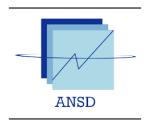
REPUBLIQUE DU SENEGAL



Un Peuple - Un But - Une Foi

Agence nationale de la Statistique et de la démographie



Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye



Projet statistique sur R : Evaluation

Rédigé par :

Sous la supervision de :

RASAMOELINA Nihaviana Albert Paulinah Élève ISE1 ECO

M. Aboubacar HEMA

Research analyst à IFPRI

Année académique 2024-2025

Introduction

Ce rapport présente une analyse statistique complète réalisée dans le cadre de l'examen de projet statistique sur R pour ISE année 1. L'objectif est d'analyser un ensemble de données relatives liées aux déplacements forcés au Sud-Soudan en 2023. Ce rapport suit les instructions fournies dans le sujet d'examen et comprend une analyse de consistance des données, le calcul d'indicateurs clés, des analyses socio-démographiques et des visualisations spatiales.

```
      used (Mb) gc trigger (Mb) max used (Mb)

      Ncells 1297381 69.3
      2226337 118.9
      2226337 118.9

      Vcells 2932261 22.4
      8388608 64.0 4162680 31.8
```

I. Importation et Analyse de consistance des bases

1. Importation des jeux de données

La base Base_Individus.dta compte 144 variables pour nrow(base_individus) observations. Alors que la base Base_Principale.dta compte 1312 variables pour nrow(base_principale) observations.

2. Analyse de consistance des bases de données

Cette section vise à vérifier la qualité des données importées et à appliquer les transformations nécessaires pour assurer leur cohérence avant l'analyse. Nous utiliserons principalement les packages janitor et tidyverse pour cette tâche.

2.1. Analyse préliminaire des bases de données

Commençons par examiner les dimensions et les caractéristiques générales de nos bases de données.

Dimensions de la base individus: 22092 144

Dimensions de la base principale: 3058 1312

Nombre de valeurs manquantes par colonne (top 10) - base individus:

$\mathtt{HH} _ \mathtt{00b} _ \mathtt{month}$	HH_00b_year	HH_11b	Dis_18	Dis_15	Dis_12
22089	22077	22025	21978	21954	21879
ID_04	HH_16	ID_06	Dis_06		
21852	21760	21710	21692		

Nombre de valeurs manquantes par colonne (top 10) - base principale:

Pourcentage global de valeurs manquantes - base individus: 64.64 %

Pourcentage global de valeurs manquantes - base principale: 65.55 %

2.2 Nettoyage des bases de données avec janitor

Le package janitor offre des fonctions utiles pour nettoyer les bases de données. Appliquons quelques-unes de ces fonctions à nos bases.

Colonnes presque vides (>90% NA) dans la base individus: 36

```
[1] "HH_00a_year" "HH_00a_month" "HH_00b_year" "HH_00b_month" "HH_11b" [6] "ID 04" "ID 06" "HH 07b" "HH 10name16" "HH 10name17"
```

Colonnes presque vides (>90% NA) dans la base principale: 589

```
[1] "HH_31b1" "HH_31b2" "HH_31b3" "HH_31b4" "HH_31b5" "HH_31b6" [7] "HH_31b7" "HH_31b8" "HH_31b9" "HH_31b10"
```

Lignes presque vides (>90% NA) dans la base individus: 0

Lignes presque vides (>90% NA) dans la base principale: 0

Exemple de noms de colonnes avant/après nettoyage - base individus:

Exemple de noms de colonnes avant/après nettoyage - base principale:

2.3 Suppression des colonnes et lignes inutiles

Sur la base de l'analyse ci-dessus, nous pouvons décider de supprimer certaines colonnes ou lignes qui contiennent principalement des valeurs manquantes. Cependant, il est important de ne pas supprimer des données qui pourraient être pertinentes pour l'analyse.

Dimensions de la base individus avant: 22092 144

Dimensions de la base individus après: 22092 116

Dimensions de la base principale avant: 3058 1312

Dimensions de la base principale après: 3058 873

Dimensions de la base individus après suppression des lignes: 22092 116

Dimensions de la base principale après suppression des lignes: 3058 873

2.4. Traitement des doublons

Vérifions s'il existe des doublons dans les bases de données.

```
Nombre de lignes avec des IDs dupliqués dans la base individus : 2971
```

Nombre de lignes dupliquées dans la base principale : 0

2.5. Cohérence entre les bases de données

Vérifions la cohérence entre les bases de données Ind et Prin, notamment la correspondance des identifiants.

```
Nombre d'IDs dans Ind qui n'existent pas dans la base principale: 0
```

Nombre d'IDs dans Prin qui n'existent pas dans la base individus: 0

2.6. Traitement des valeurs aberrantes

Examinons certaines variables numériques clés pour détecter d'éventuelles valeurs aberrantes.

Nombre de valeurs aberrantes pour l'âge: 0

2.7. Standardisation des types de données

S'assurer que les variables ont le type approprié.

2.8. Résumé du nettoyage

Résumé des modifications apportées:

```
1. Colonnes supprimées (>95% NA) - base individus: 28
```

```
2. Colonnes supprimées (>95% NA) - base principale: 439
```

```
3. Lignes supprimées (>95% NA) - base individus: 0
```

4. Lignes supprimées (>95% NA) - base principale: 0

```
Pourcentage final de valeurs manquantes - base individus: 56.83 %
```

Pourcentage final de valeurs manquantes - base principale: 48.97 %

2.9. Finalisation des bases de données nettoyées

Remplaçons les bases de données originales par les versions nettoyées pour la suite de l'analyse.

```
used (Mb) gc trigger (Mb) max used (Mb)
Ncells 1398835 74.8 2226337 118.9 2226337 118.9
Vcells 15597224 119.0 36073949 275.3 29994958 228.9
```

3. Exploration des données

3.1. Structure des bases de données

Pour mieux comprendre les données, examinons la structure des deux bases et les principales variables disponibles.

[1] 22092 116

[1] 3058 873

3.2. Variables catégorielles importantes

Examinons les variables catégorielles importantes comme le sexe, le statut marital, etc.

Male Female <NA> 10821 11251 20

Male Female <NA> 48.98153178 50.92793772 0.09053051

ted 260 NA> 580	<	S	43 ed	4	n-forma) r	married 660 widowe: 423		wid	5631 vorced 157		monoga
966 NA>	separa 1.1768 < 4.3110	1	06 ed	194640	never	3 r	married 9875068 widowe: 910193	2. ow or	wid	narried 1888647 Lvorced 7106645	25.4 di	monoga
	r 2	answer	use to	ow refu 2		do	no 956	7		yes 7021 <na> 7101</na>		
		answer 9053051			n't kno 5431830		no 393	013036	36.	<na></na>	1.78073 2.14285	
13 5 28 52	12 198 27 1	11 180 24 1	10 221 23 2	9 220 22 15	8 370 21 10	7 369 20 1	6 506 19 2	5 641 18 4	4 869 17 5	3 950 16 12	2 1098 15 9 <na></na>	1 1213 14 3 29 40

1	2	3	4	5	6
5.490675358	4.970124932	4.300199167	3.933550607	2.901502806	2.290421872
7	8	9	10	11	12
1.670287887	1.674814412	0.995835597	1.000362122	0.814774579	0.896252037
13	14	15	16	17	18
0.022632627	0.013579576	0.040738729	0.054318305	0.022632627	0.018106102
19	20	21	22	23	24
0.009053051	0.004526525	0.045265254	0.067897882	0.009053051	0.004526525
27	28	29	<na></na>		
0.004526525	0.235379323	0.181061018	68.327901503		

Government or Public	UN or NGO
1328	5212
Religious or faith-based organization	Community
167	17
Private	Other: specify
282	5
Don't know	<na></na>
10	15071

Government or Public UN or NGO 6.01122578 23.59225059 Religious or faith-based organization Community 0.75592975 0.07695093 Private Other: specify 1.27648017 0.02263263 Don't know <NA> 68.21926489 0.04526525

3.3. Statistiques descriptives des variables numériques

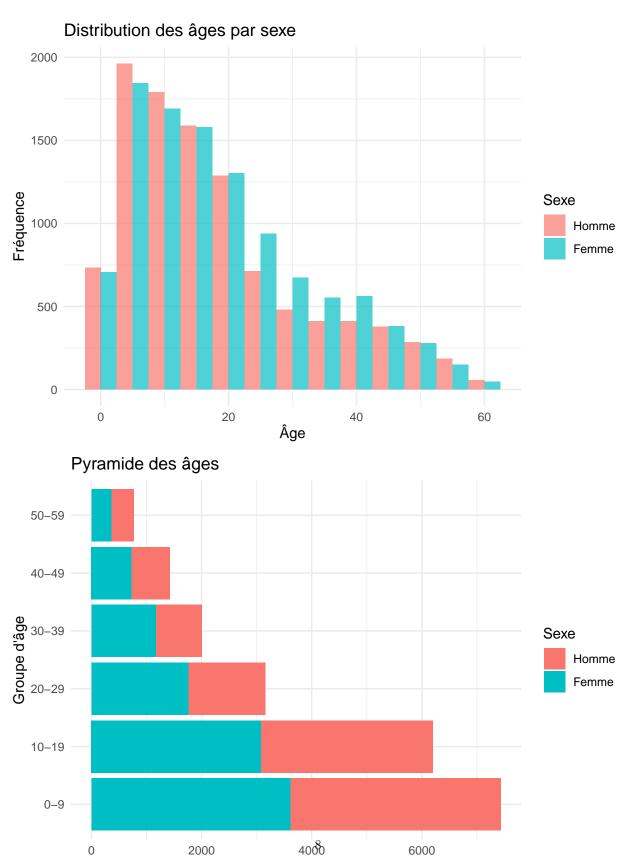
ageYears 1103

ageYears

Min. : 1.00 1st Qu.: 7.00 Median :15.00 Mean :18.55 3rd Qu.:26.00 Max. :60.00 NA's :1103

II. Analyse des données socio-démographiques

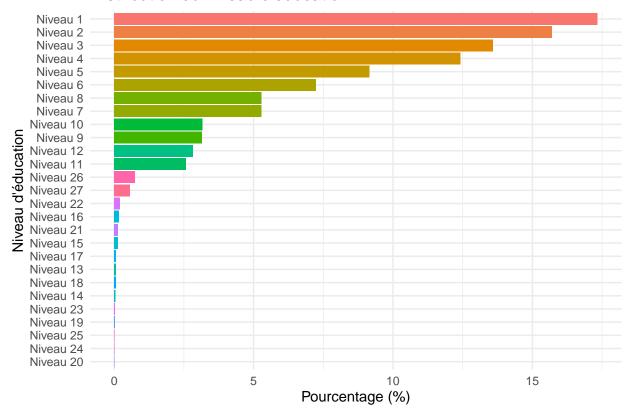
1. Répartition par âge et sexe

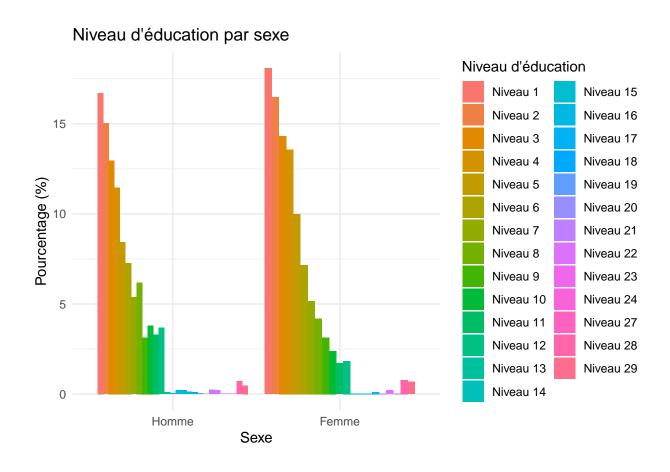


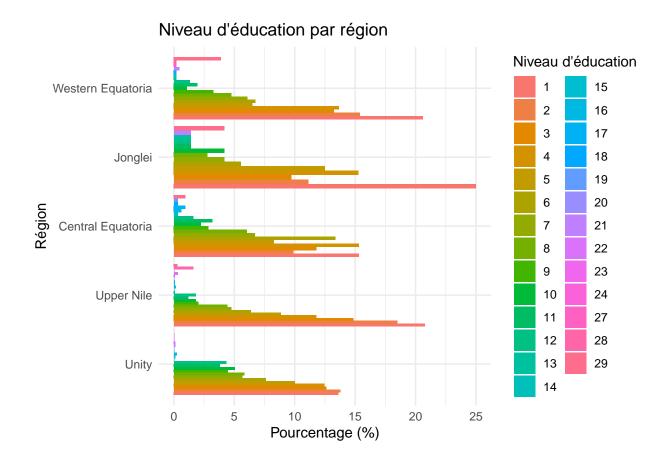
Population

##2. Analyse du niveau d'éducation

Distribution du niveau d'éducation





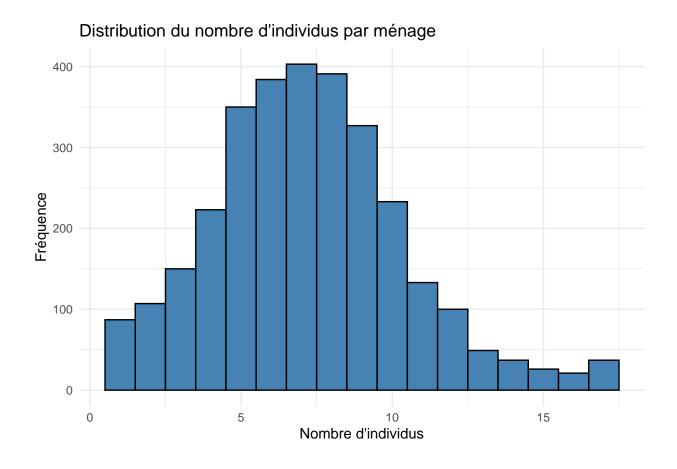


III Calcul et analyse du Crowding Index (Indice d'affluence)

Le Crowding Index ou indice d'affluence est le rapport entre le nombre de personnes vivant dans un ménage et le nombre de pièces disponibles (à l'exclusion de la cuisine et des couloirs). Cet indice est un indicateur important des conditions de vie des ménages.

1. Calcul du nombre d'individus par ménage

Min. 1st Qu. Median Mean 3rd Qu. Max. 1.000 5.000 7.000 7.224 9.000 17.000

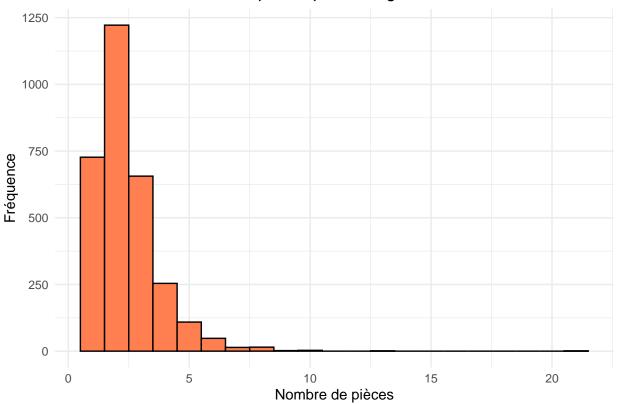


2. Statistiques descriptives sur le nombre de pièces

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 1.000 2.000 2.000 2.388 3.000 21.000 6

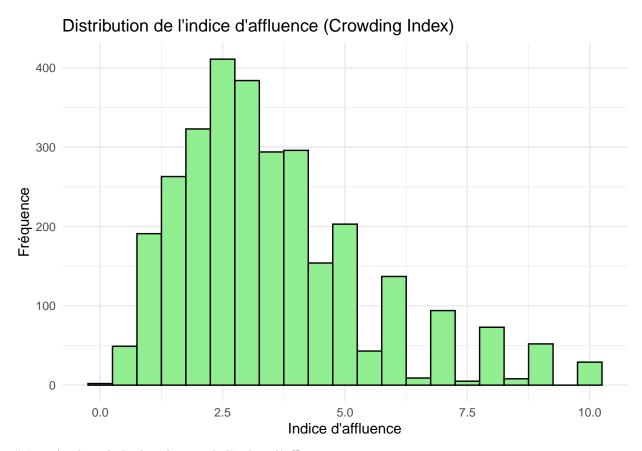
Nombre de valeurs non finies exclues : 6

Distribution du nombre de pièces par ménage

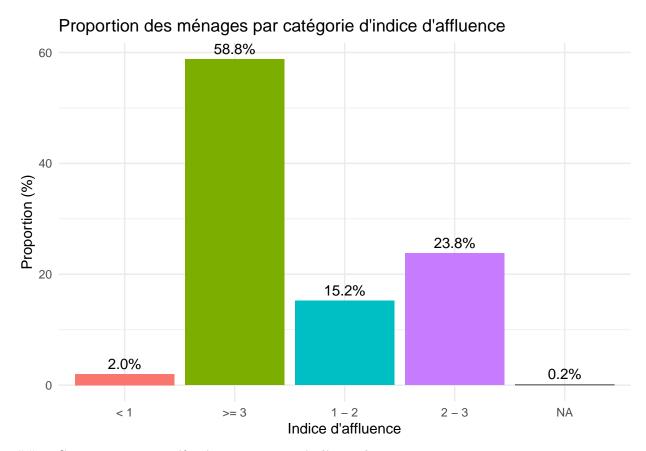


3. Calcul de l'indice d'affluence (Crowding Index)

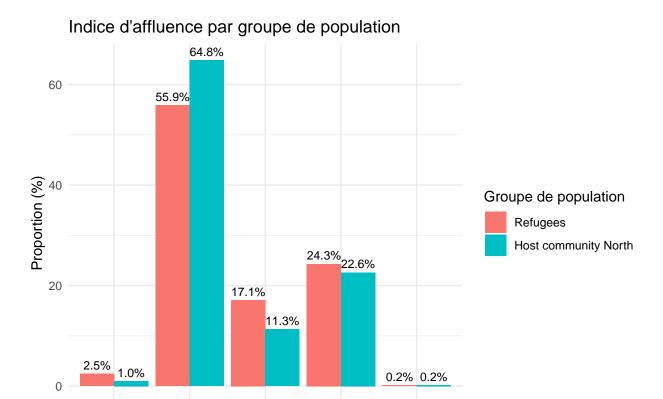
Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0.1905 2.0000 3.0000 3.5997 4.5000 15.0000 6



##4. Analyse de la distribution de l'indice d'affluence



##5. Comparaison entre réfugiés et communautés d'accueil



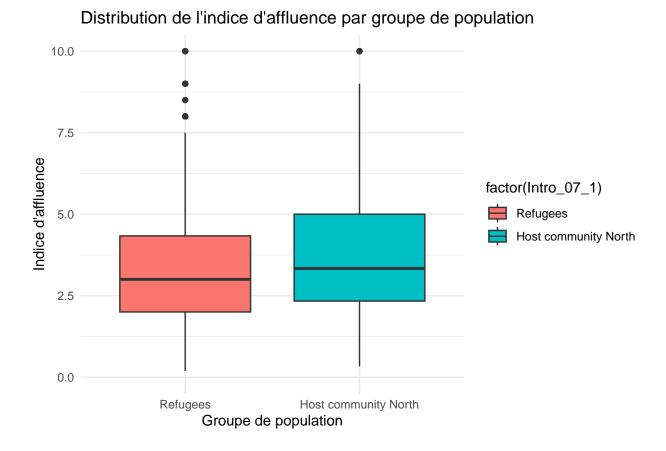
2 – 3

NA

1 - 2 Indice d'affluence

< 1

>= 3



IV Analyse de la sécurité alimentaire des déplacés internes

1. Score de Consommation Alimentaire (SCA)

Le Score de Consommation Alimentaire (SCA) est un indicateur proxy développé par le Programme Alimentaire Mondial pour mesurer la sécurité alimentaire des ménages. C'est un score composite basé sur la diversité alimentaire, la fréquence de consommation et l'importance nutritionnelle relative des groupes d'aliments consommés.

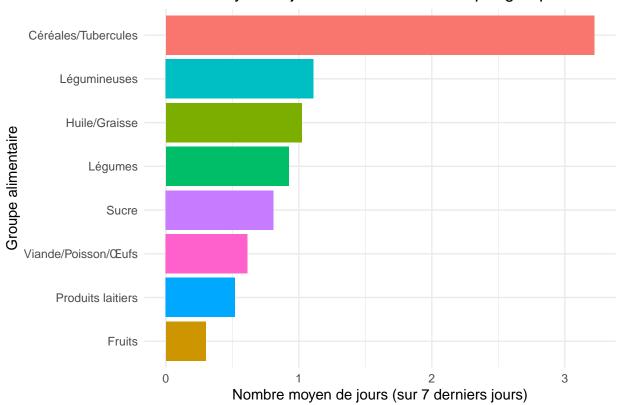
1.1. Analyse descriptive des variables composant le SCA

Table 9: Statistiques descriptives des variables du SCA

variable	moyenne	mediane	type	min	max	valides	manquants
Food_div1	3.2213491	3	NA	0	7	NA	NA
$Food_div1_ecart$	NA	NA	2.640259	NA	NA	NA	NA
$Food_div1_n$	NA	NA	NA	NA	NA	3054	4
Food_div2	1.1100917	0	NA	0	7	NA	NA
$Food_div2_ecart$	NA	NA	1.843696	NA	NA	NA	NA
$Food_div2_n$	NA	NA	NA	NA	NA	3052	6
Food_div3	0.5188463	0	NA	0	7	NA	NA
$Food_div3_ecart$	NA	NA	1.404870	NA	NA	NA	NA

variable	moyenne	mediane	type	min	max	valides	manquants
Food_div3_n	NA	NA	NA	NA	NA	3051	7
Food_div4	0.6148294	0	NA	0	7	NA	NA
$Food_div4_ecart$	NA	NA	1.280462	NA	NA	NA	NA
Food_div4_n	NA	NA	NA	NA	NA	3048	10
$Food_div5$	0.9244167	0	NA	0	7	NA	NA
$Food_div5_ecart$	NA	NA	1.746097	NA	NA	NA	NA
$Food_div5_n$	NA	NA	NA	NA	NA	3043	15
Food_div6	0.2994740	0	NA	0	7	NA	NA
$Food_div6_ecart$	NA	NA	1.139148	NA	NA	NA	NA
$Food_div6_n$	NA	NA	NA	NA	NA	3042	16
Food_div7	1.0223170	0	NA	0	7	NA	NA
$Food_div7_ecart$	NA	NA	1.907796	NA	NA	NA	NA
Food_div7_n	NA	NA	NA	NA	NA	3047	11
Food_div8	0.8103222	0	NA	0	7	NA	NA
$Food_div8_ecart$	NA	NA	1.739492	NA	NA	NA	NA
Food_div8_n	NA	NA	NA	NA	NA	3042	16

Nombre moyen de jours de consommation par groupe alimenta



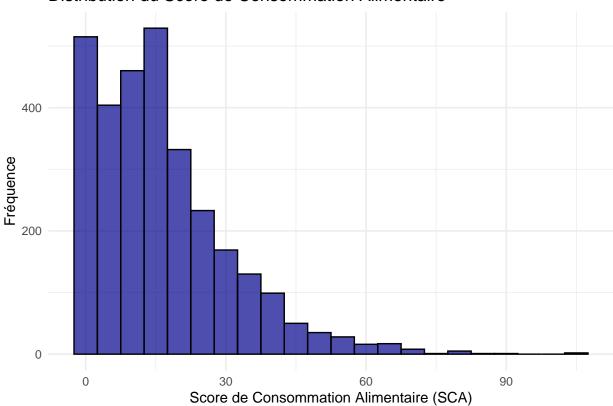
1.2. Calcul du Score de Consommation Alimentaire

Table 10: Poids attribués aux groupes alimentaires pour le calcul du SCA $\,$

groupe_alimentaire	variable	poids
Céréales/Tubercules	Food_div1	2.0
Légumineuses	$Food_div2$	3.0
Produits laitiers	Food_div3	4.0
Viande/Poisson/Œufs	$Food_div4$	4.0
Légumes	$Food_div5$	1.0
Fruits	Food_div6	1.0
Huile/Graisse	$Food_div7$	0.5
Sucre	${\rm Food_div8}$	0.5

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 0.00 6.00 14.00 16.46 23.00 107.00 23





1.3. Catégorisation du SCA selon différents seuils

Table 11: Répartition des ménages selon les seuils SCA 21/35

categorie_sca_21_35	n	proportion
Acceptable	311	10.1700458
Limite	626	20.4708960

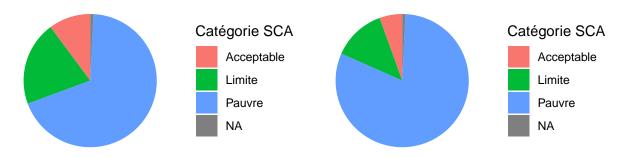
categorie_sca_21_35	n	proportion
Pauvre	2098	68.6069326
NA	23	0.7521256

Table 12: Répartition des ménages selon les seuils SCA 28/42

categorie_sca_28_42	n	proportion
Acceptable	169	5.5264879
Limite	393	12.8515370
Pauvre	2473	80.8698496
NA	23	0.7521256

Répartition selon seuils 21/35

Répartition selon seuils 28/42



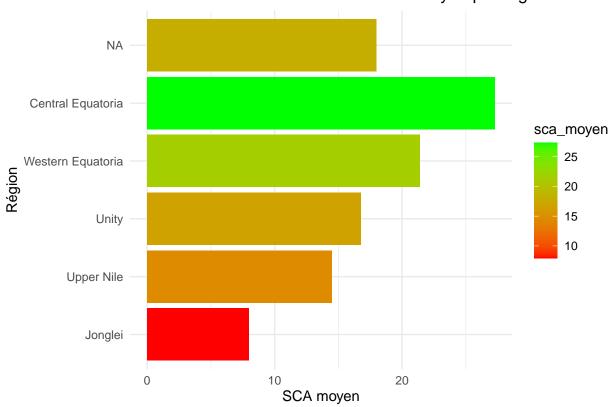
1.4. Représentation spatiale du SCA par région et département

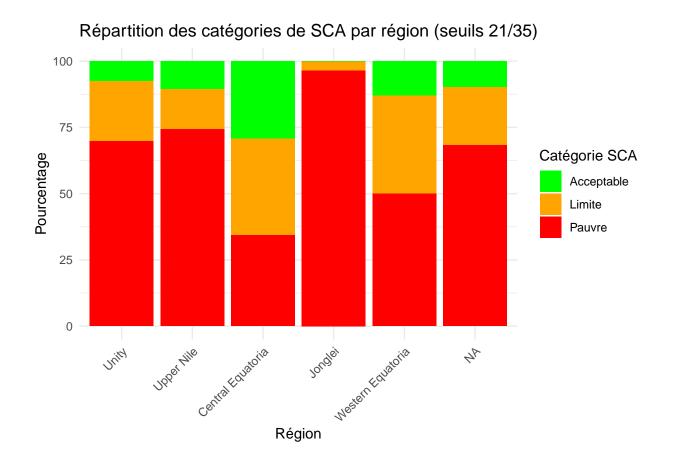
Table 13: Score de Consommation Alimentaire par région

admin1	sca_moyen sca	_median	n	pct_pauvre_21_3	35pct_limite_21_35pct	_acceptable_21_35
Unity	16.8	14.0	1064	69.9	22.6	7.5
Upper Nile	14.5	10.2	1466	74.5	15.1	10.5
Central	27.3	25.0	114	34.5	36.3	29.2
Equatoria						

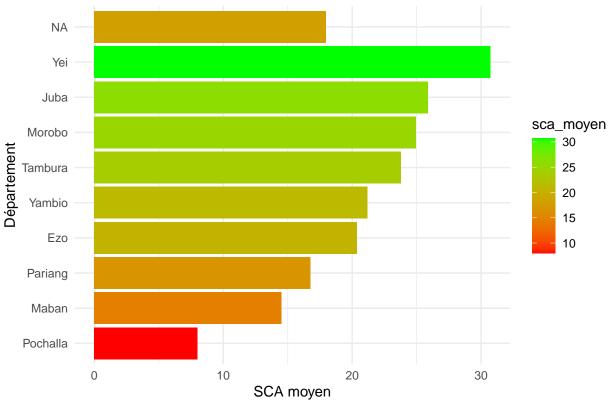
admin1	sca_moyen sca_	_median	n	pct_pauvre_21_3	5 ct_limite_21_35pct_	acceptable_21_35
Jonglei	8.0	7.0	35	96.4	3.6	0.0
Western	21.4	20.5	285	50.2	36.8	13.0
Equatoria						
NA	18.0	14.5	94	68.5	21.7	9.8







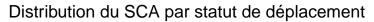


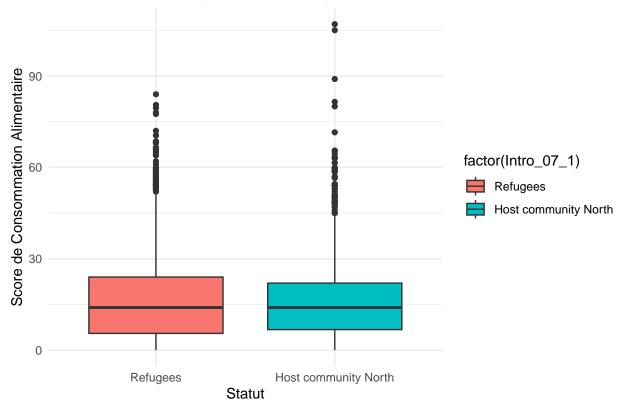


1.5. Analyse du SCA selon le statut de déplacement

Table 14: Score de Consommation Alimentaire par statut de déplacement

Intro_07_1	sca_moye rs c	a_media s c	a_ecart_t;	ype n	pct_pauvre_21pc	<u>t5_limite21_p35_</u>	_acceptable_	21_35
Refugees	16.4	14	14.2	2068	67.5	22.5	10.0	
Host	16.6	14	14.6	990	72.4	16.8	10.8	
community North								

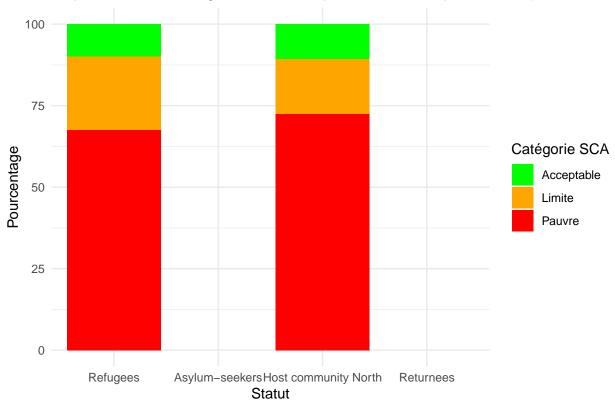




Test ANOVA pour comparer les moyennes du SCA entre les groupes:

Df Sum Sq Mean Sq F value Pr(>F)
factor(Intro_07_1) 1 19 18.53 0.091 0.764
Residuals 3033 620829 204.69
23 observations effacées parce que manquantes





V L'indice réduit des stratégies de survie (rCSI)

L'indice réduit des stratégies de survie (rCSI) est un indicateur qui mesure les comportements d'adaptation que les ménages adoptent lorsqu'ils n'ont pas accès à suffisamment de nourriture. Il est basé sur un ensemble de cinq stratégies de survie communes liées à la consommation alimentaire. Un score plus élevé indique une plus grande insécurité alimentaire.

1. Analyse descriptive des variables qui composent le rCSI

Les cinq stratégies d'adaptation communes utilisées pour calculer le rCSI sont:

- 1. Consommer des aliments moins préférés et moins chers (Food02a)
- 2. Emprunter de la nourriture ou compter sur l'aide de proches (Food05a)
- 3. Limiter la taille des portions au moment des repas (Food06a)
- 4. Réduire le nombre de repas par jour (Food08a)
- 5. Réduire la consommation des adultes pour nourrir les enfants (Food07a)

```
tibble [3,058 x 5] (S3: tbl_df/tbl/data.frame)
$ Food02a: Factor w/ 2 levels "yes","no": 1 2 1 1 1 1 1 1 1 1 ...
..- attr(*, "label")= chr "A member ate less preferred food and less expensive food [Past 30/7 days]"
$ Food05a: Factor w/ 2 levels "yes","no": 1 2 2 1 2 2 1 1 2 2 ...
..- attr(*, "label")= chr "Any member borrowed food/relied on help to get more food...[Past 30/7 days
$ Food06a: Factor w/ 2 levels "yes","no": 1 2 1 1 1 1 1 2 2 ...
```

..- attr(*, "label")= chr "Any member ate a smaller meal than you felt you needed [Past 30/7 days]" \$ Food08a: Factor w/ 2 levels "yes","no": 1 2 1 1 1 1 1 1 2- attr(*, "label")= chr "Any member ate fewer meals in a day [Past 30/7 days]"

\$ Food07a: Factor w/ 2 levels "yes","no": 1 2 1 1 1 1 2 2 1 2 ...

..- attr(*, "label")= chr "Adults ate less to have more food for children under 5? [Past 30/7 days]"

Variable: Food02a

yes no <NA> 2493 563

Variable: Food05a

no <NA> yes 1831 1222

Variable: Food06a

no <NA> yes 2600 454

Variable: Food08a

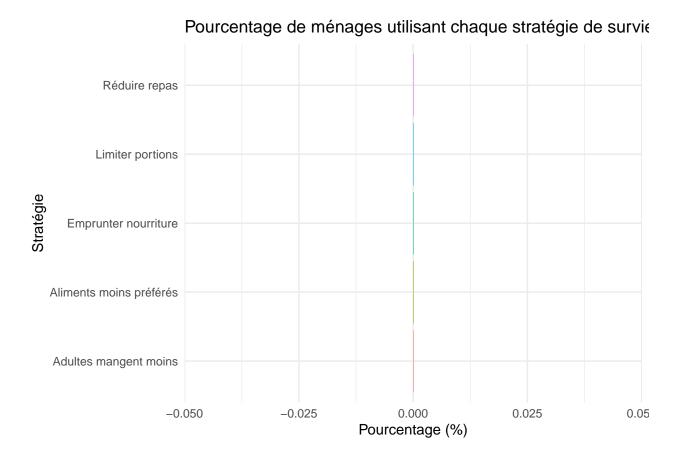
no <NA> yes 2547 510

Variable: Food07a

yes no <NA> 2145 906

Table 15: Statistiques descriptives des variables du rCSI

variable	oui	non	manquants
Food02a n	0	0	2
Food02a pct	0	NA	NA
Food05a_n	0	0	5
Food05a_pct	0	NA	NA
Food06a n	0	0	4
Food06a pct	0	NA	NA
Food08a n	0	0	1
Food08a pct	0	NA	NA
Food07a n	0	0	7
Food07a_pct	0	NA	NA

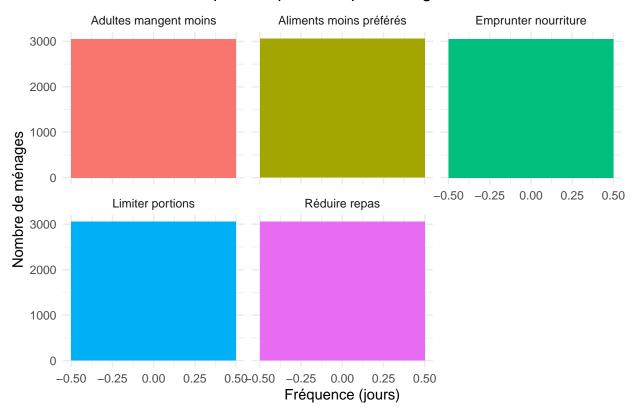


2. Création et analyse de nouvelles variables numériques pour le rCSI

Table 16: Statistiques descriptives des nouvelles variables numériques du rCSI

variable	moyenne	mediane	type	min	max
Food02a_num	0	0	NA	0	0
$Food 02a_num_ecart$	NA	NA	0	NA	NA
Food05a_num	0	0	NA	0	0
$Food05a_num_ecart$	NA	NA	0	NA	NA
Food06a_num	0	0	NA	0	0
$Food06a_num_ecart$	NA	NA	0	NA	NA
Food08a_num	0	0	NA	0	0
$Food08a_num_ecart$	NA	NA	0	NA	NA
Food07a_num	0	0	NA	0	0
$Food07a_num_ecart$	NA	NA	0	NA	NA

Distribution des fréquences pour chaque stratégie



3. Calcul de l'indice réduit des stratégies de survie (rCSI)

[1] 21

Table 17: Poids attribués aux stratégies pour le calcul du rCSI

strategie	variable	poids
Aliments moins préférés	Food02a_num	2
Emprunter nourriture	$Food05a_num$	2
Limiter portions	$Food06a_num$	3
Réduire repas	$Food08a_num$	7
Adultes mangent moins	$Food07a_num$	7

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
0 0 0 0 0 0 15

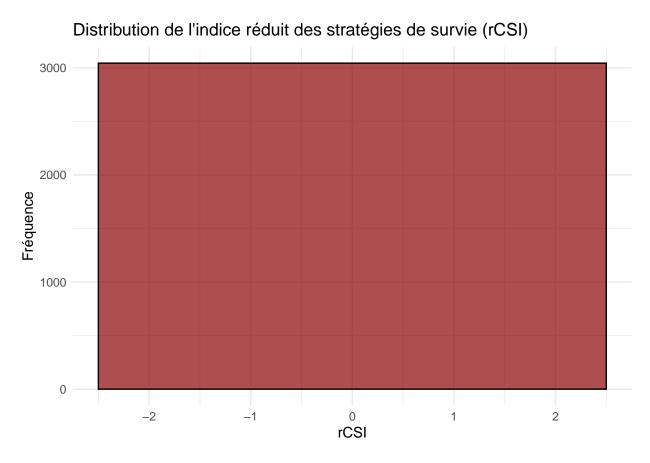
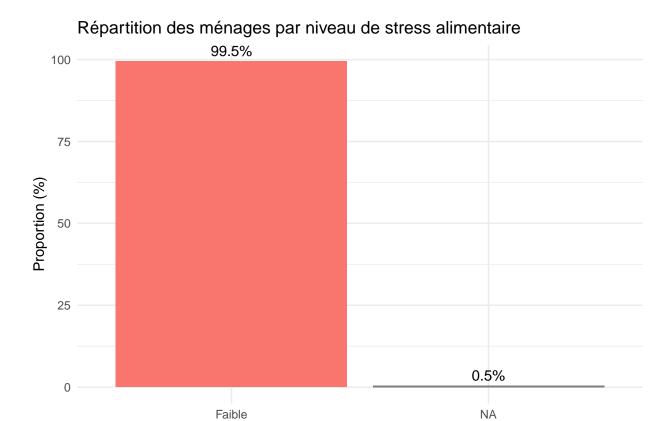


Table 18: Répartition des ménages par niveau de stress alimentaire

categorie_rcsi	n	proportion
Faible	3043	99.5094833
NA	15	0.4905167

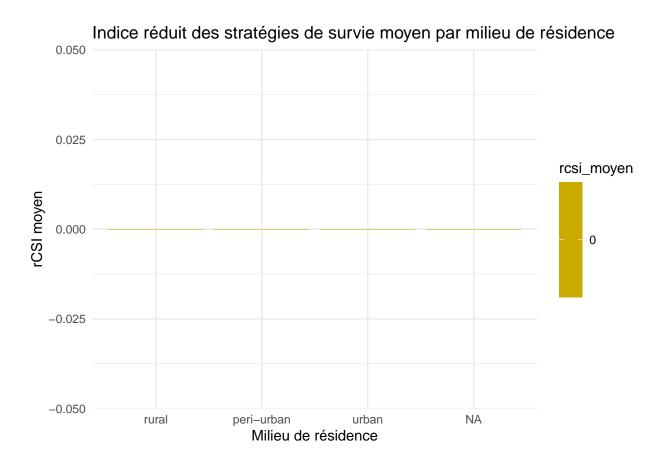


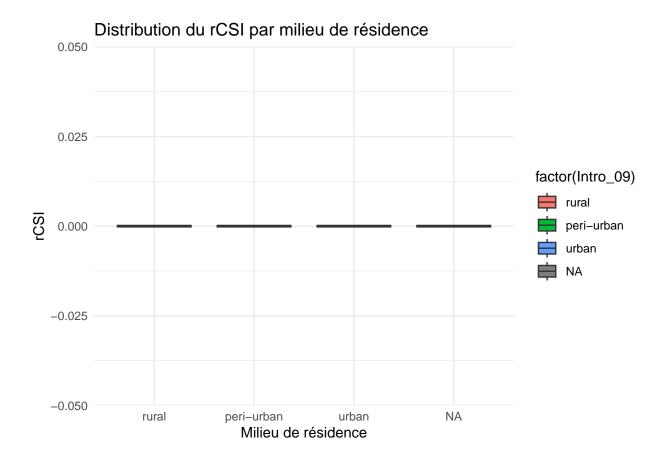
4. Représentation spatiale du rCSI selon le milieu de résidence

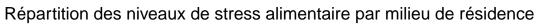
Table 19: rCSI moyen par milieu de résidence

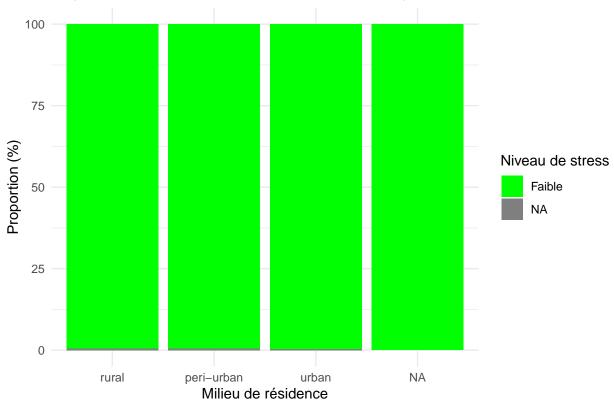
Niveau de stress alimentaire

Intro_09	rcsi_moyen	rcsi_median	rcsi_ecart_type	n
rural	0	0	0	2058
peri-urban	0	0	0	740
urban	0	0	0	258
NA	0	0	0	2

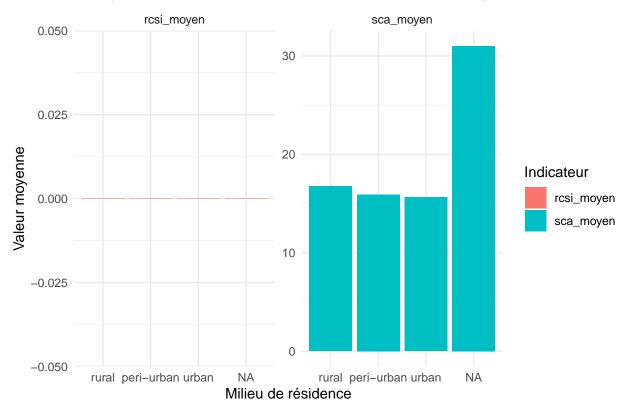








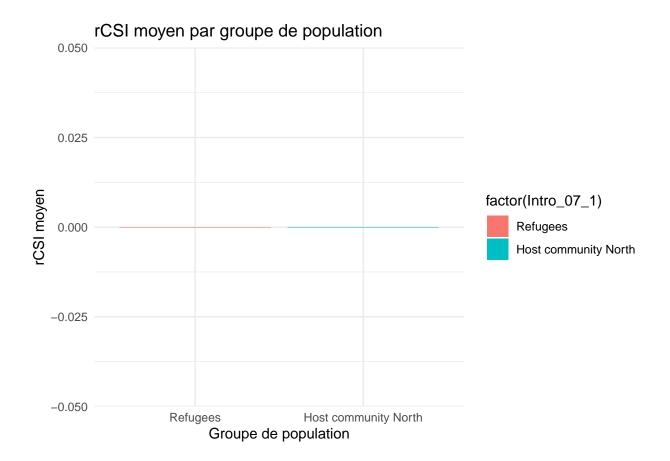
Comparaison des indicateurs de sécurité alimentaire par milieu de réside



5. Analyse croisée du rCSI avec d'autres variables socio-démographiques

Table 20: rCSI moyen par groupe de population

Intro_07_1	rcsi_moyen	rcsi_median	n	pct_eleve
Refugees	0	0	2068	0
Host community North	0	0	990	0



Test ANOVA pour comparer les moyennes du rCSI entre les groupes: Df Sum Sq Mean Sq F value Pr(>F) factor(Intro_07_1) 1 0 0 NaN NaN Residuals 3041 0 0 15 observations effacées parce que manquantes