.435%) | in original data: 2259 (98.217%)
```

```
## - 5-anonymity: 1886 (82.000%) | in original data: 2300 (100.000%)
```

```
##
```

```
## -----
```

```

## Numerical key variables: monthly_income, emergency_food_aid
##
## Disclosure risk (~100.00% in original data):
##   modified data: [0.00%; 100.00%]
##
## Current Information Loss in modified data (0.00% in original data):
##   IL1: 0.00
##   Difference of Eigenvalues: 0.000%
## -----

## Post-Randomization (PRAM):

## Variable:television

## --> final Transition-Matrix:

##           0           1
## 0 0.9993578 0.0006421921
## 1 0.4917051 0.5082949349

## Variable:car
##
## --> final Transition-Matrix:

##           0
## 0 0.905211

## Variable:improved_toilet
##
## --> final Transition-Matrix:

##           0           1
## 0 0.9951883 0.004811711
## 1 0.4763594 0.523640572

## Variable:roof
##
## --> final Transition-Matrix:

##           0           1
## 0 0.86502847 0.1349715
## 1 0.05478062 0.9452194

## Variable:wall
##
## --> final Transition-Matrix:

##           0           1
## 0 0.9982912 0.00170879
## 1 0.4895683 0.51043169

```

```
##
## Changed observations:

##      variable nrChanges percChanges
## 1    television      2      0.09
## 2         car      0      0.00
## 3 improved_toilet    23      1.00
## 4         roof    182      7.91
## 5         wall      7      0.30

## -----
```

Round numerical key variables `monthly_income` and `emergency_food_aid` to avoid exact match.

```
sdInitial@manipNumVars$monthly_income[sdInitial@manipNumVars$monthly_income <= 1 &
!is.na(sdInitial@manipNumVars$monthly_income)] <-
  round(sdInitial@manipNumVars$monthly_income[sdInitial@manipNumVars$monthly_income <= 1 &
!is.na(sdInitial@manipNumVars$monthly_income)],
        digits = 1)

sdInitial@manipNumVars$monthly_income[sdInitial@manipNumVars$monthly_income > 1 &
!is.na(sdInitial@manipNumVars$monthly_income)] <-
  round(sdInitial@manipNumVars$monthly_income[sdInitial@manipNumVars$monthly_income > 1 &
!is.na(sdInitial@manipNumVars$monthly_income)],
        digits = 0)

sdInitial@manipNumVars$emergency_food_aid<-
  round(sdInitial@manipNumVars$emergency_food_aid,
        digits = 0)
```

Local suppression

Finally, local suppression is applied to ensure 2-k anonymity.

```
sdInitial <- localSuppression(sdInitial, k = 2,
                             importance = c(6,7,8,9,1,2,3,4,5))
sdInitial
```

```
## The input dataset consists of 2300 rows and 224 variables.
## --> Categorical key variables: region_code, department_code, commune_code, village_code, cluster, l
## --> Numerical key variables: monthly_income, emergency_food_aid
## -----
```

```
## Information on categorical key variables:
##
## Reported is the number, mean size and size of the smallest category >0 for recoded variables.
## In parenthesis, the same statistics are shown for the unmodified data.
## Note: NA (missings) are counted as separate categories!
```



```

##      Key Variable Number of categories      Mean size
##      region_code          2 (2) 1150.000 (1150.000)
##      department_code       4 (4)  575.000 (575.000)
##      commune_code          4 (4)  575.000 (575.000)
##      village_code         93 (92)   24.696 (25.000)
##      cluster              95 (94)   24.426 (24.468)
##      hh_main_occupation    12 (13)  197.182 (191.583)
##      hh_educ               4 (13)  758.667 (176.923)
##      agehead               8 (73)  292.571 (31.507)
##      hhsize                5 (20)  482.000 (115.000)
##      Size of smallest (>0)
##      1148 (1148)
##      551 (551)
##      551 (551)
##      1 (3)
##      3 (3)
##      1 (1)
##      2 (1)
##      4 (1)
##      1 (1)

## -----

## Infos on 2/3-Anonymity:
##
## Number of observations violating
## - 2-anonymity: 0 (0.000%) | in original data: 2107 (91.609%)
## - 3-anonymity: 301 (13.087%) | in original data: 2259 (98.217%)
## - 5-anonymity: 1216 (52.870%) | in original data: 2300 (100.000%)
##
## -----

## Numerical key variables: monthly_income, emergency_food_aid
##
## Disclosure risk (~100.00% in original data):
##   modified data: [0.00%; 99.74%]
##
## Current Information Loss in modified data (0.00% in original data):
##   IL1: 0.00
##   Difference of Eigenvalues: 0.000%
## -----

## Post-Randomization (PRAM):

## Variable:television

## --> final Transition-Matrix:

##      0      1
## 0 0.9993578 0.0006421921
## 1 0.4917051 0.5082949349

```

```

## Variable:car
##
## --> final Transition-Matrix:

##          0
## 0 0.905211

## Variable:improved_toilet
##
## --> final Transition-Matrix:

##          0          1
## 0 0.9951883 0.004811711
## 1 0.4763594 0.523640572

## Variable:roof
##
## --> final Transition-Matrix:

##          0          1
## 0 0.86502847 0.1349715
## 1 0.05478062 0.9452194

## Variable:wall
##
## --> final Transition-Matrix:

##          0          1
## 0 0.9982912 0.00170879
## 1 0.4895683 0.51043169

##
## Changed observations:

##          variable nrChanges percChanges
## 1      television         2         0.09
## 2              car         0         0.00
## 3 improved_toilet        23         1.00
## 4              roof       182         7.91
## 5              wall         7         0.30

## -----

## Local suppression:

##          KeyVar | Suppressions (#) | Suppressions (%)
##          region_code |          0 |          0.000
##          department_code |          0 |          0.000
##          commune_code |          0 |          0.000
##          village_code |         28 |          1.217
##          cluster |          4 |          0.174
##          hh_main_occupation |        130 |          5.652
##          hh_educ |         24 |          1.043
##          agehead |        252 |         10.957
##          hhsize |        372 |         16.174

```

Save anonymized data

The variable labels are added back. Value labels are also added if the variable is altered in the anonymization process. The anonymized data is saved in several format for different software.

```
## Warning in if (cc != class(v_p)) {: the condition has length > 1 and only the  
## first element will be used
```

```
## Warning in if (cc != class(v_p)) {: the condition has length > 1 and only the  
## first element will be used
```

```
## Warning in if (cc != class(v_p)) {: the condition has length > 1 and only the  
## first element will be used
```

```
## Warning in if (cc != class(v_p)) {: the condition has length > 1 and only the  
## first element will be used
```

```
## Warning in if (cc != class(v_p)) {: the condition has length > 1 and only the  
## first element will be used
```