le cnam
Ile-de-France

CERTIFICAT DE SPÉCIALISATION ANALYSTE DE DONNÉES MASSIVES

Entreposage et fouille de données

Projet de STA211

Auteurs :

Boukary OUEDRAOGO

Bangaly CAMARA

Imad EL HAMMA

Professeur :

Mme Niang

2021-2022

$$u(x) \approx U_h = \sum_{j=1}^{N} c_j \varphi_j(x)$$

$$u''(x_i) \approx \frac{1}{h^2}[u(x_{i+1}) - 2u(x_i) + u(x_{i-1})]$$

Table des matières

Résumé

1. Introduction

Le jeu de données CNIL est issu d'une étude menée par L'Observatoire de la Qualité de l'Air Intérieur (OQAI). L'étude a été menée sur un échantillon de 567 ménages suivis sur trois années. Les données ont été collecté en fonctions de trois grands ensembles décrivant les caractéristiques des logements, les habitudes des ménages et la composition des ménages. L'objectif de l'étude est de mesurer la qualité de l'air à l'intérieur des logements. Elle dépend de la présence de nombreux polluants physiques, biologique ou chimiques dont le formaldéhyde. Ce dernier est un polluant atmosphérique omniprésent dans les logements et dont le suivi est considéré comme étant hautement prioritaire par les pouvoirs publics depuis 2006.
Le but de notre de projet est d'expliquer et de prédire le formaldéhyde à partir des caractéristiques des trois ensembles de bloc de variables cités plus haut à savoir : les habitudes des ménages, les caractéristiques des ménages et la composition des ménages. Pour ce faire, nous allons nous appuyer sur différentes méthodes

de datamining. Afin de structurer notre travail, nous le scindons en deux parties. La première correspond à une exploration de données qui est préalable à tout travail de datamining. Cette étape consiste à analyser le jeu de données en unidimensionnel, en bidimensionnel et en multidimensionnel. Dans la seconde partie nous allons utiliser une méthode d'apprentissage supervisé pour modéliser et prédire le formaldéhyde.

2. Analyse exploratoire

Cette partie sera consacrée à l'analyse exploratoire des données. Certaines variables qui sont codées comme des variables numériques mais qui en réalité sont qualitatives seront recordées en variables facteurs.
Les blocs logement et habitude sont composés respectivement de 70 variables (dont 32 variables quantitatives et 38 variables qualitatives) et 44 variables (dont 21 variables quantitatives et 23 variables qualitatives). Les caractéristiques des logements sont composées entre autres du type de logement, du type de système de ventilation, d'une information d'existence de garages attenants et des installations utilisées. Les habitudes des ménages sont décrites à travers les tâches quotidiennes des ménages notamment les tâches ménagères, la possession d'animaux de compagnie, l'utilisation de parfum et d'insecticide, etc. Le bloc ménage quant à lui contient 11 variables dont 5 variables quantitatives et 6 variables qualitatives. Ces variables décrivent la composition et la catégorie socioprofessionnelle des ménages interrogés.
Par ailleurs, nous avons supprimé de l'analyse deux variables qualitatives KVN2e12 et KVNT2e112 du bloc logement car elles n'ont pas de variabilité due à l'existence d'une seule modalité.
Pour la suite de cette partie, nous allons uniquement présenter les résultats de l'analyse exploratoire du bloc ménage. Celle des autres blocs sont présentés dans l'annexe.

2.1 Analyse univariée

Nous commençons par une analyse univariée car elle permet de résumer l'information contenu dans les variables. Elle permet également de connaître la structure, la distribution des données et de détecter les outliers. L'analyse univariée montre que l'âge des ménages varie entre 18 et 89 ans, avec une moyenne de 61 ans. Ils disposent d'un revenu variant entre 535 et 7600 euros avec un revenu moyen de 2574 euros. Toutefois 50% des ménages touchent un revenu inférieur à 2349 euros et 45% ont un revenu inférieur à 1799 euro (voir tableau **??**). Les femmes représentent environ 65% des personnes interrogées dont 50% ont un niveau de l'enseignement technique court et seulement 30% ont un diplôme de l'enseignement supérieur bac +2. Nous constatons la même structure chez les hommes. La principale source de revenus est le salaire et la pension avec respectivement 53% et 34% des sources de revenus possibles. Ils sont pour la plupart des cadres supérieurs travaillant dans des professions intermédiaires (24%) et des ouvriers (20%) (graphique 1)

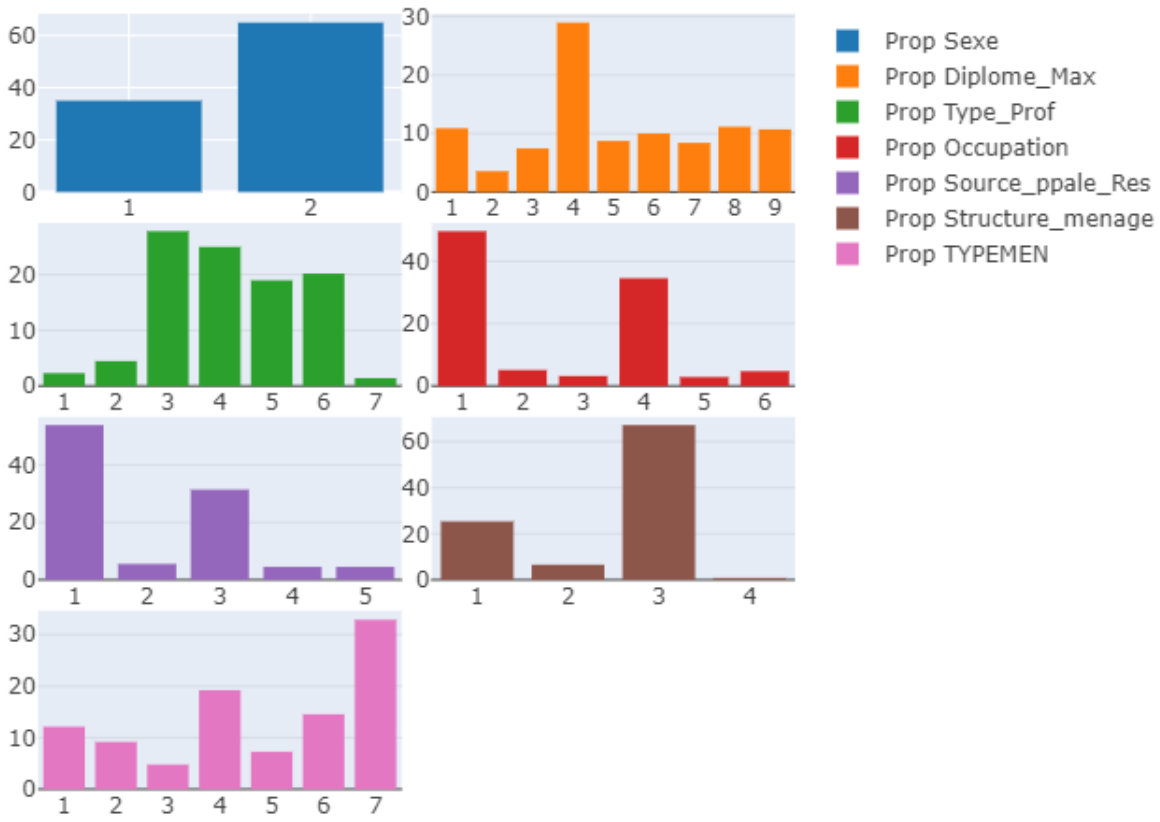

FIGURE 1 – Tri à plat des varaibles qualitatives du bloc ménage

Cette analyse univariée nous a permis de s'apercevoir que les variables quantitatives ne sont pas à la même échelle. Nous constatons également que certaines modalités des variables qualitatives sont dominantes comme les cadres supérieurs. Cependant, c'est une analyse préalable qui faudrait approfondir via une analyse bivariée.

2.2 Analyse bivariée

L'analyse de la matrice de corrélation des variables quantitatives du bloc ménage montre, globalement, une relation faible entre les variables à l'exception le nombre de personnes composant le ménage et le nombre d'enfant quel que soit l'âge(matrice de corrélation 2)

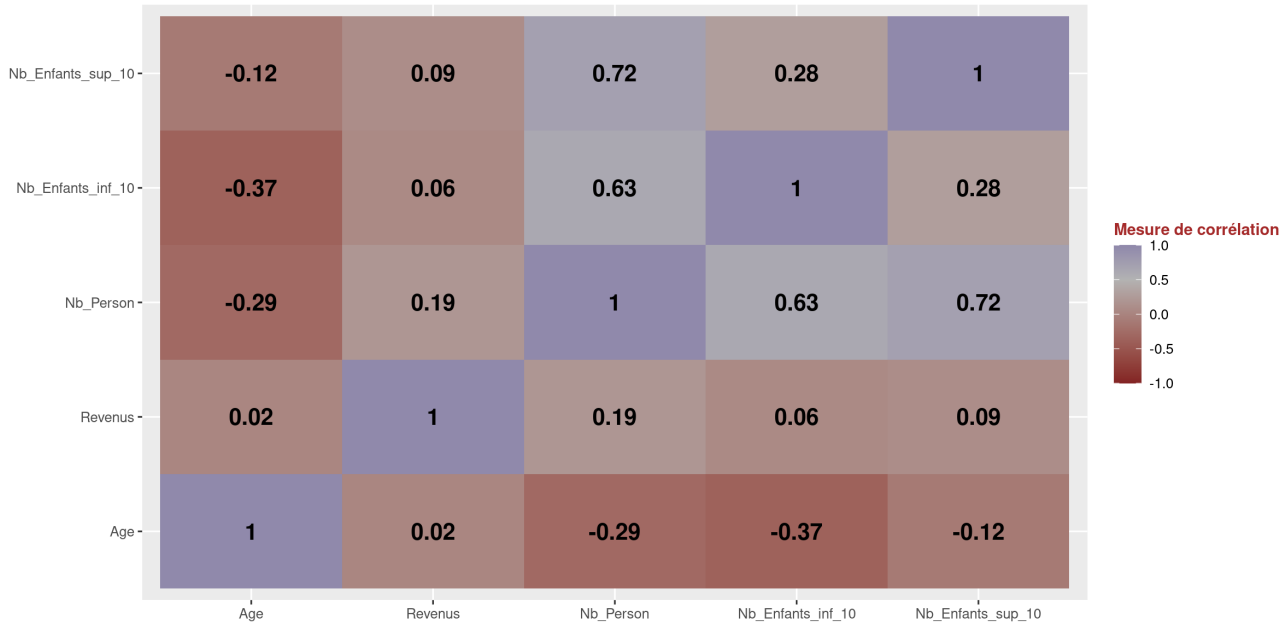

FIGURE 2 – Matrice des corrélations des variables qualitatives du bloc ménage

L'analyse du $Chi2$ de contingence et le V de Cramer permet de conclure qu'il existe une relation de dépendance entre les variables quantitatives. Toutefois, celle-ci est faible à l'exception de la relation entre la source principale de revenu et l'occupation. (tableau résultat V de cramer 2).

	sexe	type_prof	diplome_max	occupation	source_ppale_res	structure_menage
sexe	1					
type_prof	0.1634	1				
diplome_max	0.1146	0.1435	1			
occupation	0.2677	0.2370	0.1033	1		
source_ppale_res	0.1347	0.1824	0.1091	0.5634	1	
structure_menage	0.3495	0.0957	0.0750	0.1493	0.1866	1

TABLE 2 – Matrice des V de Cramer

Après avoir terminé l'analyse bidimensionnelle dans laquelle nous n'avons pas trouvé de relation forte entre les différentes variables prises deux à deux, nous passons à l'analyse multidimensionnelle.

2.3 Analyse exploration multidimensionnelle

Dans l'analyse multidimensionnelle, nous avons réalisé une ACP pour les variables quantitatives, une ACM pour les variables qualitatives et une AFM pour les variables mixtes (quantitatives et qualitatives prises ensemble)

2.3.1 Analyse en composante principale(ACP)

Les résultats de l'ACP sur les données quantitatives des ménages montrent que le premier plan factoriel exprime plus de 67% de la variabilité. La projection des cinq variables sur les deux premiers axes factoriels montre que le nombre d'enfant (quel que soit leur âge) et le nombre de personnes vivant dans le même foyer sont positivement corrélés à la première composante principale. Le revenu et l'âge sont fortement corrélés à la deuxième composante. Ainsi, deux structures se dégagent. La première structure est composée de l'âge et le revenu. La deuxième structure est formée par les variables nombre de personnes du foyer et le nombre d'enfants inférieur ou supérieur à 10 ans. La projection des individus sur le premier plan factoriel permet de distinguer quatre groupes de ménages :

- des ménages aisés, relativement vieux avec peu de personnes vivant sur le même toit ;
- des ménages aisés vivant avec plusieurs personnes sur le même toit avec ou sans enfants
- des ménages très jeune, avec des revenus modeste et peu de personnes vivant sur le même toit.

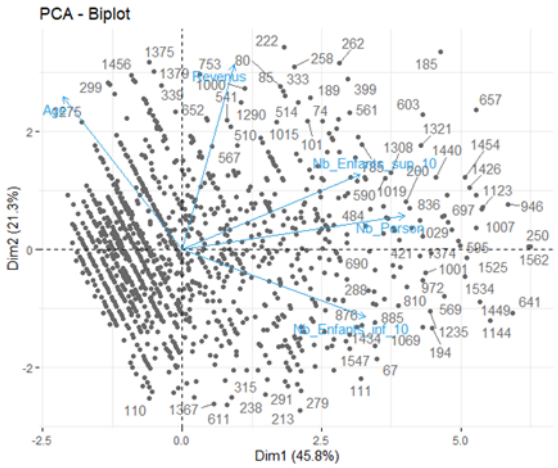

FIGURE 3 – Graphique individus/variables de l'ACP

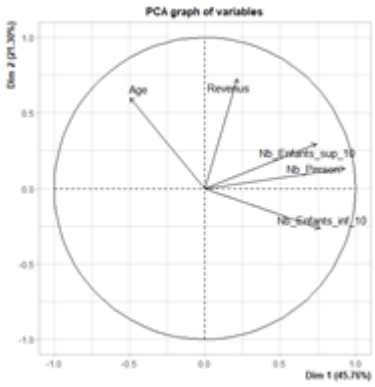

FIGURE 4 – Cercle de corrélation

2.3.2 Analyse des correspondances multiples(ACM)

Les résultats de l'ACM montrent que la première et deuxième dimension captent environ 15% de la variabilité. Les deux catégories « occupation et source principale de ressources » sont fortement corrélées (> 0.70) avec les deux dimensions. En revanche, les catégories « type de profession », « sexe », « structure de ménage » et le « diplôme le plus élevé » sont faiblement corrélées avec les deux axes. La projection des variables qualitatives les plus contributives sur les deux axes montre que : Les personnes déclarant recevoir une pension déclarent la retraite comme principale occupation. Les étudiants déclarent autres revenus comme source principale Les personnes déclarant avoir une activité professionnelle déclarent le salaire comme étant une source principale source de revenu. Et enfin, si on regarde l'allure générale du nuage des individus, c'est-à-dire les ménages, on voit qu'elle est particulière et met en évidence quatre classes de

ménages bien distinctes. La première bissectrice oppose les individus qui exercent une fonction (salariés) aux étudiants et ceux qui reçoivent un salaire à ceux dont le revenu provient d'un actif financier ou autres. Quant à elle, la deuxième bissectrice oppose ceux ont un diplôme de « fin études primaire » et ceux en «fin du second cycle enseignement général ».

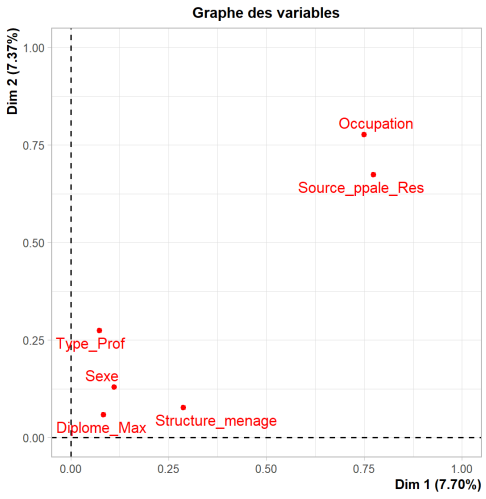

FIGURE 5 – Graphique des variables

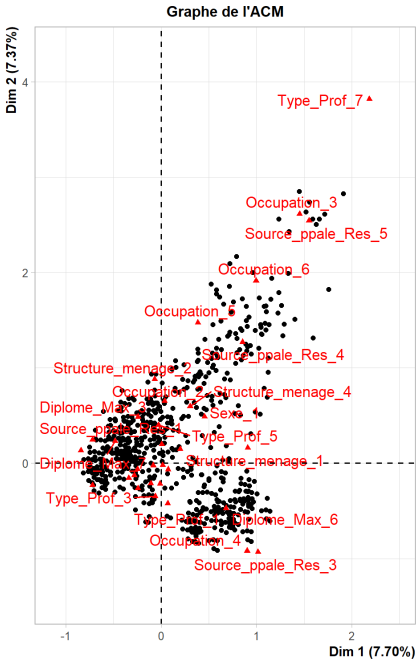

FIGURE 6 – Graphiques des modalités des variables / Individus

2.3.3 Analyse factorielle multiple(AFM)

Dans le résultat de l'analyse factorielle multiple des ménages, on constate que les coordonnées des groupes de variables qualitatives et quantitatives sont proches sur le premier axe factoriel. Sur l'axe 2 ils ont des groupes très différents sur le second axe factoriel. Cependant lorsqu'on calcul les coefficients Lg et Rv, on constate que les deux groupes de variables pris deux à deux ne sont pas significativement liées.

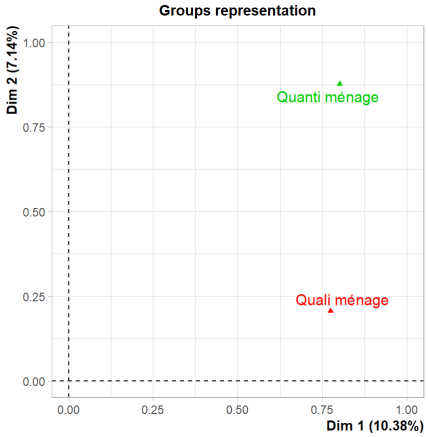

FIGURE 7 – Graphiques des groupes de l'AFM sur le bloc ménage

2.3.4 AFM sur l'ensemble des variables du jeu de données

Lorsqu'on analyse les résultats de l'AFM global, on constate que les groupes de variables quali_logement et quali_habitude sont proches sur le premier axe factoriel et un peu éloigné sur le second axe. On constate aussi que les groupes de variables du bloc logement (quantitative et qualitative) sont proches à la fois sur le premier axe factoriel et sur le second axe factoriel. Globalement, les coefficients Lg et Rv permettent de conclure que les différents blocs de variables ne sont pas significativement liés car les valeurs de ces coefficients sont faibles. Ainsi, nous pouvons conclure à la présence d'une structure par bloc dans notre dataset car ils n'ont pas de lien significatif entre eux.

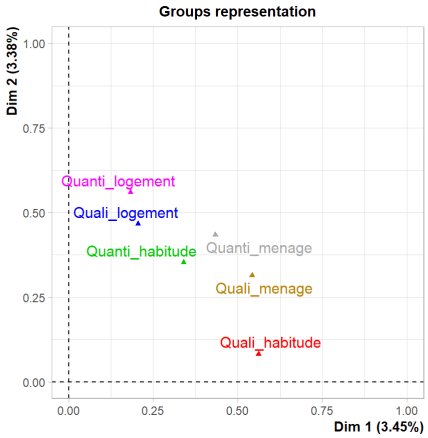

FIGURE 8 – Graphiques des groupes de l'AFM sur l'ensemble des variables

2.3.5 Classification des individus

Dans cette étape nous souhaitons voir s'il existe une structure en groupe dans les individus (ménages) du jeu de données. Pour cela nous allons réaliser une classification hiérarchique ascendante sur le jeu de données. Le but étant de voir s'il existe une structure en groupe au sein des ménages et de calculer les groupes. Comme nous avons un jeu de données mixtes, nous allons utiliser l'indice de similarité de Gower. L'objectif de cet indice consiste à mesurer dans quelle mesure deux individus sont semblables. L'indice de Gower varie entre 0 et 1. Dans le package cluster, la fonction daisy permet de calculer la distance de Gower définie par : $D_g = 1 - S_g$ avec D_g la distance de Gower et S_g l'indice de similarité de Gower. Avec la distance de Gower, deux individus sont identiques si la distance de Gower vaut 0 et ils sont totalement différents si D_g vaut 1.

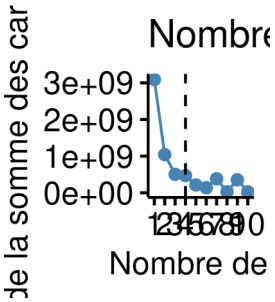

FIGURE 9 – Graphique résultant de la méthode du coude (Elbow)

Le graphique 9 indique que le nombre optimal de cluster est 4 Nous réalisons un Kmeans sur les résultats de la classification hiérarchique ascendante pour consolider les groupes.

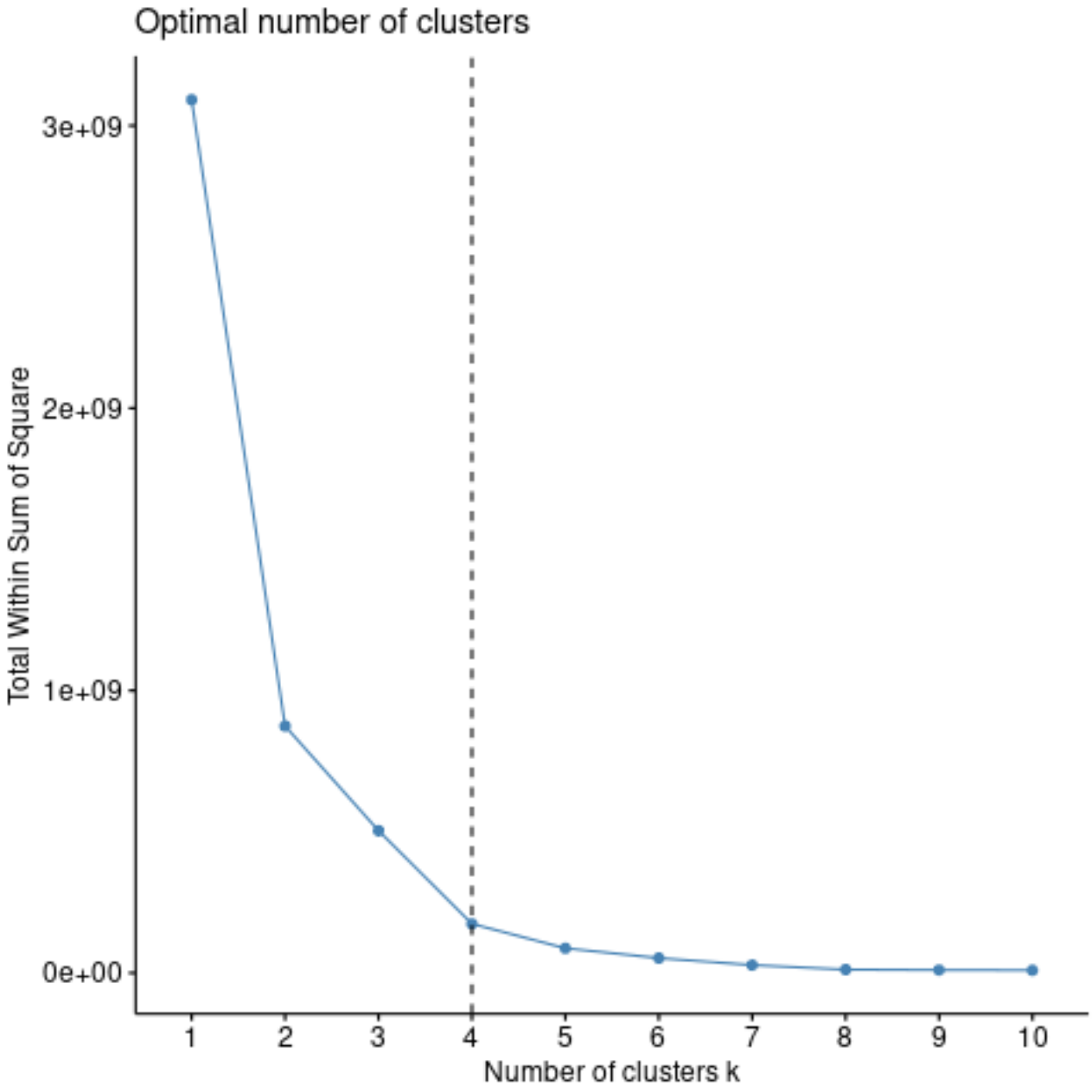

FIGURE 10 – Graphique des clusters

On observe sur le graphique 10

2.3.6 Classification des variables

Le but de la classification de variables est de regrouper ensemble des variables fortement corrélées entre elles en les séparant en classe. Nous utilisons le package ClusterOfVar. La méthode de classification des variables avec ClusterOfVar permet de créer des groupes de variables liées entre elles au sein des groupes et hétérogènes à l'extérieur des groupes. L'avantage principale de cette méthode est qu'elle prend en compte des données mixtes. Le critère d'homogénéité d'une classe est la somme des carrés des corrélations (pour les variables quantitatives) et des rapports de corrélations (pour les variables qualitatives) à une variable synthétique (quantitative) résumant au mieux les variables de la classe. La variable synthétique qui maximise ce critère est la première composante principale de l'AFDM. Deux algorithmes de classifications sont implémentés dans ce package :

1) Un algorithme de classification hiérarchique ascendante à travers la fonction hclustvar
2) Un algorithme itératif de partitionnement

Après avoir implémenté cet algorithme, nous trouvons huit clusters dont 2 sont analysés dans ce rapport. Cluster 1 : ce cluster regroupe les variables de structure des ménages, les femmes, les activités ménager (sortie des ordures), les produits de beauté (déodorant, soin visage, ...) Cluster 2 : il rassemble les variables d'habitude des ménages (jardinage, bricolage, séchage de linge dans le logement), l'utilisation d'insecticide, les habitudes de cuisson à l'eau, ...

3. Modélisation et prédiction du Formaldéhyde

3.1 Traitement des données manquantes

Dans le jeu de données la variable cible *Formaldéhyde* présente 3% de valeurs manquantes. L'imputation de ces valeurs manquantes est un préalable à l'étape de la modélisation. Nous allons utiliser la méthode d'imputation multiple par chaine de markov MICE(Multiple Imputation by Chained Equations). Cette méthode est basée sur un algorithme Monte-Carlo M

3.2 Modélisation

Les différentes méthodes de fouille de données nous ont permis voir que les individus(ménages) ont une structure en groupe donc sont hétérogènes en dehors des groupes. Nous avons pu identifier 3 groupes d'individus. Par ailleurs, la classification des variables a permis de mettre en évidence 8 clusters de variables. Cette structure en groupe des individus ne permet pas donc d'appliquer un modèle paramétrique sur l'ensemble du jeu de données. Une alternative pour utiliser des modèles paramétriques serait d'appliquer une modélisation par groupe. L'autre alternative serait d'utiliser des modèles non paramétriques comme les

arbres de décisions. Etant donné l'instabilité des résultats des arbres de décisions nous utilisons le choix sera porté sur des modèles d'ensemble de type random forest. L'inconvénient des random forest est que le calcul de l'importance des variables est biaisé lorsque celles-ci sont structurées en blocs et très corrélées au sein des blocs. Ce qui est le cas de notre jeu de données. Une solution pour pallier aux insuffisances des random forest c'est le fuzzy forest. Le fuzzy forest est une extension des random forest et qui permet de calculer l'importance des variables de manière non-biaisée. L'algorithme construit une classification de variables. Ensuite il réalise un random forest sur chaque cluster puis selectionne les variables les plus importantes de ses différents modèles pour réaliser un random forest final. L'importance des variables est calculée finalement à partir de ce dernier modèle de random forest.

3.2.1 Fuzzy forest sur les groupes issues de ClusterOfvar

3.2.2 Comparaison et selection de modèle

4. Conclusion