Analyse des Données Ménages : Sécurité Alimentaire, Résilience et Conditions de Vie

Alioune Abdou Salam Kane

2025-05-10

Table of Contents

# 1 Introduction

Ce rapport a pour objectif de traiter, analyser et visualiser les données issues d’une enquête ménages portant sur divers aspects de la sécurité alimentaire, des conditions socio-économiques et de la résilience. Il s’inscrit dans un cadre méthodologique rigoureux visant à garantir la qualité, la cohérence et la représentativité des résultats produits.

L’analyse suit une démarche structurée en plusieurs étapes, chacune répondant à des objectifs précis :

1. **Analyse de consistence des données**  
   Vérification de la structure de la base, détection des doublons, traitement des valeurs manquantes, incohérences et imputations.
2. **Analyse des données et calcul d’indicateurs**  
   Estimation des indicateurs clés liés à la sécurité alimentaire, la résilience et les stratégies d’adaptation :
   * Profil socio-démographique des ménages
   * Score de consommation alimentaire (SCA)
   * Indice réduit des stratégies de survie (rCSI)
   * Stratégies d’adaptation aux moyens d’existence (LhCSI)
   * Score de diversité alimentaire (HDDS)
   * Score de résilience auto-évaluée (SERS)
   * Régime alimentaire minimum acceptable (MAD)
3. **Analyse comparative**  
   Mise en évidence des écarts selon le sexe du chef de ménage.
4. **Outil de visualisation**  
   Proposition d’un tableau de bord interactif pour explorer les résultats (R Shiny ou autre outil pertinent).

Chaque section du rapport présentera de manière détaillée les traitements réalisés, les choix méthodologiques adoptés, les résultats obtenus ainsi que leur interprétation.

# 2 Analyse de consistence des données

## 2.1 Chargement des packages

Les bibliothèques suivantes sont activées pour permettre toutes les opérations nécessaires : - tidyverse pour manipuler les données, - haven pour importer les données STATA, - labelled pour lire les libellés de variable, - janitor pour nettoyer les noms de colonnes, - rlang pour manipuler dynamiquement les noms de variables.

Le code est masqué à l’exécution, mais ces bibliothèques sont essentielles pour garantir la reproductibilité et la lisibilité des étapes qui suivent.

## 2.2 Importation de la base et nettoyage

Nous importons la base de données principale située dans le dossier data/ au format .dta. Ensuite, nous nettoyons les noms de colonnes pour les rendre uniformes : minuscules, sans espaces, ni caractères spéciaux. Cela simplifie la manipulation future des données.

## 2.3 Détection des doublons

L’objectif est d’identifier les éventuelles lignes dupliquées. Une ligne est considérée comme un doublon si toutes ses colonnes ont exactement les mêmes valeurs qu’une autre. Ces doublons, s’ils existent, doivent être traités pour éviter de biaiser les analyses statistiques.

## 2.4 Exploration des variables et de leurs libellés

Nous générons un tableau recensant l’ensemble des variables de la base, accompagné de leur description (label). Cela facilite la compréhension des données et permet d’identifier rapidement les variables clés pour l’analyse.

## 2.5 Identification des NA justifiés (sauts logiques)

Certaines valeurs manquantes ne résultent pas d’un oubli ou d’une erreur, mais d’un saut logique dans le questionnaire. Par exemple, un ménage qui n’a pas déclaré de difficultés ne répondra pas aux questions sur les stratégies d’adaptation. Nous mettons en place des règles pour détecter ces cas et éviter de les considérer comme des anomalies.

## 2.6 Extraction des NA injustifiés

Une fois les règles logiques établies, nous filtrons les lignes où des variables contiennent des NA injustifiés. Ces NA sont potentiellement problématiques car ils ne devraient pas exister selon la logique du questionnaire. Ils sont donc marqués pour traitement ultérieur.

## 2.7 Imputation simple des NA suspects

Les NA identifiés comme injustifiés sont imputés selon une méthode simple mais efficace : - Si la variable est **numérique**, on remplace les NA par la **médiane** de la variable concernée. Cette méthode est robuste face aux valeurs extrêmes. - Si la variable est **catégorielle**, on remplace les NA par la **modalité la plus fréquente** (le mode). Cela permet de conserver une logique cohérente dans les réponses.

Cette stratégie permet de corriger les données de manière transparente, tout en maintenant une certaine fidélité au profil moyen des ménages observés.

# 3 Analyse de consistence des données

## 3.1 Définition

L’analyse de consistence vise à évaluer la **qualité interne des données**. Il s’agit de détecter :

* des valeurs manquantes problématiques,
* des doublons,
* des incohérences logiques (par exemple, une personne déclarée “homme” et “enceinte”),
* des valeurs aberrantes (comme un âge de 150 ans),
* ou encore des erreurs de saisie typographiques ou hors des échelles prévues.

Cette première étape est cruciale : elle garantit que les résultats ultérieurs reposeront sur une base saine et fiable. Toute anomalie non corrigée pourrait fausser les indicateurs calculés, altérer les interprétations et compromettre la prise de décision.

## 3.2 Chargement des packages

Les bibliothèques suivantes sont activées pour permettre toutes les opérations nécessaires : - tidyverse pour manipuler les données, - haven pour importer les données STATA, - labelled pour lire les libellés de variable, - janitor pour nettoyer les noms de colonnes, - rlang pour manipuler dynamiquement les noms de variables.

Le code est masqué à l’exécution, mais ces bibliothèques sont essentielles pour garantir la reproductibilité et la lisibilité des étapes qui suivent.

## 3.3 Importation de la base et nettoyage

Nous importons la base de données principale située dans le dossier data/ au format .dta. Ensuite, nous nettoyons les noms de colonnes pour les rendre uniformes : minuscules, sans espaces, ni caractères spéciaux. Cela simplifie la manipulation future des données.

## 3.4 Détection des doublons

L’objectif est d’identifier les éventuelles lignes dupliquées. Une ligne est considérée comme un doublon si toutes ses colonnes ont exactement les mêmes valeurs qu’une autre. Ces doublons, s’ils existent, doivent être traités pour éviter de biaiser les analyses statistiques.

## 3.5 Exploration des variables et de leurs libellés

Nous générons un tableau recensant l’ensemble des variables de la base, accompagné de leur description (label). Cela facilite la compréhension des données et permet d’identifier rapidement les variables clés pour l’analyse.

## 3.6 Identification des NA justifiés (sauts logiques)

Certaines valeurs manquantes ne résultent pas d’un oubli ou d’une erreur, mais d’un saut logique dans le questionnaire. Par exemple, un ménage qui n’a pas déclaré de difficultés ne répondra pas aux questions sur les stratégies d’adaptation. Nous mettons en place des règles pour détecter ces cas et éviter de les considérer comme des anomalies.

## 3.7 Extraction des NA injustifiés

Une fois les règles logiques établies, nous filtrons les lignes où des variables contiennent des NA injustifiés. Ces NA sont potentiellement problématiques car ils ne devraient pas exister selon la logique du questionnaire. Ils sont donc marqués pour traitement ultérieur.

## 3.8 Imputation simple des NA suspects

Les NA identifiés comme injustifiés sont imputés selon une méthode simple mais efficace : - Si la variable est **numérique**, on remplace les NA par la **médiane** de la variable concernée. Cette méthode est robuste face aux valeurs extrêmes. - Si la variable est **catégorielle**, on remplace les NA par la **modalité la plus fréquente** (le mode). Cela permet de conserver une logique cohérente dans les réponses.

Cette stratégie permet de corriger les données de manière transparente, tout en maintenant une certaine fidélité au profil moyen des ménages observés.

## 3.9 Vérification des incohérences logiques

Au-delà des valeurs manquantes, certaines réponses peuvent être **présentes mais contradictoires**. Il s’agit de cas où les données renseignées ne respectent pas la logique interne du questionnaire. Nous détaillons ci-dessous les principales incohérences recherchées, ainsi que les principes de correction appliqués. Nous précisions par la même occasion qu’avant d’opérer toute modification sur la base originale nous l’avons dupliqué en amont.

### 3.9.1 Taille du ménage vs somme des tranches d’âge

**Logique** : Le nombre total de membres déclarés (hh\_size) doit correspondre exactement à la somme des effectifs par tranche d’âge.  
**Correction** : En cas d’écart, nous faisons confiance aux tranches d’âge (plus détaillées) et ajustons la variable hh\_size en conséquence.

Après vérification , il a été constaté qu’aucune incohérence n’a été détecté concernant les tailles de ménage.

### 3.9.2 Difficultés déclarées vs stratégies d’adaptation

**Logique** : Si un ménage déclare ne pas avoir rencontré de difficulté (sers\_difficultes = 1 ou 2), il ne devrait pas indiquer de stratégies d’adaptation (réduction des repas, vente d’actifs, etc.).  
**Correction** : Les valeurs renseignées dans les variables de stratégie sont remplacées par NA, car ces questions n’auraient pas dû être posées.

### 3.9.3 Consommation alimentaire vs source d’approvisionnement

**Logique** : Lorsqu’un ménage déclare avoir consommé un aliment spécifique (œufs, produits laitiers, fruits…), il doit également indiquer une source d’approvisionnement correspondante (\*\_s\_rf).  
**Correction** : Si cette source est absente ou égale à 0 alors que la consommation > 0, on attribue une valeur minimale (1) pour refléter l’obtention effective de l’aliment.

Ces vérifications garantissent une cohérence interne des données, primordiale pour la fiabilité des analyses ultérieures.

# 4 Analyse des données et calcul d’indicateurs

## 4.1 Analyse socio-démographique des ménages

Cette section vise à dresser un **profil socio-démographique des chefs de ménage**, stratifié par **sexe** (femme/homme), à travers trois variables essentielles :

* **Âge du chef de ménage** (en années),
* **Taille du ménage** (nombre total de membres),
* **Diplôme le plus élevé obtenu par le chef de ménage**.

L’objectif est de mieux comprendre les caractéristiques de base de la population enquêtée et d’identifier d’éventuelles disparités structurelles selon le sexe du chef de ménage.

### 4.1.1 Construction du tableau descriptif

Le tableau ci-dessous a été généré avec la fonction tbl\_summary() du package {gtsummary}. Il fournit :

* Pour les variables **numériques** : la **moyenne** et l’**écart-type**,
* Pour les variables **catégorielles** : l’**effectif** et la **proportion (%)** dans chaque groupe.

Les résultats sont **stratifiés par sexe** afin de comparer les caractéristiques entre les ménages dirigés par des femmes et ceux dirigés par des hommes. Le style est adapté à une sortie propre pour Word grâce à {flextable}.

Les résultats indiquent des différences notables selon le sexe du chef de ménage.

* **Âge du chef de ménage** : les hommes sont en moyenne plus âgés (45,17 ans) que les femmes (40,09 ans), ce qui peut refléter une structure familiale plus traditionnelle où les hommes prennent la tête du ménage à un âge plus avancé.
* **Taille du ménage** : les ménages dirigés par des hommes sont légèrement plus grands (7,75 personnes en moyenne) que ceux dirigés par des femmes (7,29), ce qui peut être lié à des structures familiales plus étendues ou à la cohabitation intergénérationnelle.
* **Diplôme le plus élevé** : les disparités de genre sont marquées. Une majorité de femmes sont alphabétisées ou ont suivi un enseignement coranique (50 %), mais une proportion importante (44 %) n’a aucun diplôme. En revanche, chez les hommes, seuls 21 % sont sans diplôme, et 71 % ont atteint au moins un niveau d’alphabétisation ou coranique. Les niveaux d’enseignement supérieur restent très faibles dans les deux groupes.

En résumé, les chefs de ménage hommes apparaissent globalement plus instruits, plus âgés et à la tête de foyers légèrement plus grands que leurs homologues féminins. Ces écarts peuvent avoir des implications sur l’accès aux ressources, les stratégies de subsistance et la vulnérabilité des ménages.

### 4.1.2 Distribution de la fréquence de consommation des différentes denrées

*(7 derniers jours – Score de Consommation Alimentaire, SCA)*

La majorité des ménages consomment quotidiennement des céréales. La fréquence de consommation de légumineuses, produits laitiers et protéines animales augmente légèrement avec le niveau d’éducation, suggérant une meilleure connaissance ou un meilleur accès à une alimentation diversifiée chez les chefs de ménage instruits.

## 4.2 3. Calcul et présentation de l’indicateur rCSI

Le **Reduced Coping Strategies Index (rCSI)** mesure la fréquence d’adoption de stratégies comportementales face à l’insécurité alimentaire, sur la base de 5 actions décrites par le PAM. Nous utilisons ici la version pondérée (somme des poids = 21) pour refléter la gravité relative de chaque stratégie.

* **Tableau 1** : résumé descriptif du nombre moyen de jours (sur 7) pendant lesquels chaque stratégie a été utilisée.
* **Tableau 2** : poids attribués à chaque stratégie pour le calcul du score pondéré.
* **Tableau 3** : distribution du score rCSI pondéré dans l’échantillon.

Ces tableaux sont interprétables sans connaissance préalable du questionnaire : ils donnent la fréquence moyenne d’adoption des stratégies et expliquent clairement la méthode de pondération.