UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

The clustering of (un)healthy behaviors and the link to behavioral intentions: The Flemish Prevention Barometer

Author: **Léa Jacquet**
Supervisors: **Pr. Catherine Legrand, Dr. Elise Braekman**
Readers: **Pr. Bernadette Govaerts, Pr. Johan Segers**
Academic year 2022-2023

UCLouvain

Institut de Statistique, Biostatistique et Actuariat

The clustering of (un)healthy behaviors and the link to behavioral intentions: The Flemish Prevention Barometer

Membres du jury:

Promotrice:
Pr. Catherine LEGRAND

Co-promotrice:
Dr. Elise BRAEKMAN

Lecteur·ice·s:
Pr. Bernadette GOVAERTS
Pr. Johan SEGERS

Mémoire présenté en vue de l'obtention du master en sciences des donnés (orientation statistiques) par:

Léa JACQUET

Louvain-La-Neuve
Juin 2023

Preface.

This dissertation has been prepared in partial fulfillment of the requirements for the master degree in data sciences and statistics delivered by the Université Catholique de Louvain.

Acknowledgment.

I would like to express my sincere gratitude to my co-promoters, Catherine Legrand and Elise Braekman, for taking the time to guide me through this work, sharing their relevant feedback and supporting me. Additionally, I would like to thank Sciensano for trusting me to carry out this research and enabling me to give meaning to my master thesis. I would also like to thank Bernadette Govaerts and Johan Segers for accepting to read this work. Finally, I would like to take this opportunity to thank my family, especially my sister and my mother, and friends for giving me 5 wonderful years at Uclouvain and for supporting me throughout.

Contents

Notations

Notation	Meaning of notation
n	Number of observations
$\mathbf{x}_i$	A vector representing a data point in a P-dimensional space
$\mathbf{c}_k$	A vector representing the centroid of cluster k in P-dimensional space
$d(\mathbf{x}_i, \mathbf{c}_k)$	A distance between data point $\mathbf{x}_i$ and centroid $\mathbf{c}_k$
C_k	Represents the clusters of data
I_a	Intraclass inertia
I_c	Interclass inertia
$\mathbf{c}$	The centre of gravity of the entire dataset
$sil(\mathbf{x}_i)$	Silhouette coefficient for data point $\mathbf{x}_i$
S_{ij}	Gower's similarity coefficient between observations i and j
ω_p	Weight of variable p
δ_{ijp}	Binary variable indicating if i and j can be compared based on variable p
L_p	The number of possible categories for the p-th variable
s_{ijp}	Degree of similarity between observations i and j along a specific variable p
g_{ij}	Dissimilarity measure between observations i and j based on the Gower distance
$w_{i,k}$	Membership value of data point x_i for cluster k
m	The fuzzifier parameter
L_p	Number of possible categories of variable p
x_{ipl}	Observed value of the p-th categorical variable
$\pi_{\mathbf{k}}$	Individual class membership probability for individual i to belong to class k
π_{pkl}	Class-conditional probability that an individual belonging to class k will produce the outcome l on variable p
p_k	Class membership probability of class k at the population level
λ	The maximum log-likelihood of the model
ϕ	The number of parameters
y_{ik}	Indicator taking value 1 if the i-th observation belongs to category k, and 0 otherwise
∇l	The gradient vector
H	The Hessian matrix
$Wald_{pl}$	The Wald statistic
$\boldsymbol{\beta}$	A matrix containing the parameter of the multinomial logistic regression

List of the notations used in this report

Chapter 1

Introduction

Over the past few decades, unhealthy behaviours have become a growing public health concern. In 2018, the World Health Organization has released a report emphasizing the growing number of cases of non-communicable diseases, which accounts for 71% of the worldwide deaths [1]. There is agreement that these non-communicable diseases are closely related to unhealthy behaviors, such as excessive alcohol consumption, physical inactivity, poor diet habits, etc. [1]. While it is well known that engaging in a one unhealthy behavior has a negative impact on health, these behaviors tend to co-exist having synergistic effects. In other words, their combined impact on health is greater than the sum of their individual effects. A study conducted by Ford et al. in 2012 demonstrated that the risk of all-cause mortality decreases as the number of healthy behaviors increases [3]. This suggests that multiple unhealthy behaviors can drastically increase the risk of diseases and reduce the quality of life.

It is widely acknowledged that unhealthy behaviors are not uniformly distributed across the population as they are influenced by different socio-demographic factors. So, understanding how these behaviors are clustered and distributed within the population is of major importance for developing strategies to encourage multiple health behavior changes [3].

Several researches have examined the topic of multiple risk behaviors using different grouping methods (latent class analysis, clustering methods, principal component analysis, etc.). Main findings are consistent across the studies: smoking, alcohol consumption, poor diet and physical inactivity tend to co-occur together. Noble et al. (2015) showed that men and people being socially disadvantaged are more often engaged in multiple risk behaviors [4]. On the other hand, Hobbs et al. (2019) showed that the sex has no significant influence on the clusters, possibly due to the consideration of a wider range of health behaviors [5].

Although many researches have investigated the co-occurrence of unhealthy behaviors and its link with socio-demographic factors, this topic has not been yet studied in the Flemish population to our knowledge. Therefore, this is one of the objectives of the Flemish Prevention Barometer conducted by Sciensano. In addition, it aims to explore the intentions to change unhealthy behaviors. According to the literature, individuals are more likely to succeed in changing their behavior if the intention to change comes from themselves [6]. Indeed, it is widely accepted that intention to change is an important determinant of behavior and that the absence of intention often results in a lack of action [3]. Thus, understanding relationships

between intentions is essential for promoting positive lifestyle changes and creating effective health policies.

The aims of the Flemish Prevention Barometer are to investigate:

- The identification of behavioral clusters in the Flemish population.
- The association between clusters and socio-demographic characteristics.
- The willingness to change unhealthy behaviors and the association between intentions within each cluster.

There is evidence of the positive impact of implementing multiple behavior change interventions, and the above information are crucial for planning effective prevention and intervention strategies [3]. Therefore, conclusions of this work may influence public health policies in the Flemish region. By determining clusters of behaviors in the Flemish population and exploring their links with socio-demographic factors and intentions to change, the study may give valuable information for designing interventions with a maximal impact.

The contributions of this work are multiple. In chapter 2, we give a general overview of the methodology used: it explains the objectives of the Flemish Prevention Barometer, the general study design, the database description and the statistical methods used. Chapter 3 provides a technical insight of the statistical methods. It discusses potential clustering methods and their adjustments to handle categorical variables, the multinomial logistic regression and the techniques used to assess the relationship between clusters and intentions to change. As we encountered missing values, this chapter describes some possible ways to handle them. Chapter 4 focuses on the results of the statistical analyses (presented in chapter 3) applied to our dataset. Among other things, it gives information about the clusters found in the Flemish population. Chapter 5 examines the limitations and the possible improvements of this study. Finally, Chapter 6 concludes by summarizing this work and presenting the results of interest of this study.

Chapter 2

Methodology

2.1 The Flemish Prevention Barometer

Unhealthy behaviors such as smoking, poor nutrition and physical inactivity cause many chronic diseases which lead to preventable death and disability. Given their detrimental health effects, preventing unhealthy behaviors and promoting healthy behavaiors is an important public health goal. The goal of the Flemish Prevention Barometer follows this objective by collecting information on health-related behaviors and on determinants that cause these behaviors in the adult Flemish population. It also focuses on multiple-risk behaviors as a single risk behavior can have a major impact on the health but a combination of these behaviors drastically increases their impact [7].

2.1.1 Study population and sampling

This study is cross-sectional and targets a sample size of 4000 individuals in the general population of Flanders aged 18 years and above. Note that there is one exception: persons living in institutions (ex: residential care centers) are excluded for practical reasons. To select the target sample population, the National Register is used. The method used to improve the representativeness of the sample of the Flemish region population consisted of a combination of different sampling techniques. Firstly, the population was stratified by health care region, there are 14 "regional city" care regions in Flanders. Secondly, the population was stratified by sex as recorded in the National Register. This means that a double stratification was performed and 28 strata were obtained. The number of individuals selected within each strata was proportional to the actual number of inhabitants within that strata. A stepwise selection procedure based on age and education level was used to select individuals within the strata. This ensured that adults from all age groups and education categories were selected [7].

2.1.2 Data collection

The data we use have been collected by Sciensano from February 2022 to October 2022. In practice, an invitation to participate to the survey was sent by postal mail with a link to the online questionnaire. If the selected person did not participate, a reminder with a paper questionnaire and postage-paid envelope was sent 14 days later. Selected individuals were not obliged to participate, and a proportion of those selected did not participate. To achieve the target of 4000 participants, a system of substitution of non-participants was employed where

individuals with similar characteristics were selected as substitute individuals. A maximum of 5 reserve individuals were selected for each initially selected individual. If the selected person still did not answer 14 days after the reminder, an invitation was sent to the first reserve individual. Data collection was stopped when 4,000 participants were reached (actually, 4011 surveys were completed). The distribution of the realized sample in terms of sex, age, and education level is conform to the distribution in the Flemish population. During the first and the second waves of data collection, some particular groups were less likely to participate (low-educated men between 18 and 57 years old and middle-educated men between 18 and 37 years old). To solve this issue and to have a balanced representation of all groups, a conditional incentive of 10 euros was offered to those groups [7].

2.1.3 Participation

From Table 6.0.1 in the Appendix section, some information about the participation rate in the Prevention Barometer and the sample characteristics can be retrieved. Note that information about the participation of the initially selected individuals and the reserve individuals are not available in this table. The overall participation rate is 20.1% and more than a third of the participants preferred to take the paper questionnaire.

2.2 Data information

The original dataset contains 4011 observations and 84 variables, going from health and intention behaviors to socio-demographic factors. One drawback of the future methods that will be used is their limited ability to distinguish between relevant and irrelevant variables for the given task. They may not accurately determine the importance of each variable to the desired outcome, leading to crucial consideration of which variables to include in the clustering analysis [8]. Variables selection can be done subjectively or objectively. Objective methods rely on empirical data while subjective approaches are based on expert judgements or other published studies [8]. In this work, we select 22 variables that are aligned with the purpose of this study in a subjective way. There are 8 variables about health behaviors, 7 about socio-demographic characteristics and 7 about behaviors change.

2.2.1 Health related behaviors

- *Physical activity* was self-reported by asking how many hours/minutes of moderate/intense physical activity the individual did in the past 7 days. The WHO (*World Health Organization*) recommends at least 150 minutes a week of moderate to intense physical activity per week [9]. Therefore, the variable is dichotomized as fulfilling the WHO recommendation (PHY = 1) or as not fulfilling the WHO recommendation (PHY = 2).

- *Alcohol consumption* was assessed by asking the individual about their frequency of alcohol consumption per year and per week in the past 12 months. The variable is dichotomized, according to the recommendation of the CSS (*Conseil supérieur de la santé*), as excessive if the alcohol consumption is higher than 10 standardunits per week (ALC = 1) or as not excessive otherwise (ALC = 2) [10].

- *Smoking* was determined by asking if the individual smokes tobacco (not vaping or e-cigarettes) daily or almost daily. The variable is dichotomized as current smoker (SMO

= 1) or as not current smoker (SMO = 2).

- *Cannabis* was identified by asking if the individual has used cannabis in the past 12 months. The variable is dichotomized as cannabis user (CAN = 1) or as not cannabis user (CAN = 2).

- *Sitting time* was self-reported by asking the individual how many hours/minutes they sit or recline on a normal day. Initially, this variable was continuous but to simplify the analysis, it has been transformed into a categorical variable. Note that there is no recommended cut off by the WHO [9]. However, we use <=8 hours per day as a cut off as it is generally done in other studies. It is dichotomized as engaging in a sedentary behavior (SIT = 1) or not (SIT = 2).

- *Fruit/Vegetable* was determined by asking how many fruits and vegetables the individual eats per week. Initially, they were 2 distinct variables but as they are strongly associated with each other, we have decided to group them into a single one. This new variable is then dichotomized as eating fruits/vegetables daily (FRU_VEG = 1) or as not eating fruits/vegetables daily (FRU_VEG = 2). This is based on the recommendation from the CSS [11].

- *Snack* was assessed by asking how many snacks the individual eats per week. The variable is dichotomized as eating snacks daily (SNA = 1), which is an unhealthy behavior, or as not eating snacks daily (SNA = 2). This is based on the recommendation from the CSS [11].

- *Soda* is determined in the same way as the other healthy food variables. It is dichotomized as drinking sodas daily (SOD = 1) or as not drinking sodas daily (SOD = 2). This is based on the recommendation from the CSS [11].

Healthy behavior	Variable definition	Unhealthy behavior definition
Physical activity	Minutes of doing moderate and/or intense physical activity in the past 7 days	$<$ 150 minutes per week of moderate to intense physical activity [9]
Alcohol	Consumption of standard units of alcohol in the past 7 days	$>$ 10 standard units of alcohol per week [10]
Smoking	Current cigarette smoker	There is no cut off as daily or occasional smoking is bad in general
Cannabis	Cannabis use in the past 12 months	There is no cut off as using cannabis occasionally is bad in general
Sitting time	Total time sitting or reclining in a day	$\geq$ 8 hours per day [9]
Fruits/Vegetables	Eating fruit(s) and/or vegetable(s) daily	Not eating fruits or vegetables daily
Snack	Eating snack(s) daily	Daily
Soda	Drinking soda(s) daily	Daily

Table 2.2.1: Health behaviors selection

2.2.2 Socio-demographic variables

The following socio-demographic factors are evaluated: sex, age, Body Mass Index (BMI), spoken language at home, household income, household type and education level. The table below lists questions used to assess socio-demographic characteristics and how they are categorized in the dataset. Note that it is important to point out that these are information reported by the participants. In addition, for the income variable, the perception of the difficulties of living on the household income may be different for each participant.

Demographics	Question	Recoding into categories
Sex (SEX)	What's your gender?	(1) = man (2) = woman (3) = other
Age (AGE7)	What's your birth year?	(1) = 18-24 years old (2) = 25-34 years old (3) = 35-44 years old (4) = 45-54 years old (5) = 55-64 years old (6) = 65-74 years old (7) = 75+ years old
BMI (BMI_CAT)	How tall are you? What is your weight? The Body Mass Index is then calculated with the weight in kgs and the square of the height in meters.	(1) = underweight category (BMI $\leq$ 18.5) (2) = normal category (18.5 $<$ BMI $\geq$ 24.9) (3) = overweight category (25 $\geq$ BMI $\geq$ 29.9) (4) = obesity category (BMI $\geq$ 30)
Language (NDLS)	What language or languages do you usually speak at home?	(1) = speaking Dutch at home (2) = not speaking Dutch at home
Income (IN)	Are you finding it difficult or easy to live with your household income?	(1) = with great difficulties (2) = with difficulties (3) = with some difficulties (4) = fairly easily (5) = easily (6) = very easily
Household type (HHTYPE)	How do you live?	(1) = alone (2) = alone with child(ren) (3) = as a couple without child(ren) (4) = as a couple with child(ren) (5) = with my parent(s), family, friends or acquaintances
Education level (EDUC)	What is your highest degree or certificate?	(1) = low education level (2) = middle education level (3) = higher education level

Table 2.2.2: Socio-demographic characteristics selection

2.2.3 Intention variables

In addition to the health related behaviors, there 7 are variables about the intention to change unhealthy behaviors within the future 12 months. An example on how the question is formulated in the survey can be found in the Appendix section (Fig. 6.0.1).

- *ALC_FUTURE_LESS* is the intention to reduce alcohol consumption within the next 12 months. This question was asked to all participants reported having drunk at least one unit of alcohol within the past 12 months.

- *ALC_FUTURE_STOP* is the intention to stop alcohol consumption within the next 12 months. This question was asked to all participants who reported having drunk at least one unit of alcohol within the past 12 months.

- *SMO_FUTURE_STOP* is the intention to quit smoking within the next 12 months. This question was asked to all participants who reported being current smokers.

- *CAN_FUTURE_STOP* is the intention to stop using cannabis within the next 12 months. This question was asked of all participants who reported using cannabis at least once a month.

- *PHY_FUTURE_MORE* is the intention to do more physical activity within the next 12 months. This question was asked to all the participants.

- *SIT_FUTURE_LESS* is the intention to spend less time sitting within the next 12 months. This question was asked to all the participants.

- *EAT_FUTURE_BETTER* is the intention to eat healthier within the next 12 months. This question was asked to all the participants.

They can take 3 values: (-3) Not applicable, (1) Yes, (2) No.

2.2.4 Impact of dichotomising continuous variables

In many sectors, it is a common approach to categorize continuous variables. It has the advantages to simplify the analysis and the interpretation/presentation of the results. In the health sector, for instance, there is the necessity to classify individuals as having or not an attribute in order to detect a disease [12]. In this work, variables were dichotomized prior the analysis to define what constitutes an unhealthy behavior and to purse specific objectives. However, it can also lead to several issues. Firstly, dichotomising results in a loss of information and thus the statistical power to detect any relationships between the variable and observation outcome is reduced. Secondly, this approach can lead to serious underestimation of the variation of the outcome between groups and a significant variability may be included within each group. Individuals close to the cut off point, but on either side of it, are defined as very different rather than very close [12]. It is then important to be aware, for the rest of the analysis, that results could be more reliable if variables were not dichotomized.

2.3 Statistical methods

As it has been proven that multiple risk-behaviors have a considerable negative impact on health, one goal of the Prevention Barometer is to investigate their combinations. The main objective of this study is to define clusters by applying clustering methods on the sample. The idea behind is to identify relevant groups with distinguishable patterns of risk behaviors. Three methods of clustering are compared in order to get the most relevant clustering: K-means, fuzzy K-means and latent class analysis. Based on different assumptions of the methods presented in Section 3.1, on the readability of the results and on the assumptions of the methods, one will be chosen.

Before conducting the clustering analysis, it is important to get an overview of the sample. Descriptive statistics are used to explore socio-demographic characteristics of the sample, the frequency of each risk behaviors and the engagement in multiple risk behaviors. This analysis enables to identify missing values and should be investigated through a sensitivity analysis. Then, a multinomial logistic regression is performed to examine whether cluster are associated with socio-demographic factors. Finally, an association analysis within clusters is performed in order to investigate the willingness of individuals to change multiple unhealthy behaviors. It relies on a fundamental test: the Fisher's Exact test.

Note that all statistical analysis are performed in R. Different R packages are needed in order to run these techniques. The *pam()* function from the *cluster* package [13], the *fanny()* function from the *cluster* package [14] and the *poLCA()* function from the *poLCA* package [15] enable to implement the different clustering techniques. The multinominal logistic regression is implemented with the *multinom* function from the *nnet* R package [16]. Finally, to performed a Fisher's Exact test, we use the *fisher.test()* function from the *stats* package [17].

Chapter 3

Statistical analysis

3.1 Clustering analysis

The interest in clustering analysis techniques has increased these passed years and have been applied to a wide range of topics. In the context of our study, clustering analysis is applied to public health to group individuals engaging in (un)healthy behaviors. There are several ways to define what a clustering analysis is. However, all definitions share the same fundamental approach: clusters represent groups of similar observations. The main idea of these methods is to classify data based on all variables of interest into K subgroups such that data present in the same class are more similar with each other than with the ones in the other classes [18]. There are several criteria to obtain the optimal partition between data. The Between- and Within-clusters homogeneity are 2 of the main criteria: the partition should maximize the former one and should minimize the latter one to achieve greatest separation between clusters [19].

There exist many clustering methods which may give different results and make the interpretation depending on the method. In the rest of the analysis, we mainly focus on 3 methods: K-means clustering, fuzzy K-means clustering and Latent Class Analysis.

3.1.1 K-means clustering

The K-means is one of the most popular unsupervised clustering algorithm for handling continuous data [20]. It is a type of partition clustering. Partition clustering algorithms are methods that divise the data points into a predetermined number of partitions, which represent clusters. There exist different partitioning clustering algorithms with different set of rules to divise the data [20].

The main principle of the K-means clustering is to allocate a set of data into K homogeneous clusters, where K is determined prior. It is an iterative algorithm and it starts by choosing arbitrarily K centroids, representing the K clusters. Then, each data point is assigned to its nearest centroid, using a certain distance measurement. At each iteration, new centroids are calculated based on the clusters defined at the previous iteration and the allocation of the data is updated. This loop ends when there is no more change in the centroids position, that is, when the convergence criterion is met [20].

The following notations will be used :

- n is the number of observations
- P is the number of variables
- K is the number of clusters
- $\mathbf{x}_i \in \mathbb{R}^P$ is a vector representing a data point in a P-dimensional space ($i = 1, ..., n$)
- $\mathbf{c}_k \in \mathbb{R}^P$ is a vector in a P-dimensional space containing the coordinates of the centroid of cluster k ($k = 1, ..., K$)
- $d(\mathbf{x}_i, \mathbf{c}_k)$ is the distance measure between data point $\mathbf{x_i}$ and centroid $\mathbf{c_k}$
- C_k represents the clusters of data ($k = 1, ..., K$)

C_k is defined by using a specific distance metric and a set of K centroids as:

$$C_k = \{\mathbf{x_i} : d(\mathbf{x_i}, \mathbf{c_k}) \leq d(\mathbf{x_i}, \mathbf{c_s})\} \tag{3.1.1}$$

$\forall\ s \in 1, ..., K$

The principal goal of K-means algorithm is to find a partition that minimizes the intraclass inertia [22].

$$I_a = \frac{1}{n} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{c}_k)^2 \tag{3.1.2}$$

Maximizing the interclass intertia I_c, which quantifies the between-cluster variance, can help ensuring greater separation between clusters. The interclass inertia is defined as:

$$I_c = \sum_{k=1}^{K} \frac{|C_k|}{n} d(\mathbf{c_k}, \mathbf{c}) \tag{3.1.3}$$

where $\mathbf{c}$ is the centre of gravity of the entire dataset denoted by $\mathbf{c} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and where $|C_k|$ indicates the number of data points in cluter k.

The algorithm follows this pseudo-code:

Algorithm 1: K-means clustering algorithm
Input: $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ in $\mathbb{R}^p$ //set of n data points K //a priori fixed number of clusters
Output: $C = \{C_1, C_2, ..., C_K\}$ // set of K clusters
Steps: 1. Choose arbitrarily $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K$ in $\mathbb{R}^p$, the K initial centroids 2. Do until convergence: - For each i $\in \{1, 2, ..., n\}$, assign $\mathbf{x}_i$ to the cluster with the closest centroid; - For each k $\in \{1, 2, ..., K\}$ calculate the new centroid $\mathbf{c}_k$

Table 3.1.1: Pseudo-code for K-means

New centroids are calculated with the following formula:

$$\mathbf{c}_k = \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i \tag{3.1.4}$$

where $\mathbf{c}_k = (c_{1k}, c_{2k}, ..., c_{pk})$ and $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{pi})$.

The K-means clustering method goes fast, is efficient and works well for large datasets. However, one major disadvantage is that the final results are sensitive to the initial position and number of centroids. It is therefore recommended to run this algorithm several times with different initial centroids and to select the model which gives the lowest intraclass inertia [20].

Average silhouette coefficient method

As already explained, the output of the K-means algorithm directly depends on the a priori fixed number of clusters, the K-value. However, it is generally difficult to determine the optimal K-value and one possible method that can be used is the Silhouette coefficient method [23]. The main idea behind it is to measure how similar an observation is to the data points of its cluster based on the concepts of cohesion and separation. The former one calculates the average distance between the observation and the other points in its cluster, that is, the degree of similarity between observations in the same cluster. The latter one calculates the average distance between the observation and the other points in the next nearest cluster, that is, the separation between the observation and the data points in the nearest cluster. The silhouette coefficient value ranges between -1, indicating that the data point is probably in the wrong cluster, and 1, indicating that the data point is well-clustered. By calculating the Silhouette coefficient for each data point $\mathbf{x}_i$, we can estimate the overall quality of the clustering. It can be defined as [23]:

$$sil(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}} \tag{3.1.5}$$

where the cohesion and the separation factors are respectively defined as:

$$a(\mathbf{x}_i) = \frac{1}{|C_k| - 1} \sum_{\mathbf{x}_j \in C_k, j \neq i} d(\mathbf{x}_i, \mathbf{x}_j) \tag{3.1.6}$$

$$b(\mathbf{x}_i) = \frac{1}{|C_{k'}|} \sum_{\mathbf{x}_j \in Ck'} d(\mathbf{x}_i, \mathbf{x}_j) \tag{3.1.7}$$

where $C_{k'}$ corresponds to the closest cluster of $\mathbf{x}_i$ for which $\mathbf{x}_i$ is not part of it. .

To determine the optimal number of clusters, we consider the average silhouette coefficient, that is, the mean value of $sil(\mathbf{x}_i)$ for all data points. The idea is then to choose the K-value that corresponds to the maximal average silhouette coefficient. Figure 3.1.1 is an illustration of the average silhouette coefficient method. It displays the average silhouette coefficient value on the y-axis and K-values on the x-axis.

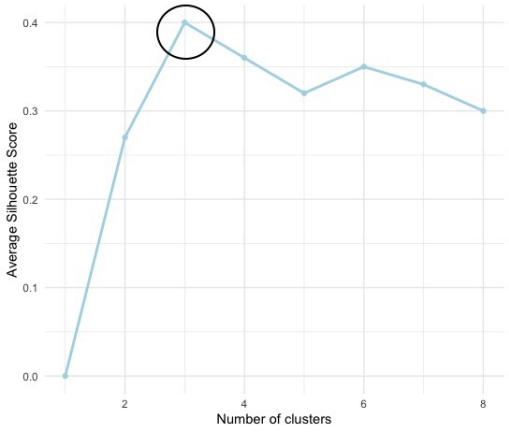

Figure 3.1.1: Average silhouette coefficient method illustration

Distance metric

The choice of the distance is arbitrary and the Euclidean distance is usually used in the presence of only continuous variables. However, it cannot be applied to our case as the dataset contains mainly categorical variables. To handle this issue, the Gower distance is proposed as a similarity coefficient for evaluating distances between mixed-data. It defines the similarity between observations i and j as [24]:

$$S_{ij} = \frac{\sum_{p=1}^{P} s_{ijp}\delta_{ijp}\omega_p}{\sum_{p=1}^{P} \delta_{ijp}\omega_p} \tag{3.1.8}$$

- P is the number of variables.
- ω_p the weight of variable p. Choosing a reasonable set of weights is difficult and one possible weighting is to give equal weights to all variables [25].
- δ_{ijp} a binary variable. It is equal to 1 if i and j can be compared based on variable p otherwise it is equal to 0.

- s_{ijp} is an index that reveals the degree of similitude between observations i and j along a specific variable p. The index ranges between 0 and 1 (0 indicating that the 2 observations are very different, while 1 indicating that the 2 observations are very similar). In the case of a categorical/binary variable: $s_{ijp} = 0$ or $s_{ijp} = 1$. In the case of a continuous variable: $s_{ijp} = 1 - \frac{|x_{ip} - x_{jp}|}{max(x_{lp}) - min(x_{lp})}$, where the denominator corresponds to the range of values of the variable p across all data points (excluding missing values).

The Gower coefficient is a measure of similarity, which takes values between 0 and 1, while the K-means clustering relies on distance measurements. The Gower similarity coefficient should then be replaced by its dissimilarity measure. The dissimilarity measure between observations i and j based on the Gower coefficient is defined as:

$$g_{ij} = 1 - S_{ij} \tag{3.1.9}$$

In their article, Ali et Massmoudi (2013) compare the K-means clustering algorithm using different metrics. They found out that the Gower and the Manhatan distances perform better than the Euclidean and Chebyshev distances on categorical/mixed data type [21].

Example

Here is a simple example illustrating the calculation of the Gower distance and the comparison of two Gower distances.

	A	B	C
1	Peer	Belgium	4.0
2	Peer	France	6.0
3	Apple	Belgium	8.0

Table 3.1.2: Illustration of the Gower distance - Dataset with 3 observations and 3 variables

The dataset contains 2 categorical variables (A,B) and 1 continuous variable (C). Note that in this example, equal weights are used ($\omega_p = 1$).
The formula of the Gower distance between observation 1 and observation 2:

$$S_{12} = \frac{\sum_{p=1}^{P} s_{12p}\delta_{12p}\omega_p}{\sum_{p=1}^{P} \delta_{12p}\omega_p}$$

For variable A, both observations have the value "Peer", resulting in $\delta_{12A} = 1$ and $s_{12A} = 1$. For variable B, although both observations contain a value, resulting in $\delta_{12B} = 1$, they have different values, which implies $s_{12B} = 0$. For variable C, both observations have a value, so $\delta_{12C} = 1$. As C is continuous, s_{12C} is calculated with the formula defined above, resulting in $s_{12C} = 1 - \frac{|4-6|}{8-4} = 0.5$. The Gower distance between observation 1 and observation 2 is:

$$S_{12} = \frac{(1 \cdot 1) + (0 \cdot 1) + (0.5 \cdot 1)}{(1 + 1 + 1)} = \frac{1.5}{3} = 0.5$$

The same procedure can be applied to calculate the Gower distance between observation 1 and observation 3. Knowing that $s_{13A} = 0$, $s_{13B} = 1$, $s_{13C} = 1 - \frac{|8-4|}{8-4} = 0$ and that $\delta_{13A} = 1$, $\delta_{13B} = 1$, $\delta_{13C} = 1$, the Gower distance between observation 1 and observation 3 is:

$$S_{13} = \frac{(0 \cdot 1) + (1 \cdot 1) + (0 \cdot 1)}{(1 + 1 + 1)} = \frac{1}{3} \approx 0.33$$

From the above calculations, it can be concluded that observations with a higher Gower similarity coefficient (observation 1 and 2) are more similar across the considered variables than observations with a lower Gower similarity coefficient (observation 1 and 3).

K-means using the Gower distance

When using the K-means with the Gower distance, the algorithm is applied on the Gower dissimilarity matrix, it results in a K-medoids problem where centroids (medoids) are actual points of the dataset. Now, the objective function aims to minimize the intra cluster dissimilarities of data points [48]:

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{c}_k) \tag{3.1.10}$$

To update medoids, we swap the medoid with all other points and choose the point as a new medoid that minimizes the objective function. If it reduces the objective function, then the swap is performed [48].

3.1.2 Fuzzy K-means

The K-means algorithm is referred to as "hard clustering" which implies that one observation belongs to one and only one cluster. In contrast, the fuzzy K-means is referred to as "soft clustering" which allows observations to potentially belong to multiple clusters. In this case, a data point is assigned to a cluster based on its membership value which is its degree of belongingness to that particular cluster [25]. The idea behind the Fuzzy k-means algorithm is to minimize the intra-cluster variance:

$$\sum_{i=1}^{n} \sum_{j=1}^{K} (w_{i,k})^m d(\mathbf{x}_i, \mathbf{c}_k) \tag{3.1.11}$$

where $w_{i,k}$ is the membership value of data point $\mathbf{x}_i$ to belong the cluster k and m is the fuzzifier ($m \in \mathbb{R}^+$ and $m > 1$), an exponent controlling the degree of fuzziness in $w_{i,k}$. It is interesting to note the link between the objective function of the K-means and the one of the Fuzzy K-means. The former one aims to minimize the intra-class inertia, which is the sum of squared distances. The latter one aims to minimize the intra-class variance, which is the weighted sum of distances.

The membership value is related to the inverse distance to the centroid of the cluster:

$$w_{ik} = \frac{1}{d(\mathbf{x}_i, \mathbf{c}_k)} \tag{3.1.12}$$

In fuzzy clustering algorithms, the sum of the membership values for one observation should be equal to 1 ($\sum_{k=1}^{K} w_{i,k} = 1$). To ensure this requirement, membership values need to be fuzzyfied using the parameter m and then normalized [25]:

$$w_{ik} = \frac{1}{\sum_{u=1}^{K} (\frac{d(\mathbf{x}_i, \mathbf{c}_k)}{d(\mathbf{x}_i, \mathbf{c}_u)})^{\frac{2}{m-1}}} \tag{3.1.13}$$

The fuzzifier parameter, m, influences the degree of ambiguity in a cluster. When m increases, the degrees of belongingness $w_{i,k}$ become smaller and are more uniformly distributed between clusters. When m approaches 1, the denominator of Eq. 3.1.13 becomes larger and $w_{i,k}$ tends to be either 0 or 1, in function of the relative distances of $\mathbf{x}_i$ to $\mathbf{c}_k$ and to $\mathbf{c}_u$. In this case, the fuzzy K-means acts like a hard clustering algorithm resulting in non-overlapping clusters [25]. The choice of m has been investigated by many studies but no agreement was found. Chan and Cheung (1992) [26] proposed to set the value of m to [1.25:1.75] while Bezdek (1993) demonstrated that $m = 2$ is optimal [27].

The algorithm follows this pseudo-code:

Algorithm 1: Fuzzy k-means clustering algorithm
Input:
$X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ in $\mathbb{R}^p$ //set of n data points
K //a priori fixed number of clusters
m //a priori fuzzier value
Output:
$C = \{C_1, C_2, ..., C_K\}$ // set of K clusters
W // partition matrix of membership values $w_{i,k}$ for i=1,...,n and k=1,...,K
Steps:
1. Choose arbitrarily $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K$ in $\mathbb{R}^p$, the K initial centroids
2. For each i $\in \{1, 2, ..., n\}$, assign randomly $w_{i,k} \in [0; 1]$for each observation
3. Do until convergence:
- For each k $\in \{1, 2, ..., K\}$ calculate the new centroids $\mathbf{c}_k$ (Eq. 3.1.14);
- For each i $\in \{1, 2, ..., n\}$, assign new $w_{i,k}$ for being in clusters (Eq. 3.1.13)
4. Assign each observation to the cluster with the highest membership value.

Table 3.1.3: Pseudo-code for Fuzzy k-means

Finally, centroids can be calculated as follow:

$$\mathbf{c}_k = \frac{\sum_{i=1}^{n} \mathbf{x}_i (w_{ik})^m}{\sum_{i=1}^{n} (w_{ik})^m} \tag{3.1.14}$$

Note that centroids are calculated by including all observations, weighted by their membership value, to allow for the contribution of each data point.

Fuzzy K-means using the Gower distance

When using the fuzzy K-means directly on a dissimilarity matrix, the objective function (Eq. 3.1.11) is transformed so that it no longer depends on centroids [48] :

$$\sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ik}^m w_{jk}^m d(\mathbf{x_i}, \mathbf{x_j})}{2 \sum_{i=1}^{n} w_{ik}^m} \tag{3.1.15}$$

Membership values of each point in each cluster are updated iteratively in order to minimize the objective function that does not depend on centroids anymore. The algorithm stops when the objective function do not decrease further or if a maximum of iteration is reached. The explanation of the minimization of the objective function and the update of membership values are out of the scope of this master thesis, but can be retrieved in [48].

3.1.3 Latent Class Analysis

Latent Class Analysis (LCA) is a useful statistical method to handle multivariate categorical data. The motivation behind it is that the categorical variables, which are observable, can be influenced by non observable variables/classes. It uses probabilities to identify distinct, non-overlapping groups based on an unobserved variable. It estimates two probabilities:

- The probability of an individual to belong to a class. This probability is calculated based on the observed responses of an individual to categorical variables.
- The probability of an individual, who belongs to a class, to give a certain response to a categorical variable.

By calculating these probabilities, LCA enables to identify different groups and to understand the relationships between observed and unobservable variables [28].

In simple words, the basic idea of LCA is to identify the latent variable V_L that can explain all the association between the P observed variables, using the minimal number of classes. This is implied by the assumption of local independence. This assumption means that once the latent class variable is taken into account, the relationships among the observed variables are assumed to be independent [28]. For example, in a study examining substance use among young people, this assumption would imply that the association between alcohol and tobacco use is explained by their association with the latent class variable rather than by a direct causal relationship between alcohol and tobacco use.

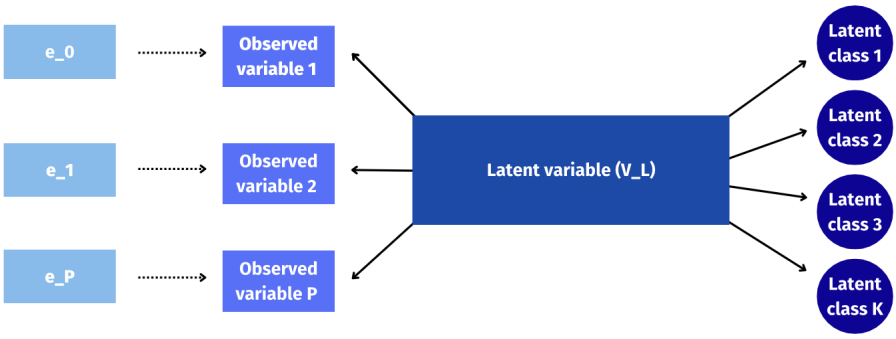

Figure 3.1.2: Latent class modeling - visualization

Figure 3.1.2 illustrates an hypothetical latent class model using P observed variables. The latent variable (V_L) is not directly observable but exists hypothetically with the LCA. It represents the joint unobserved characteristics of the observed variables. Note that LCA can handle measurements errors in continuous variables (e_1, e_2,...,e_P) [29]. However, this out of the scope of this work and will not be considered further. Finally, the latent variable is the

element that differentiates latent classes: two up to K classes can be created from the latent variable [29]. Note that the latent variable is not directly used in the model definition, but through the probabilities of class membership.

Model definition

To explain the model in a more statistical way, Linzer and Lewis (2011) define several equations [15]. First, they define the probability that an individual i, who belongs to a specific class k, generates a certain set of P outcomes on the observed variables is:

$$P(\mathbf{x}_i|\pi_k) = \prod_{p=1}^{P}\prod_{l=1}^{L_p}(\pi_{pkl})^{Y_{ipl}} \tag{3.1.16}$$

where

- $\mathbf{x}_i$ is a set of outcomes for individual i.
- π_k is the individual class membership probability for individual i to belong to class k.
- P is the number of categorical variables.
- L_p is the number of possible categories of variable p.
- x_{ipl} is the observed value of the p-th categorical variable. It is equal to 1 if individual i answers l to variable p, otherwise it is equal to 0 ($p = 1, ..., P$ and $l = 1, ..., L_p$).
- π_{pkl} is the class-conditional probability that an individual belonging to class k will produce the outcome l on variable p.

Note that this formula assumes conditional independence. Then, it defines the probability of observing a specific set of outcomes $\mathbf{x}_i$ for individual i conditionally on the individual class membership probabilities π (vector containing π_k values) and on the class membership probabilities at the population level $\mathbf{p}$ (vector containing p_k values) as:

$$P(\mathbf{x}_i|\pi, \mathbf{p}) = \sum_{k=1}^{K} p_k \prod_{p=1}^{P}\prod_{l=1}^{L_p}(\pi_{pkl})^{x_{ipl}} \tag{3.1.17}$$

where p_k is the class membership probability of class k at the population level [15].

Finally, the Bayes' formula is used to calculate the posterior probability that each individual is associated to each class, conditionally on the observed values of the categorical variables:

$$P(k|\mathbf{x}_i) = \frac{p_k[P(\mathbf{x}_i|\pi_k)]}{\sum_{q=1}^{K} p_q[P(\mathbf{x}_i|\pi_q)]} \tag{3.1.18}$$

The latent class model estimates these two parameters: p_k and π_{pkl}. They are estimated by the maximization of the log-likelihood function with respect to p_k and π_{pkl}:

$$log(L) = \sum_{i=1}^{n} ln \sum_{k=1}^{K} p_k \prod_{p=1}^{P}\prod_{l=1}^{L_p}(\pi_{pkl})^{x_{ipl}} \tag{3.1.19}$$

Two methods are commonly used to maximize the log-likelihood: the Expectation Maximization (EM) algorithm and the Newton-Raphson (NR) algorithm. The EM algorithm works well when the model is influenced by unobserved latent variables. Additionally, it is a stable method, robust to initial values and convergences to a local maximum likelihood [30]. Furthermore, this method is implemented in most software packages, including the R poLCA package. This algorithm is an iterative algorithm which enables the refining of estimates with in the E-step and M-step. The E-step enables to calculate the "missing" class membership probabilities given the current estimates and the observed data. The M-step enables to re-estimate the parameter estimates given the class membership probabilities obtained in the E-step [15]. The algorithm stops when convergence is achieved.

This is the pseudo-code of the Expectation-Maximization algorithm:

Algorithm 3: Expectation-maximization

1. Let $\hat{p}_k^t$ and $\hat{\pi}_{pkl}^t$ denote arbitrary initial values
2. **E-step**: Calculate the class membership probabilities with Eq. 3.1.18 with $\hat{p}_k^t$ and $\hat{\pi}_{pkl}^t$
3. **M-step**: Based on the posterior probabilities calculated in the E-step, update the estimates by maximizing the log-likelihood function with:
   3.1. $\hat{p}_k^{t+1} = \sum_{i=1}^{n} \hat{P}(k|\mathbf{x}_i)$
   3.2. $\hat{\pi}_{pkl}^{t+1} = \frac{\sum_{i=1}^{n} \mathbf{x}_{pi} \hat{P}(k|\mathbf{x}_i)}{\sum_{i=1}^{n} \hat{P}(k|\mathbf{x}_i)}$
4. Repeat steps 2-3 until a stopping criterion is met.

Table 3.1.4: Pseudo-code for the expectation-maximization algorithm

Number of classes

One of the drawbacks of the LCA is the determination of the number of classes. Indeed, as for the K-means and Fuzzy K-means, a number of classes should be fixed in advance. However, LCA proposes different statistical tools that can be used to find this optimal number. There is no consensus on which criterion is the best but the most commonly cited is the Bayesian information criterion (BIC) [15]. The model-fit generally increases with the number of latent classes but there is a potential risk of over-fitting the error in the data. The risk of having an over-fitted model is that it is less generalizable and less replicable. So, to have a balance between under- and over-fitting, there exist parsimony criteria that penalize the number of parameters to be estimated in the log-likelihood. Indeed, models that minimize the BIC or the AIC (Akaike information criterion) are preferred [15]. The main difference between these 2 information criteria is that the BIC penalizes the addition of parameters to the model as a function of sample size, resulting in a preference for simpler models with fewer classes. In opposition, the AIC tends to select more complex models with more classes when n increases. The sample size is not taken into account in the calculation of the AIC [32].

$$BIC = -2\lambda + \phi ln(n) \tag{3.1.20}$$

$$AIC = -2\lambda + 2\phi \tag{3.1.21}$$

where λ is the maximum log-likehood of the model and ϕ the number of parameters.

3.2 Multinomial logistic regression

In the context of this study, multinomial logistic regression is used to assess the association between socio-demographic factors and the clusters identified in the previous analysis. The logistic regression model has become a widely accepted method of analyzing the association between predictor variables and a binary outcome variable. The multinomial logistic regression model is its generalization when the outcome variable has more than 2 possible outcomes [33].

Multinomial logistic regression aims to report the probability of an observation to belong to each category of the dependent variable, given the values of the independent variables. Note that the dependent variable should be nominal while independent variables can be continuous or categorical. The model estimates this probability for each category using a set of parameters, which can be estimated by the maximum likelihood method [33]. The estimated coefficients, their p-values and their confidence intervals enable to determine the significance and direction of the relationship between the independent variables and the dependent variable.

A logit in the context of multinomial logistic is the logarithm of the odds ratio for each category compared to a reference category. To construct the logit with n independent observations (X), P independent variables and one dependent variable (C) with K possible outcomes, one category is chosen as the reference level. It is important to note that any category can be used as the reference level and that all the other logits are created regarding to this reference level. A multinomial logistic regression model can be defined as [32]:

$$P[C = k|X = \mathbf{x}_i] = \frac{\exp\left(\beta_{0k} + \beta_{1k}x_{1i} + ... + \beta_{Pk}x_{Pi}\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\beta_{0k} + \beta_{1k}x_{1i} + ... + \beta_{Pk}x_{Pi}\right)} \tag{3.2.1}$$

where l=1,...,$K-1$

$$P[C = K|X = \mathbf{x}_i] = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\left(\beta_{0k} + \beta_{1k}x_{1i} + ... + \beta_{Pk}x_{Pi}\right)} \tag{3.2.2}$$

The logarithm of the odds ratio of category l compared to the reference category L_d is then:

$$log[\frac{P[C = k|X = \mathbf{x}_i]}{P[C = K|X = \mathbf{x}_i]}] = \beta_{0k} + \beta_{1k}x_{1i} + ... + \beta_{Pk}x_{Pi} \tag{3.2.3}$$

where $k = 1, ..., K-1$.

Parameter estimation

In a multinomial logistic regression, parameters are estimated with the maximum likelihood method (MLE). The idea is to maximize the log-likelihood function of the multinomial logistic regression model [34]:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_{ik} \sum_{p=0}^{P} x_{pi}\beta_{pk} - \log \left(1 + \sum_{p=0}^{P} \exp(x_{pi}\beta_{pj}) \right) \right) \tag{3.2.4}$$

where y_{ik} is an indicator taking value 1 if the i-th observation belong to category k and 0 otherwise. Note that x_{0i} is generally set to 1, to enable the intercept β_{0k} to be estimated separately.

The idea is then to find the $\boldsymbol{\beta}$ that maximizes the Eq. 3.2.4. This can be done using the iterative Newton-Raphson algorithm which involves calculating the first and the second derivatives of the log-likelihood function.

Here is the pseudo-code of the Newton-Raphson:

Algorithm 4: Newton-Raphson

1. Let $\hat{\boldsymbol{\beta}}^t$ denote an initial guess
2. Calculate the gradient (∇l) and the Hessian matrix ($H_{l(\hat{\boldsymbol{\beta}})}$) with respect to $\hat{\boldsymbol{\beta}}^t$
3. Update $\hat{\boldsymbol{\beta}}^t$ with: $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^t - H^{-1}_{l(\hat{\boldsymbol{\beta}})} \nabla l$
4. Repeat steps 2-4 until a stopping criterion is met.

Table 3.2.1: Pseudo-code for the Newton-Raphson algorithm

The Gradient matrix contains all the first partial derivatives of the function, while the Hessian matrix contains all second partial derivatives of the function with respect to the parameters of interest.

Wald statistics

The Wald statistic is one available tool to assess the significance of the coefficients [35]. The idea behind this test is to determine if the value of a coefficient differs significantly from 0. The Wald statistic for a coefficient is:

$$Wald_{pl} = \left(\frac{\hat{\beta}_{pl}}{SE(\hat{\beta}_{pl})} \right)^2 \tag{3.2.5}$$

It follows a χ^2 distribution with 1 degree of freedom. The null hypothesis is rejected if the p-value is lower then a certain threshold (usually 0.05), meaning that the coefficient has a significant impact on the model. This test is easy to calculate but may not give reliable results when the sample size is small [35].

Interpretation

Odds ratio are often used to interpret the outcomes of multinomial logistic regression models. They refer to the probability of the event on the probability of the nonevent ($OR = \frac{\pi}{1-\pi}$). In the case of a multinomial logistic regression, odds ratios correspond to the exponential of the model's coefficients.

A multinomial logistic regression produces $K-1$ (K is the reference level) separate comparisons with their own regression coefficients. When the independent variable p changes by 1 unit while keeping the other independent variables constant, the odds ratio of one specific outcome value C=k over the reference group C=K changes by the corresponding parameter estimate β_{kp}. If $OR > 1.0$ ($OR < 1.0$), an increase (decrease) in the independent variable increases (decreases) the probability of the dependent variable to be in the outcome C=K (compared to the reference level) [34].

3.3 Association analysis

The term "association" can have different definitions in function of the context. Here, it refers to the relationship between 2 qualitative variables [36]. We use this method to analyse the intentions to change unhealthy behavior and the possible association between intentions.

There exist many statistical tests for assessing the relationship between 2 categorical variables and the choice of it mainly depends on the size of the sample and distribution of the data, as well as the way data are paired (refer to Fig 6.0.2 in the Appendix section) [36].

We focus here on the Fisher's Exact Test. The main idea behind this tests is to check the independence between 2 categorical variables that is, to check if the level of the first variable does not determine the level the second variable [36].

3.3.1 Fisher's Exact test

The aim of the Fisher's exact test is to assess if there is an association between 2 categorical variables. The idea behind this test is to determine every possible frequency combinations as the observed table. Then, it calculates its probability under the null hypothesis of independence. Note that this test gives the exact probability, that is why it is said to be exact [36]. This probability is calculated based on an hypergeometric distribution:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3.3.1}$$

where a, b, c and d are the individual frequencies of each cell in a 2x2 contingency table. While considering all possible contingency tables, we maintain row and column totals constant.

The null hypothesis of the Fisher's Exact test states that there is no association, and a low p-value enables to reject this null hypothesis. Note that this test is designed for any sample size but is particularly useful for small sample size. Indeed, it becomes quickly computer intensive when the sample size is large. In this case, a chi-squared test is more appropriate as the approximation works well and is much less computer intensive. However, it is never wrong to use it, even if the sample size is big [36].

3.3.2 Benjamin-Hochberg procedure

When performing simultaneously many tests, multiple comparisons problems arise and increase the probability of obtaining false-positive results. Traditional multiple comparison

procedures are useful tools to control the probability of committing Type I error [37]. However, when the sample size is large, these methods have shown a certain lack of ability to detect real differences. Benjamini-Hochberg criterion has been proposed as a new criterion to control the overall Type I error: the false discovery rate which is the expected proportion of incorrect rejections that is, the proportion of falsely rejected true null hypotheses on the total number of rejected null hypotheses. The Benjamini-Hochberg (BH) procedure controls the overall Type I error by adjusting the p-values and can be defined as follows [37]:

1. Let $\{H_1, H_2,...,H_R\}$ be the R null hypothesis to be tested and let $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(R)}$ be the ordered p-values for $r = 1, ..., R$.
2. Let define a significance threshold for each r null hypothesis as: $q_r = (\frac{r}{R})\alpha$ where α is the desired overall Type I error rate ($\in [0; 1]$).
3. Find the largest index r_0 for which $P_r \leq q_r$.
4. If r_0 exists, we reject the null hypothesis $H_{(2)},...,H_{(r_0)}$.

The Benjamini-Hochberg procedure will be used in this study to correct for multiple comparisons when assessing the associations between (un)healthy behaviors, between (un)healthy behaviors and socio-demographic factors, and between intentions to change unhealthy behaviors.

3.4 Missing values

Missing values are frequent in large-scale studies as ours. They can be encounter in many situations and are referred as item non-response or unit non-response. The former one corresponds to situations in which a part of the answer is missing that is, when a question is skipped in a survey for example. The latter one corresponds to situations in which no answers are available for a subject that was in the sample [38]. Efron (1994) elaborates a broad definition of missing data: it refers to a class of problems made difficult by the absence of some portions of a familiar data structure [38]. Even if missing data has become of interest, handling them is still not a simple task. Indeed, there exist some guidelines on the methods to be used but not for all cases [39] .

3.4.1 Missing data mechanisms

Before looking at strategies to handle msising data, it is important to understand the type of missing values present in the dataset. Three types of missing data can be distinguished: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [41]. In the first case, data is considered to be MCAR when the missingness is unrelated to the observed and unobserved data. A typical example would be a question accidentally skipped by someone in a survey [41]. In the second case, data is considered to be MAR when the missingness is independent of the unobserved data, but is dependent of the observed data. For instance, in a survey, the age is an observed variable and older people are less likely to respond to age-related questions [41]. In the last case, data is considered to be MNAR when the missingness is dependent on unobserved data so, factors that cannot be quantified. For example, it is the case when questions related to sensitive topics (drug

use, income,...) in a survey are skipped because the person is not willing to share a social undesirable answer. The last one is the most frequent case and is also the most complex to handle [41].

3.4.2 Strategies

Selecting a strategy mainly depends on the pattern of missing data. However, it is a difficult task as MAR and MNAR cannot be properly tested and as MCAR is rarely encountered [41]. In this work, the complete case strategy has been investigated as well as two methods of multiple imputation. The main idea behind the complete case strategy is to include in the analysis only observations with available data for all variables. This is a simple strategy to implement but it can reduce significantly the size of the dataset and can lead to bias results when data are not MCAR [41]. The idea behind imputation methods is to infer missing values from the available data. Single imputation fills in each missing values by an estimated value (such as the mean, the regression coefficient, etc.) and produces a single imputed dataset [40]. Multiple imputation differs from single imputation by filling in missing values multiple times, each time with a different estimated value. Several datasets are then produced with different estimated values and are combined at the end to get a single inference. The advantage is that it quantifies the uncertainty and allows for variability. However, it generally cannot handle all the bias introduced by the missing values if data are MNAR [41]. In the case of categorical data, methods of imputation can result in an over-representation of certain categories. This is because, they are often based on classification techniques which assign to missing values the value of the categories considered most likely. This can lead to a bias in imputed data towards the more common categories [41]. A sensitivity analysis is therefore conducted as it reveals the robustness of the results under different data missing assumptions. Ideally, results should be the similar under each scenario to provide reassurance that missing data do not affect the results and are less likely to introduce significant biased [42].

Complete case analysis

The complete case analysis (CCA) is the simplest and the most straightforward method to treat missing values. It is generally known as listwise deletion. Individuals with at least one missing value are excluded from the dataset and only those with no missing values on any variables are retained [41]. The main disadvantage of the CCA is the possible loss of information by discarding incomplete observations. This can lead to a loss of precision in inference and can also induce bias [43]. When data are MCAR, the subset of complete cases is a random sample drawn from the larger population. Estimates derived from it are thus less likely to be biased. However, when data are MAR or MNAR, using the complete case produces biased results. [43]

Multivariate imputation by chained equations

The first imputation method that is investigated is the multivariate imputation by chained equations (MICE). This method does not make a specific assumption about the joint distribution of the data, but relies on the MAR assumption. If data are not MAR, it can result in some bias [44]. It involves running a series of regression models, where each variable with missing data is modeled conditionally on the other variables in the dataset. This allows each variable to be modeled based on its unique distribution, with binary variables being modeled

using logistic regression and continuous variables being modeled using linear regression, for instance [44]. In other words, this approach involves fitting a set of conditional densities to the data instead of a multivariate distribution. The MICE package in R is applied to the data and follows the following steps [44]:

1. Provide an initial estimate for missing values using for example the mode imputation. These are referred as 'place holders'.
2. For one variable, the 'place holder' is put back to missing.
3. This step is a regression step where the variable in step 2 is the dependent variable and all the others are predictors variables.
4. The fitted model constructed in step 3 is used to predict the missing values of the selected variable. An update is done to replace the missing values by the predicted values in the dataset.
5. Steps 2-4 are repeated for each variable containing missing values. This process is repeated a specified number of times or until convergence.

At the end, only final imputations are kept, which form the final imputed dataset. The number of iterations to be performed is usually defined at 10. However, to find the optimal number, further research should be conducted under different assumptions [44].

Random forest

The second imputation method that has been investigated is the random forest (RF) approach described by Golino et Gomes (2016) [39]. A random forest is a non-linear and non-parametric method that can be used for prediction. Its main advantages is that it relies on minimal assumptions about the structure of the data and it can easily be applied on mixed data types and on complex data structures [45]. The main idea behind the random forest algorithm for multiple imputation is to train a model on the observed part and then to predict the missing part. This training and prediction process is repeated iteratively until a stopping criterion is met. The missForest package in R follows the following steps [45]:

1. Provide an initial estimate for missing values (categorical/numerical) in using for example the mode/mean imputation.

2 We store the previous imputed matrix containing the imputed values from the previous iteration.

3 We iterate through the variables containing missing values in an ascending order (so, starting with the ones having the lowest amount of missing values).

3.1 A random forest model is fitted to predict the non-missing values of the variable using the observed values of the other variables.

3.2 The trained random forest model from step 3.1 is used to predict the missing values.

3.3 Update the imputed matrix with the predicted missing values from step 3.2.

4. Update the stopping criterion.

5. Steps 2-4 are repeated until a stopping criterion is met.

Stekhoven and Bühlmann (2012) compares the random forest approach with other multiple imputation methods and in particular, with the MICE method [45]. By applying the different methods on several medical datasets, the random forest for multiple imputation outperformed. In addition to the advantages listed above, Stekhoven and Bühlmann (2012) found that the random forest for multiple imputation can be applied to high-dimensional datasets, while producing good results rapidly. On the other hands, it suffers from a lack of interpretability [45].

Chapter 4

Results

4.1 Data set-up

This chapter describes the analysis of the results obtained from the different parts of the study. First, a descriptive analysis of the missing values and the variables is described. Then, the core analysis is separated into 3 sections: the clustering of unhealthy behaviors, the association with socio-demographic factors and the link with intentions.

4.1.1 Missing values

As described in Section 3.4, missing values are frequently encountered and difficult to handle. So, before starting the main analysis, missing values should be investigated. Our dataset (with 22 variables and 4011 observations) contains 2558 missing values for a total of 88242 values. This means that there are 2.89% of item non-responses, which is a relatively small percentage. However, we are mainly interested in having observations with no items non-responses, that is observations with no missing values. In the dataset, there are 1113 rows containing at least one missing value for a total of 4011. This means that 27.74% of the observations are not complete. Due to this high percentage, missing values cannot be ignored. Moreover, 18 out of the 22 (81,8%) variables contain missing values. The first step is to investigate the missingness distribution mechanism (MAR, MCAR or MNAR).

Figure 4.1.1 and Figure 6.0.3 (in the Appendix section) are two graphs that aim to give a visual description of the missingness. From these 2 graphs, it can be seen that the variable about physical activity (PHY) has the biggest amount of missing values. Indeed, it contains 758 missing values out of the 4011 observations (=18.9%). This can be explained by the difficulty to answer questions in the survey related to physical activity. Undoubtedly, measuring the total number of minutes spent doing moderate and intense physical activity in the past 7 days is not clear-cut. The second variable with the biggest number of missing values is the variable about sedentary behavior (SIT). This is also probably due to the difficulty to report the total minutes remaining sitting per day in the past 7 days. The other variables contain less than 5% of missing values, which is a relatively small percentage.

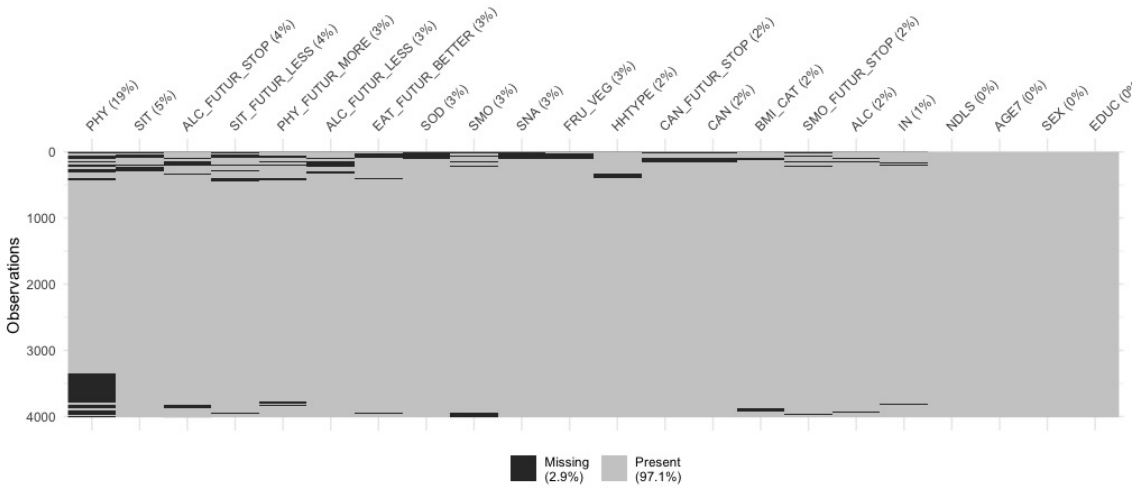

Figure 4.1.1: Missingness visualization

It is also interesting to look at the dependencies between missing values, or in other words, to the missingness pattern. Figure 4.1.2 shows the 5 variables with the largest amount of missing values and their dependencies. The left bar chart shows the number of missing values for the 5 variables. The matrix in the middle shows the intersections between variables: a dot is present when a variable is part of the missing data combination. The top bar chart indicates the number of observations having this particular missing value combination. Missing data are certainly not MCAR as they do not seem to be randomly dispersed between variables (Figure 4.2.1). However, it is difficult to determine whether it is MAR or MNAR based only on the observed data. Indeed, missing data are unknown by definition, and it cannot be used to completely assess whether the missingness depends on the missing data itself (MNAR) or only on the observed data (MAR) [41].

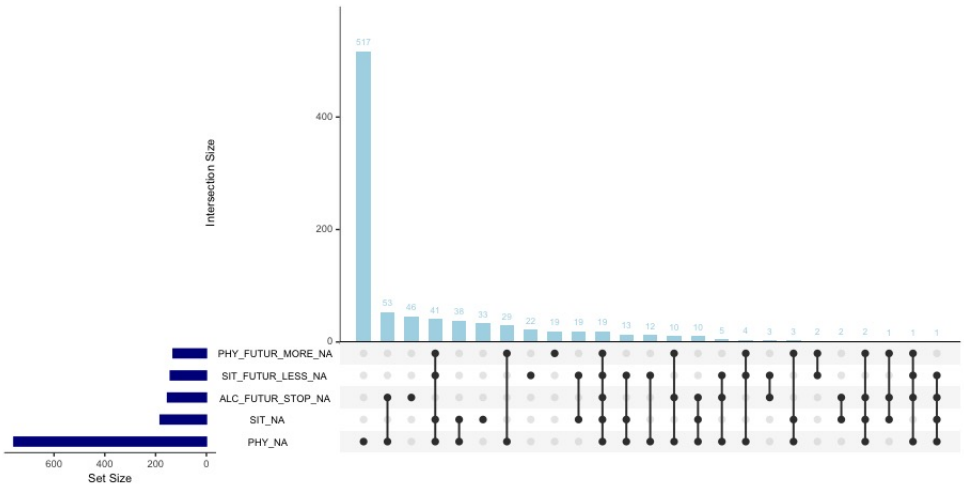

Figure 4.1.2: Missingness pattern

As described in Section 3.4, a sensitivity analysis is conducted in order to compare different methods to handle missing values: MICE algorithm, Random Forest algorithm and Complete

Case analysis. Each of them creates a new dataset on which clustering methods are applied. In order to distinguish the new datasets, a name is assigned to them: mice dataset, forest dataset and complete dataset. Results from the clustering methods will be presented for all three datasets to determine the most suitable one for further analyses.

4.1.2 Descriptive analysis

The first fundamental step in any statistical analysis is to explore conscientiously the data. In Section 2.2, 31 out of the 84 categorical variables have been selected to suit the purpose of the study. Note that intentions variables will be explored in Section 4.4. For categorical variables, one needs to explore their frequencies. Categories with a low number of observations can cause issues if included in the model. Indeed, they give insufficient information about their distribution. The initial sample size is 4011 however, 1169 observations contain at least 1 missing value.

Health related behaviors

Table 4.1.1 displays the outcomes of the descriptive statistics for the health related behaviors of the sample (refer to Table 2.2.1 to have the definition of (un)healthy behaviors). Among all the selected (un)healthy behaviors, daily snacking is the most prevalent. Indeed, 26.2% of the participants report eating snack everyday. In terms of fruit and vegetable consumption, 25.4% of those surveyed reported insufficient consumption. More than a fifth of the sample is inactive: 22.7% are physical inactive and 20.9% engage in a sedentary behavior. Daily soda consumption represents 14.5% of the population while smoking concerns 13.6% of the sample. Finally, a tenth of the sample reported to have a risky alcohol consumption (10.1%) and less then one twentieth (4.4%) reported cannabis usage within the past 12 months.

Behaviors	n(%)	Behaviors	n(%)
Cannabis		Sitting	
yes	175 (4.4%)	yes	2990 (74.6%)
no	3755 (93.6%)	no	839 (20.9%)
no answer	81 (2.0%)	no answer	182 (4.5%)
Smoking		Soda	
yes	546 (13.6%)	yes	581 (14.5%)
no	3352 (83.6%)	no	3315 (82.6%)
no answer	113 (2.8%)	no answer	115 (2.9%)
Alcohol		Snack	
yes	407 (10.1%)	yes	1051 (26.2%)
no	3541 (88.3%)	no	2848 (71.0%)
no answer	63 (1.6%)	no answer	112 (2.8%)
Physical activity		Fruits or vegetables	
yes	2344 (58.4%)	yes	2890 (72.1%)
no	909 (22.7%)	no	1018 (25.4%)
no answer	758 (18.9%)	no answer	103 (2.6%)

Table 4.1.1: Health behaviors of the sample

One important objective of this study is to investigate multiple risk behaviors. Figure 4.1.3 shows the original sample distribution in terms of the unhealthy behavior combinations. Most people in the sample engage in only one unhealthy behavior (35.08%). The unhealthy behavior with the highest prevalence in the 1-risky behavior category is daily snacking. It is followed by the no risky behavior category. Indeed, based on all their answers to the survey, 25.90% of the subjects reported to have no unhealthy behaviors. The 2-unhealthy behaviors still represents a significant percentage of the sample (22.14%). The combination with the highest prevalence is the inactive lifestyle: being physically inactive and having a sedentary behavior. The 3-risky behaviors category accounts for 11.17% of the participants. The combination with the highest prevalence is: being physical inactivity, being sedentary and insufficient fruits and vegetables daily consumption. The 4-, 5-, 6- and 7-risky behaviors categories constitute a small minority of the population, containing respectively 4.04 %, 1.37 %, 0.22% and .05% of the participants. Finally, only one person engages in all the explored unhealthy behaviors. Figure 6.0.4, Figure 6.0.5 and Figure 6.0.6 in the Appendix section show respectively the distribution of the unhealthy behavior combinations in the forest sample, the mice sample and complete sample. The main difference lies in the percentage of the no risk behavior category, which is lower in the mice and complete samples than in the other two samples (original and forest samples). In addition, the mice and complete samples contain a slightly higher percentage of people engaging in two unhealthy behaviors.

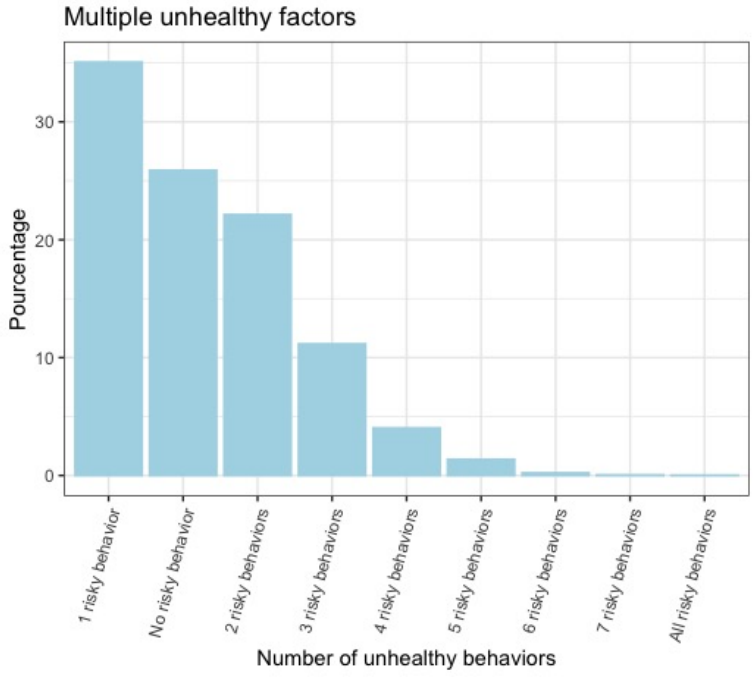

Figure 4.1.3: Original sample distribution in terms of unhealthy behaviors combinations

Socio-demographic factors

Table 4.1.2 displays the outcomes of the descriptive statistics for the socio-demographic characteristics of the sample. The original sample contains the same proportion of men and women (49.5% and 50.3%). All age categories are significantly represented. However, most of the par-

ticipants are between 55 and 64 years old (19.8%). Concerning the BMI, 45.6% fall in the normal category, while the majority of the participants is classified as overweight or obese. Most participants affirmed to speak Dutch at home (93.9%). All levels of education are well represented: 40.9% of the participants have a diploma of higher education (higher education level), 33.9% have a diploma of higher secondary education (middle education level) and 25.2% have a diploma of the second grade of secondary education or lower (lower education level). More then a quarter of the sample (28.6%) report having difficulties to live with their household income. At last, the majority of the participants live as a couple or as a couple with children (67.8%).

Demographics	n(%)	Demographics	n(%)
Sex		Education	
Male	1986 (49.5%)	lower level	1010 (25.2%)
Female	2018 (50.3%)	middle level	1359 (33.9%)
Other	7 (0.2%)	higher level	1642 (40.9%)
Age		Income	
18–24	376 (9.4%)	great difficulties	97 (2.4%)
25–34	545 (13.6%)	difficulties	278 (6.9%)
35–44	580 (14.5%)	some difficulties	774 (19.3%)
45–54	655 (16.3%)	fairly easily	1373 (34.3%)
55–64	794 (19.8%)	easily	1111 (27.7%)
65–74	619 (15.4%)	very easily	329 (8.2%)
75+	442 (11.0%)	no answer	49 (1.2%)
BMI		Household type	
underweight	88 (2.2%)	alone	542 (13.5%)
normal	1830 (45.6%)	alone + child(ren)	141 (3.5%)
overweight	1359 (33.9%)	couple	1490 (37.1%)
obesity	662 (16.5%)	couple + child(ren)	1232 (30.7%)
no answer	72 (1.8%)	with parent(s), etc.	521 (13.1%)
Dutch			
yes	3767 (93.9%)		
no	244 (6.1%)		

Table 4.1.2: Demographics characteristics of the sample

Associations between health behaviors and socio-demographic factors

To determine if categorical variables are associated, a Chi-Squared test is performed. Table 4.1.3 displays the corrected p-values of the tests between health behavior and socio-demographic variables of the original dataset. As explained in Section 3.3, if the p-value is lower then a certain threshold, the hypothesis of independence should be rejected. This tables shows that only 9 pairs of variables present no significant association at a 0.05 level: *household type-snack*, *education-physical activity*, *bmi-smoking*, *dutch-cannabis*, *dutch-smoking*, *dutch-alcohol*, *dutch-sitting*, *sex-snack* and *age-snack*. Same conclusions can be drawn by performing this test on the other datasets (refer to Table 6.0.2, Table 6.0.3 and Table 6.0.4 in the Appendix section).

Variables	H. type	Educ.	Income	BMI	Dutch	Sex	Age
Cannabis	<0.001*	<0.001*	<0.001*	<0.001*	0.348	<0.001*	<0.001*
Smoking	<0.001*	<0.001*	<0.001*	0.652	0.652	<0.001*	<0.001*
Alcohol	<0.001*	<0.001*	<0.001*	<0.001*	0.489	<0.001*	<0.001*
Phy.Act.	<0.001*	0.457	<0.001*	<0.001*	0.038*	<0.001*	<0.001*
Sitting	<0.001*	<0.001*	<0.001*	0.014*	0.211	<0.001*	<0.001*
Fru.Veg.	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
Snack	0.197	<0.001*	<0.001*	<0.001*	<0.001*	0.062	0.325
Soda	<0.001*	<0.00*1	0.001*	0.001*	<0.001*	<0.001*	<0.001*

Table 4.1.3: Association between variables of the original dataset: p-value of the Chi Squared test

In Table 6.0.5 in the Appendix section, relationships between health behavior variables are evaluated. All corrected p-values fall under 0.05, indicating association between these variables.

4.2 Objective 1 - Clustering of unhealthy behaviors

We perform clustering analysis on the original dataset where 8 health related behaviors are selected (in Section 2.2). Note that in order to treat missing values, the 3 methods described in Section 3.4 are applied on the original dataset resulting in 3 new datasets. A sensitivity analysis is then conducted to determine if there are significant differences in the results when handling missing values using various techniques. Specifically, it consists of applying clustering methods to the 3 new datasets (the mice dataset, the forest dataset and the complete dataset) and comparing their results.

4.2.1 K-means clustering

We run the K-means algorithm using the Gower dissimilarity measure on the 3 datasets. It is important to note that the algorithm is not directly applied on the datasets, but on the Gower dissimilarity matrices which contain the distance between all pairs of data points. We decided to use equal weights in the distance calculation (recall Section 3.1.2) since data are of the same type and we do not have prior knowledge to assigned different weights to variables. Before looking at the results, we determine the optimal K value using the average silhouette method.

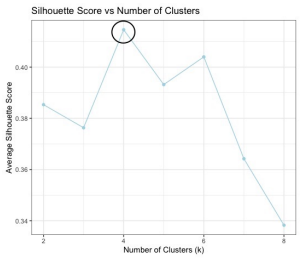

Figure 4.2.1: Average silhouette - Complete dataset for K-means

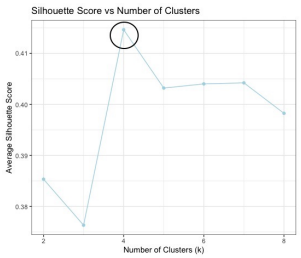

Figure 4.2.2: Average silhouette - Mice dataset for K-means

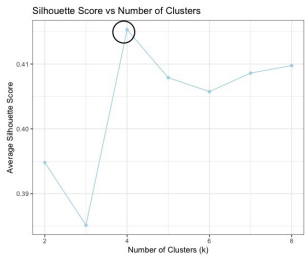

Figure 4.2.3: Average silhouette - Forest dataset for K-means

The above graphs illustrate the average silhouette method for each datasets. Recall that higher silhouette values correspond to well-clustered data points (refer to Section 3.1.1). In each case, it determines K=4 as the optimal number of clusters. The K-means algorithm can be then conducted on each datasets with K=4 as input.

Comparing the results of the K-means applied on each dataset enables to learn more about the sensitivity of the data and the clusters to missing values. Figure 4.2.4, Figure 4.2.5 and Figure 4.2.6 display the percentage of individuals belonging to a particular cluster who engage in unhealthy behaviors. By looking broadly at the graphs, four similar clusters are found, with small variations in the percentages, regardless of the scenario.

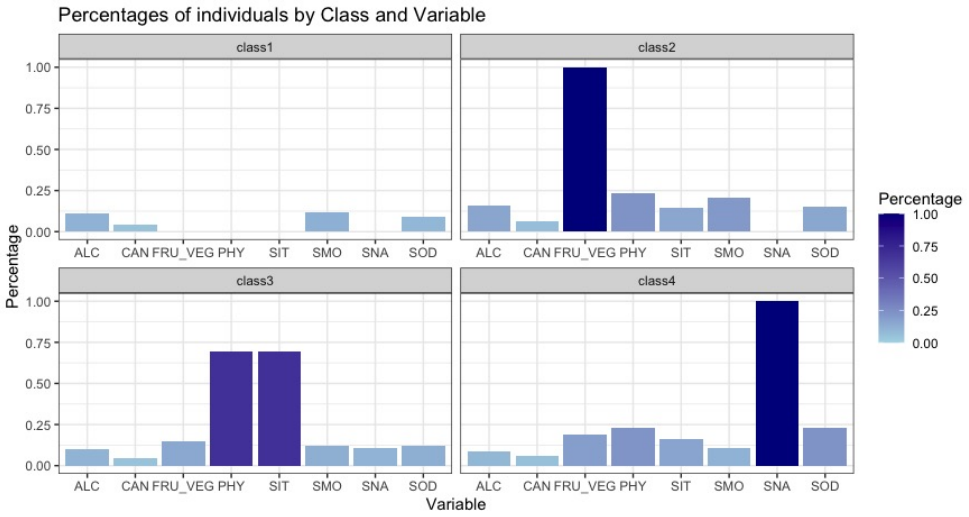

Figure 4.2.4: Percentages of individuals engaging in unhealthy behaviors within each clusters - Complete dataset for K-means

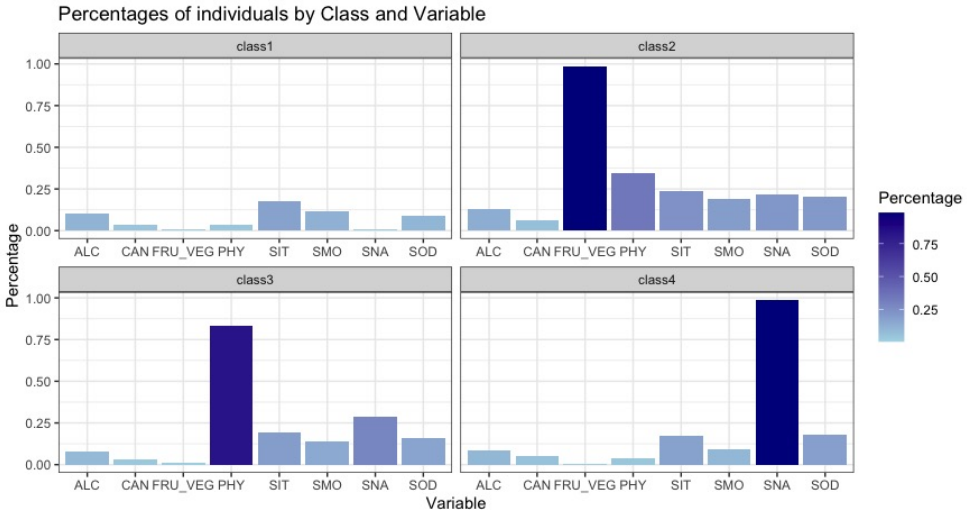

Figure 4.2.5: Percentages of individuals engaging in unhealthy behaviors within each clusters - Mice dataset for K-means

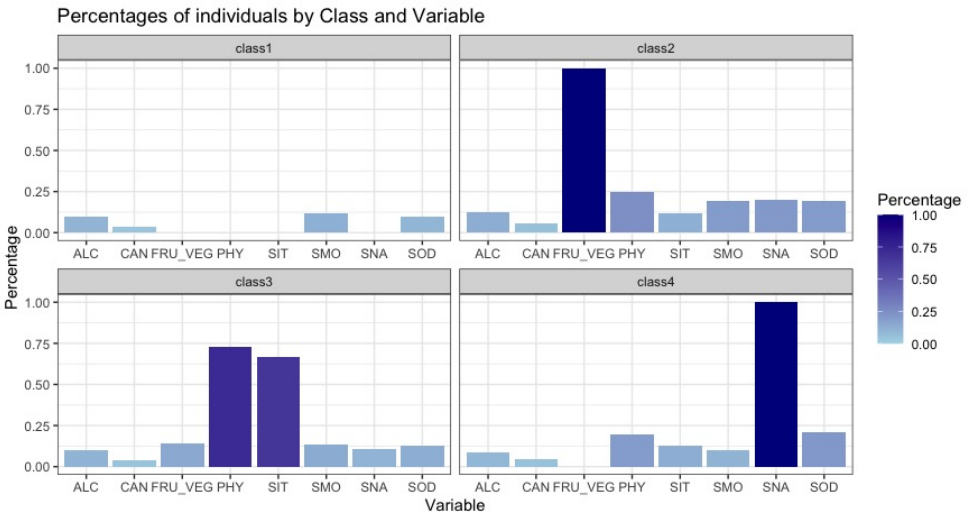

Figure 4.2.6: Percentages of individuals engaging in unhealthy behaviors within each clusters - Forest dataset for K-means

Based on the above graphs and on Table 6.0.6, Table 6.0.7 and Table 6.0.8 in the Appendix section, clusters using the K-means algorithm can be described as:

Cluster 1: This cluster contains 932 individuals under the complete dataset (31.98%), 1562 under the mice dataset (38.94%) and 1338 under the forest dataset (33.36%). The cluster is characterized by small percentages of individuals engaging in unhealthy behaviors.

Cluster 2: This cluster contains 508 individuals under the complete dataset (17.43%), 1047 under the mice dataset (26.10%) and 911 under the forest dataset (22.71%). Individuals belonging to this cluster are mainly characterized by insufficient fruits and vegetables consumption.

Cluster 3: This cluster contains 760 individuals under the complete dataset (26.08%), 781 under the mice dataset (19.47%) and 973 under the forest dataset (25.26%). Individuals belonging to this cluster are mainly characterized by inactive behaviors (sedentary and physical inactivity behaviors). It can be seen that under the mice dataset, a smaller proportion of individuals engage in sedentary behavior (around 20.00%).

Cluster 4: This cluster contains 714 individuals under the complete dataset (24.50%), 621 under the mice dataset (15.48%) and 789 under the forest dataset (19.67%). Individuals belonging to this cluster mostly engage in daily snacking.

As a conclusion of the K-means analysis part, results are consistent under the 3 different datasets. However, in order to validate this conclusion, the sensitivity analysis should also be conducted under the 2 other clustering methods.

4.2.2 Fuzzy K-means

As explained in Section 3.1.2, the fuzzy K-means differs from the K-means by the possibility for a data point to belong to several clusters. Recall that the fuzzy K-means method assigns cluster membership values to observations. Then, observations are allocated to the cluster with the highest membership value. This method gives potentially better results for overlapped dataset than the K-means [25]. We run the algorithm with a fuzzier parameter of $m = 1.5$ and the number of classes $K = 4$. Recall that when m approaches 1, the method tends to act like a hard clustering approaches. However, setting $m > 1.5$ has revealed that some clusters have no data points exclusively assigned to them, while some individuals have non-zero membership values for these empty clusters. The following results are obtained:

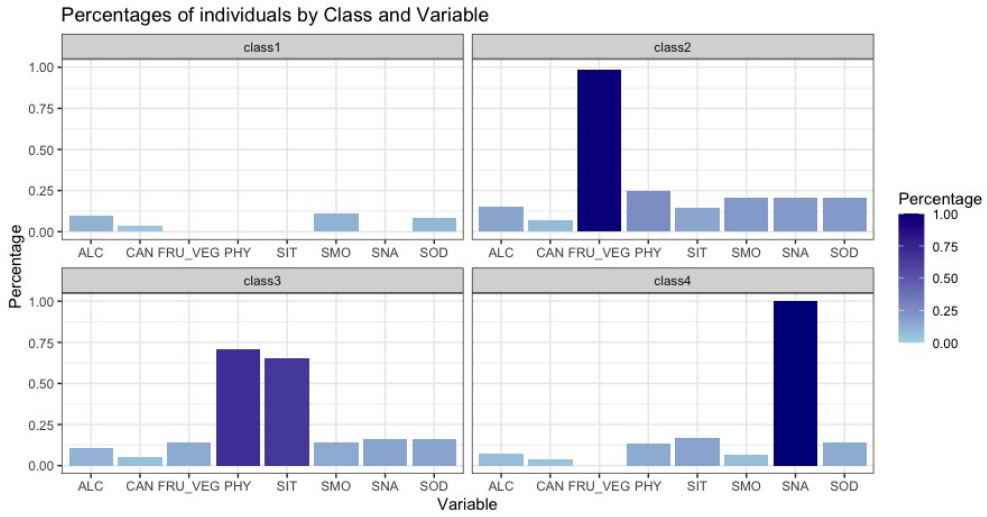

Figure 4.2.7: Percentages of individuals engaging in unhealthy behaviors within each clusters - Complete dataset for fuzzy K-means

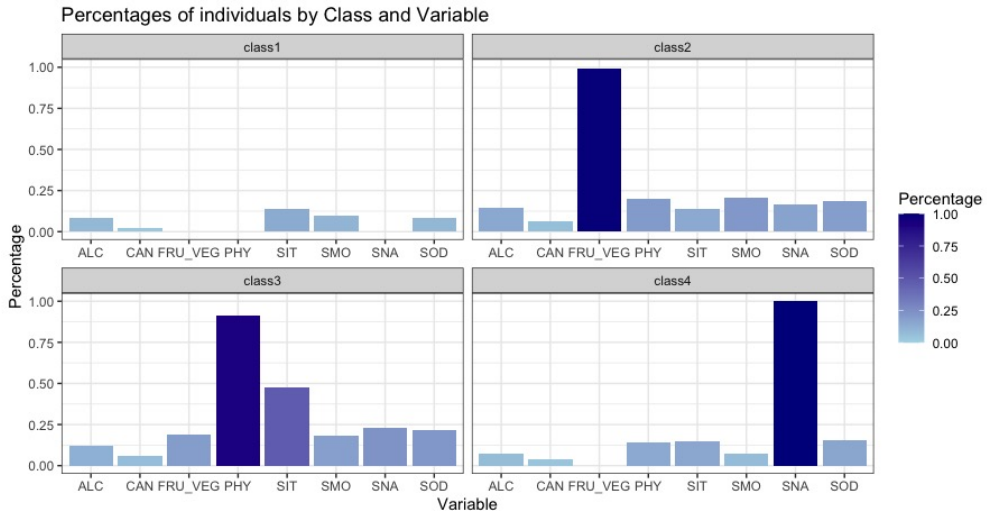

Figure 4.2.8: Percentages of individuals engaging in unhealthy behaviors within each clusters - Mice dataset for fuzzy K-means

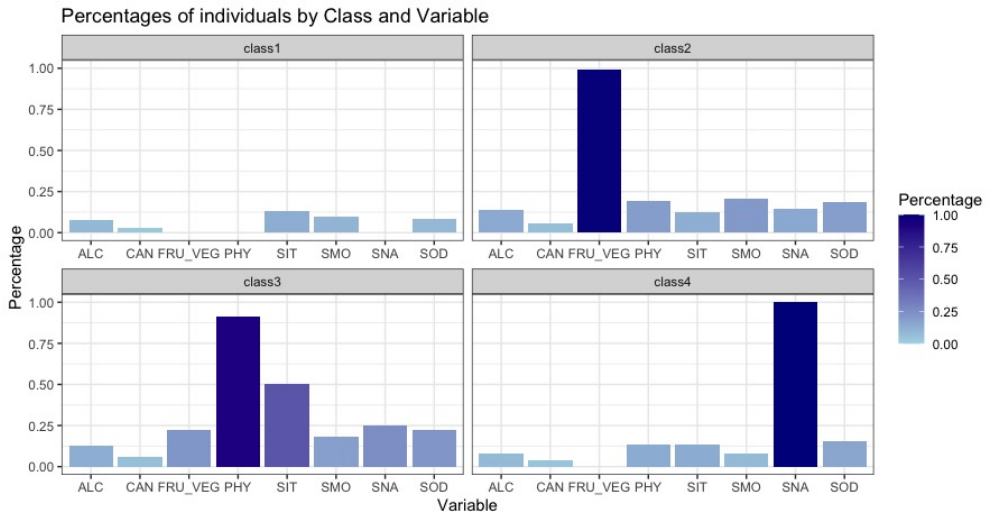

Figure 4.2.9: Percentages of individuals engaging in unhealthy behaviors within each clusters - Forest dataset for fuzzy K-means

These graphs present the percentages of individuals who engage in unhealthy behaviors within each cluster under the different datasets. Based on these graphs and on Table 6.0.9, Table 6.0.10 and Table 6.0.11, this method gives clusters which can be described as:

Cluster 1: This cluster 923 individuals under the complete dataset (31.67%), 1479 under the mice dataset (36.87%) and 1526 under the forest dataset (38.05%). It is characterized by small percentages of individuals engaging in unhealthy behaviors.

Cluster 2: This cluster contains 650 individuals under the complete dataset (22.31%), 877 under the mice dataset (21.86%) and 861 under the forest dataset (21.47%). This class is mainly characterized by an insufficient fruits/vegetables consumption.

Cluster 3: This cluster contains 813 individuals under the complete dataset (27.90%), 939 under the mice dataset (23.41%) and 899 under the forest dataset (22.41%). This class is characterized by an inactive lifestyle (insufficient physical activity and sedentary behavior).

Cluster 4: This cluster contains 528 individuals under the complete dataset (18.12%), 716 under the mice dataset (17.85%) and 725 under the forest dataset (18.08%). It is mainly characterized by eating snacks daily.

It can be seen that results obtained with the K-means and the fuzzy K-means are very similar. Indeed, the only major difference lies in the variation in the proportion of the sedentary behavior under the mice dataset: 19.3% when using the K-means while 47.5% when using the fuzzy K-means. Therefore, we can conclude that there is minimal overlap between clusters.

4.2.3 Latent class analysis

In this section, we present the results of the LCA applied on our 3 different datasets. By examining the conditional probabilities of categorical variables within each class, we obtain information on the characteristic profiles of the identified cluster. Indeed, in the context of the LCA, probabilities indicate the likelihood an individual within a specific class of engaging in unhealthy behaviors. Before obtaining the results, we need to determine the most appropriate number of classes, using essential measures to identify the best model for our sample, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

Dataset	AIC	BIC	N smallest class
Complete dataset			
3 classes	20920.6	21076.01.85	234
4 classes	20838.82	21088.02	224
5 classes	**20800.05**	**21063.05**	101
Forest dataset			
3 classes	28393.07	28556.78	306
4 classes	28220.2	28440.59	294
5 classes	**28156.88**	**28433.94**	112
Mice dataset			
3 classes	28575.07	28738.79	319
4 classes	28481.22	28765.6	277
5 classes	**28432.61**	**28709.67**	121

Table 4.2.1: Validation indices - LCA

For each dataset, the 5-class model produces the best model fit according to information criteria as their values decrease with the number of classes. Note that in this analysis, LCA with 6 classes gives higher values for the information criterion and less readable classes. Once again, comparing the results under the 3 datasets is part of the sensitivity analysis about missing data. The following graphs and the tables in the Appendix section (Table 6.0.12, Table 6.0.13 and Table 6.0.14) present the probabilities of unhealthy behaviors within each class under the different scenarios. Regardless of the dataset, the same types of class can be found.

Cluster 1: The share of the population in this class is estimated at 8.03% under the complete dataset, 11.4% under the mice dataset and 8.23% under the forest dataset. This class is characterized by people with poor diet habits (daily snacking, soda consumption and insufficient fruits and vegetables consumption). Note that there is a slight variation in the probabilities of the soda (SOD) and snack (SNA) variables when comparing the different data sets. In the complete and forest datasets, snacks account for around 60% and soft drinks for 100%. However, in the mice dataset, snacks represent 51.99% and soft drinks 62.42%.

Cluster 2: The share of the population in this class is estimated at 13.73% under the complete dataset, 13.72% under the mice dataset and 12.60% under the forest dataset. This class is characterized by an inactive lifestyle and by insufficient fruits and vegetables consumption. Note that there are differences in the probabilities for the physical activity and sedentary (PHY lower and SIT higher) variables when using the mice dataset compared to the other datasets. These differences can be explained by a higher number of missing values for these variables.

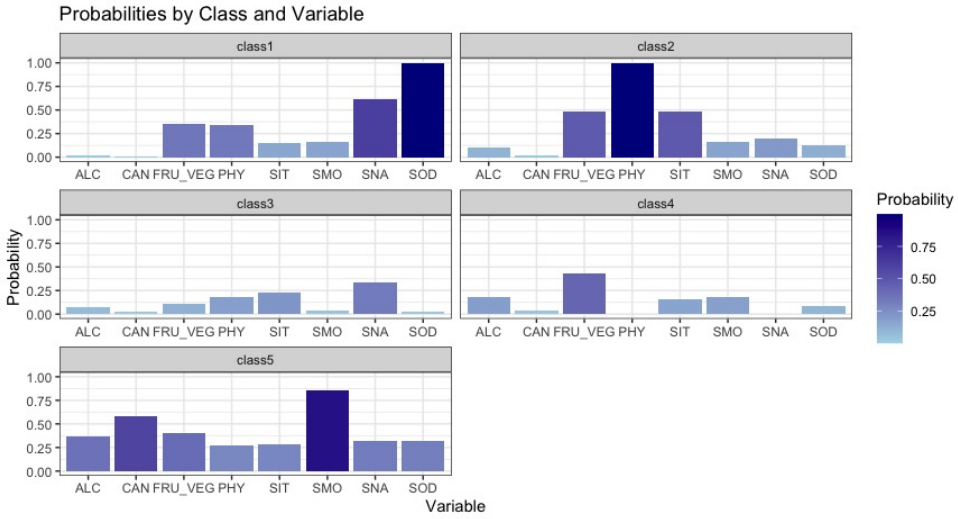

Figure 4.2.10: Probabilities of unhealthy behaviors to belong to clusters - Complete dataset for LCA

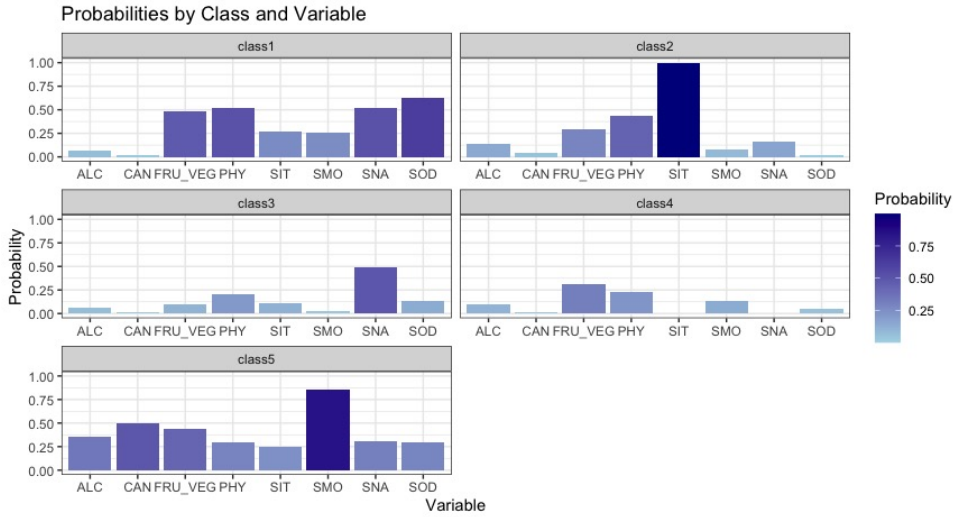

Figure 4.2.11: Probabilities of unhealthy behaviors to belong to clusters - Mice dataset for LCA

Cluster 3: The share of the population in this class is estimated at 54.17% under the complete dataset, 34.88% under the mice dataset and 54.76% under the forest dataset. This class is characterized by small probabilities of engaging in unhealthy behaviors. Note that there is an exception: about a quarter of people in this cluster eat snack daily.

Cluster 4: The share of the population in this class is estimated at 19.35% under the complete dataset, 13.72% under the mice dataset and 20.71% under the forest dataset. This class is characterized by small probabilities of engaging in unhealthy behaviors. However, one half of people in this cluster have an insufficient consumption of fruits and vegetables.

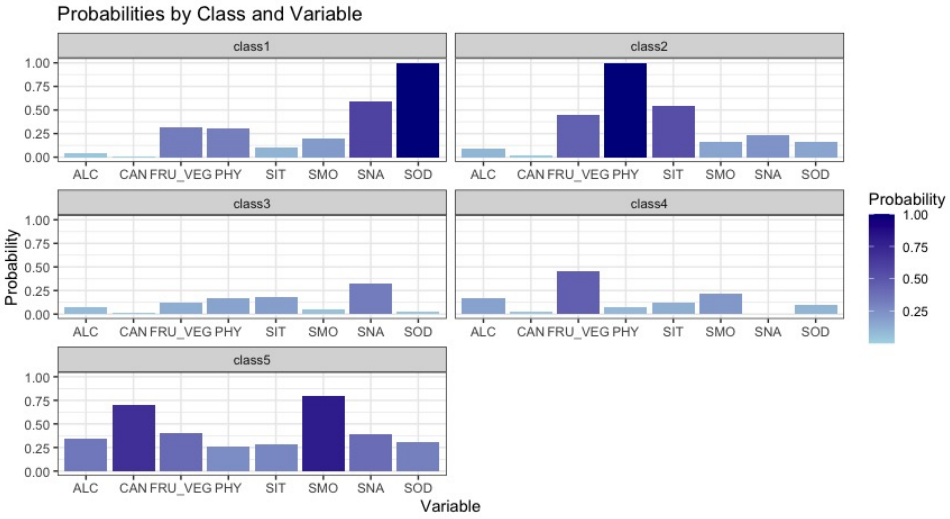

Figure 4.2.12: Probabilities of unhealthy behaviors to belong to clusters - Forest dataset for LCA

Cluster 5: The share of the population in this class is estimated at 4.72% under the complete dataset, 5.00% under the mice dataset and 3.73% under the forest dataset. This class is characterized by higher probabilities of engaging in unhealthy behaviors. Note that this cluster is specifically defined by substance usage (cannabis, alcohol and smoking).

4.2.4 Model selection and cluster profiles

Sensitivity analysis

There is unfortunately no agreement on how to summarize the information from a sensitivity analysis into a decision [46]. It is important to recall the assumptions under which the methods are applied. The MICE method relies on the MAR assumption while the complete case analysis method relies on the MCAR assumption (refer to Section 3.3). As the random forest method for imputation is a non-parametric method, it makes less assumptions about the data and is more robust to the missing data patterns. From the results of the clustering techniques, it can be seen that they are similar. Indeed, under each scenario, similar clusters are found, with only slight variations in probabilities. However, it is important to note that the probabilities for the sedentary and physical activity variables differ somewhat under the three datasets. The most noticeable difference is in cluster 2 of the LCA, where sedentary behavior constitutes 99,97% of the individuals in the mice dataset, 53.80% in the forest dataset and 83.80 in the complete dataset. Additionally, physical inactivity accounts for 43.93% in the mice dataset while it accounts for 100% in both of the complete and forest datasets. This can be explained by the fact that these two variables contains more missing values. From this sensitivity analysis, we can conclude that missing data do not have a significant influence on the results and on the behavioral groups. For the rest of the analysis, we decide to work under the random forest scenario as it relies on less assumptions and enables to work on the entire dataset.

Model comparison

Direct comparison between the K-means, the fuzzy K-means and the LCA can be challenging due to their fundamental different approach. Indeed, there is no agreement on common objective criteria for assessing the validity of clustering methods across all scenarios. It is therefore important to refer to the assumptions of the models, the interpretability of the results and the objective of the study to determine the method to be used [47]. The K-means and the fuzzy K-means are methods originally designed for clustering continuous variables even if they can be adapted to mixed/categorical variables by using the Gower distance. In contrast, the LCA is designed to handle multivariate categorical variables. In terms of interpretability, the K-means and fuzzy K-means methods provide a direct interpretation through centroids, while the LCA provides a better understanding of the data by identifying latent classes and probabilistic association with categorical variables. Additionally, the major difference between the results of the different methods is that the LCA results in 5 clusters, in which one of them is characterized by the unhealthy diet behaviors. This cluster is interesting for the objective of the study as this group can provide useful information for targeted intervention policies to tackle unhealthy dietary habits. Therefore, the LCA is the choosen clustering method for the rest of the analysis.

Cluster profiles

Clusters can be named on the basis of unhealthy behaviors with a high probability of conducting the behavior as:

- The **healthy lifestyle** is the largest cluster, containing 2246 people with a lower risk of engaging in unhealthy behaviors. The percentages of people engaging in cannabis use, smoking, alcohol consumption, soda consumption and insufficient fruits/vegetables consumption are relatively low (respectively 1.95%, 2.05%, 8.15%, 0.01% and 3.12%). Sedentary behavior, physical inactivity and snacking are prevalent in this cluster, accounting respectively for 15.89%, 15.72% and 30.10% of the cluster.

- The **unhealthy diet** cluster includes 437 individuals with a higher risk of engaging in unhealthy diet habits. Within this cluster, 100% of individuals consume soda daily, 49.66% consume snacks daily and 28.83% have an insufficient consumption of fruits/vegetables. Additionally, 26.31 % engage in insufficient physical activity and 21.51% are current smokers. Very few people in this cluster engage in sedentary behavior, excessive alcohol consumption or cannabis use.

- The **substance usage** cluster is the smallest clusters, representing 112 people. The majority engage in cannabis use (95.53%), smoking (87.50%) and alcohol consumption (37.5%). The percentages of people engaging in the other unhealthy behaviors remain significant, ranging from 27.67% to 42.85%.

- The **mixed lifestyle** cluster includes 629 individuals who have mixed healthy behaviors. In this cluster, percentage of cannabis use is low (2.70%), while a significant part of the individuals engage in alcohol consumption (18.76%) and smoking (31.32%). Globally, people in this cluster lead active lives, with relatively low percentage being sedentary (12.72%) and none being physically inactive. Although few people consume daily soda (4.29%) or snacks (0.00%), a majority eat insufficient fruits/vegetables (77.90%).

- The **inactive lifestyle** cluster contains 587 individuals. All individuals engage in insufficient physical activity and 63.03% engage in sedentary behavior. A majority eats insufficient fruits/vegetables (54.00%) and a significant proportion consumes snacks (23.33%) and soda (14.48%) daily. Smoking concerns 18.91% of the cluster, while alcohol consumption and cannabis use concern respectively 10.39% and 1.36% of the cluster.

Figure 4.2.13 enables an intuitive illustration of the distribution of the health related behaviors of the sample among the different identified clusters. This visualization highlights the distinct patterns. For instance, in the unhealthy diet cluster, it can be seen that light blue is prevalent (corresponding to an unhealthy behavior) for the diet variables (snack, soda and fruits & vegetables consumption). The displayed numbers give an idea of the size of the clusters. However, it is important to note that the conditional probabilities slightly differ from the real percentage of individuals assigned to a specific cluster on the basis of the highest probability.

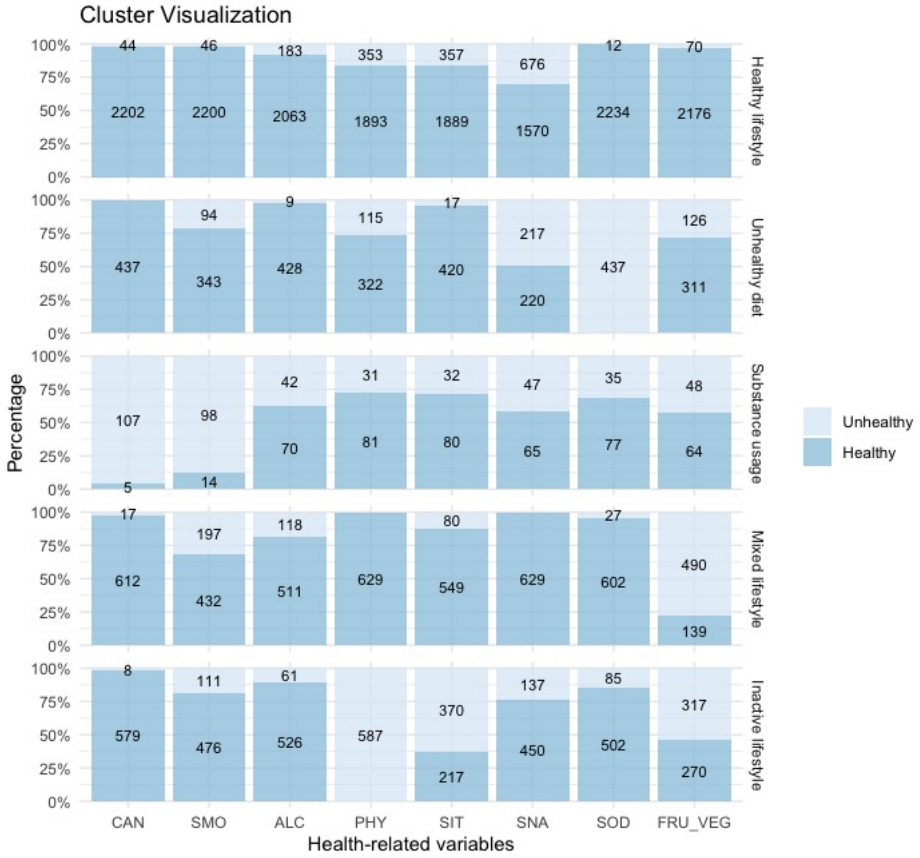

Figure 4.2.13: Distribution of health related behaviors of the sample within each cluster

4.3 Objective 2 - Cluster characterisation

From now, we consider clusters obtained with the LCA under the forest dataset. We perform multinomial logistic regression to analyse the association of socio-demographic factors with the clusters. Cluster 1, which corresponds to the healthy lifestyle group, is taken as the reference group and is compared to the other clusters to assess the relationship between socio-demographic variables and clusters. The reference categories for each socio-demographic variable were chosen on the basis of their size and interest. Table 4.3.1 shows the results of this analysis by displaying the odds ratios and their 95% confidence interval. To assess the significance of parameters, the Wald test is performed and bolded results indicate statistical significance at a 0.05 level. To understand these results, it is important to bear in mind that these odds ratios must be interpreted in terms of comparison with the reference group (healthy lifestyle) and the reference categories for each variable.

Demographics	Unhealthy diet	Substance usage	Mixed lifestyle	Inactive lifestyle
Sex				
Male	1 (reference)	1 (reference)	1 (reference)	1 (reference)
Female	**0.44 (0.35-0.55)**	**0.31 (0.20-0.49)**	**0.42(0.35-0.51)**	**0.78 (0.64-0.94)**
Other	0.38 (0.04-3.66)	1.25 (0.19-13.11)	0.40 (0.04-3.69)	0.001 (<0.01->99)
Education				
lower level	1.32 (0.99-1.75)	**0.29 (0.11-0.77)**	1.06 (0.83-1.36)	1.19 (0.91-1.56)
middle level	1 (reference)	1 (reference)	1 (reference)	1 (reference)
higher level	**0.37 (0.28-0.49)**	**0.34 (0.21-0.54)**	**0.53 (0.42-0.67)**	0.85 (0.67-1.07)
Age				
18–24	**1.95 (1.07-3.57)**	**2.75 (1.08-6.94)**	1.22 (0.70-2.11)	1.42 (0.81-2.49)
25–34	**1.62 (1.08-2.43)**	**2.56 (1.20-5.46)**	1.28 (0.88-1.86)	**1.66 (1.17-2.36)**
35–44	1.15 (0.79-1.67)	1.94 (0.94-4.01)	1.06 (0.76-1.49)	1.27 (0.92-1.75)
45–54	1 (reference)	1 (reference)	1 (reference)	1 (reference)
55–64	**0.62 (0.20-0.51)**	**0.20 (0.06-0.67)**	1.08 (0.79-1.49)	0.90 (0.65-1.25)
65–74	**0.32 (0.26-0.67)**	0.19 (<0.01->99)	0.70 (0.48-1.01)	**0.39 (0.26-0.59)**
75+	**0.42 (0.11-0.42)**	<0.01 (<0.01->99)	**0.42 (0.27-0.65)**	**0.63 (0.42-0.95)**
Income				
great difficulties	**2.12 (1.12-4.00)**	**3.53 (1.08-11.50)**	1.31 (0.69-2.49)	**2.42 (1.41-4.14)**
difficulties	1.23 (0.80-1.89)	**2.40 (1.00-5.73)**	1.414 (0.98-2.03)	1.23 (0.84-1.80)
some difficulties	**1.369 (1.03-1.81)**	1.24 (0.66-2.33)	**1.32 (1.02-1.71)**	1.11 (0.84-1.45)
fairly easily	1 (reference)	1 (reference)	1 (reference)	1 (reference)
easily	**0.74 (0.55-0.98)**	1.25 (0.75-2.09)	0.94 (0.74-1.19)	0.89 (0.69-1.13)
very easily	**0.39 (0.22-0.69)**	1.22 (0.59-2.54)	0.87 (0.60-1.27)	0.97 (0.68-0.1.40)
BMI				
underweight	1.00 (0.43-2.31)	1.87 (0.71-4.90)	1.19 (0.63-2.27)	1.41 (0.76-2.63)
normal	1 (reference)	1 (reference)	1 (reference)	1 (reference)
overweight	1.25 (0.97-1.61)	1.34 (0.84-2.13)	1.06 (0.86-1.31)	**1.49 (1.19-1.86)**
obesity	**1.35 (1.00-1.84)**	0.16 (0.06-1.11)	0.87 (0.66-1.15)	**1.90 (1.47-2.47)**
Household type				
alone	0.95 (0.66-1.37)	**2.47 (1.17-5.23)**	**1.34 (1.00-1.79)**	**1.87 (1.40-2.49)**
alone + child	1.10 (0.59-2.06)	1.00 (0.26-3.84)	**1.93 (1.20-3.10)**	1.45 (0.85-2.45)
couple	1 (reference)	1 (reference)	1 (reference)	1 (reference)
couple + child	1.13 (0.82-1.54)	0.65 (0.31-1.34)	0.96 (0.74-1.26)	1.26 (0.96-1.66)
with parent(s), etc.	1.12 (0.68-1.86)	**2.18 (1.01-14.69)**	1.33 (0.85-1.36)	1.21 (0.77-1.91)
Dutch				
yes	1 (reference)	1 (reference)	1 (reference)	1 (reference)
no	**0.54 (0.31-0.92)**	1.30 (0.61-2.77)	1.17 (0.81-1.69)	1.27 (0.88-1.82)

Table 4.3.1: Socio-demographic factors associated with clusters - OR(95%)

Concerning sex, women have a significantly lower odds ratio in all four comparisons. In the unhealthy diet cluster, women are less likely to belong to it compared to men (OR = 0.44, 95%CI: 0.35-0.55). Similarly, in the substance usage cluster, women have a significantly lower odds ratio (OR = 0.31, 95%CI: 0.205-0.49). For the mixed lifestyle cluster, women also show

a lower odds ratio (OR = 0.42, 95%CI: 0.35-0.51), indicating a lower probability of being part of this cluster. At last, women also show a lower odds ratio compared to men in the inactive lifestyle cluster (OR = 0.78, 95%CI: 0.64-0.94).

Age is another factor of important influence. In the unhealthy diet cluster, younger age categories (18-24 and 25-34) have higher odds ratios, indicating a higher probability of belonging to this cluster, while older age categories (55-64, 65-74, and 75+) show lower odds ratios, indicating a lower probability of belonging the unhealthy diet cluster. Similar conclusions can be drawn in the substance usage cluster, where the 18-24 and 25-34 age categories have higher odds ratios, and the 55-64 age category shows a lower odds ratio. For the mixed lifestyle cluster, only the 75+ age category shows a significant lower odds ratio (OR = 0.42, 95%CI: 0.27-0.65). In the inactive lifestyle cluster, the 65-74 and 75+ age categories have lower odds ratios, while the 24-35 age category shows a higher odds ratio (OR = 1.66, 95%CI: 1.17-2.36).

Education level is associated with the probability of belonging to different clusters. In the unhealthy diet cluster and in the mixed, individuals with a higher education level have a significantly a lower odds ratio compared to those with a middle education level (OR = 0.37, 95%CI: 0.28-0.49 and OR = 0.53, 95%CI: 0.42-0.67 respectively). For the substance usage cluster, higher and lower educated individuals are 3 times less likely to be in this cluster (OR = 0.34 and 0R = 0.29 respectively).

Income has also influence on clusters. The unhealthy diet cluster is inversely related to the ability to end month: individuals having great difficulties and difficulties to end the month show higher odds ratios to belong to this cluster, while individuals who can at least easily end month show lower odds ratios for this cluster. It also have an influence on the substance usage cluster with people having great difficulties and difficulties to end month show a higher probabilities for this cluster (higher odds ratios). In the inactive lifestyle cluster, individuals with great difficulties to end the month also have a higher odds ratio (OR = 2.42, 95%CI: 1.41-4.14).

Obesity is associated with the unhealthy diet cluster, as individuals who are obese have a higher odds ratio (OR = 1.35, 95%CI: 1.00-1.84). Additionally, individuals who are overweight or obesity have more probabilities to belong to the inactive lifestyle cluster (higher odds ratios).

Household type is a another factor of influence. Individuals living alone or living with parents, friends, etc. have a higher odds ratio to belong to the substance usage cluster (OR = 2.47 and OR = 2.18 respectively). Additionally, individuals living alone are also more likely to be in the mixed lifestyle and in the inactive lifestyle clusters (OR = 1.34 and OR = 1.87 respectively). Concerning individuals living alone with child(ren), they have a higher odds ratio for the mixed lifestyle cluster (OR = 1.93, 95%CI: 1.20-3.10).

Finally, language spoken at home has a significant influence on the unhealthy diet cluster: individuals who do not speak Dutch at home show a significantly lower odds ratio (OR = 0.54, 95%CI: 0.31-0.92).

Conclusion

From the results of the multinomial logistic regression and the interpretation of the odds ratio, several tendencies can be emphasize:

- Sex is a determinant feature for the characterization of the lifestyles. Women have significant a lower odds ratio than men to be in the unhealthy diet, substance usage, mixed lifestyle or inactive lifestyle cluster. This indicates that women are more likely to engage in healthy behaviors.
- The age category seems to have an important influence. The odds ratios suggest that engaging in an unhealthy diet and substance usage are inversely related to the age: younger age categories have higher odds ratios than the reference age category (45-54 years old). Additionally, older individuals, especially those who are 65+, are less likely to belong to the mixed lifestyle and inactive lifestyle clusters. This conclusion is unexpected and can be explained by the fact that the physical activity and sedentary variables contain more missing values. Indeed, Fig 6.0.7 and Fig 6.0.8 (in the Appendix section) show that non-item responses increase with age for the physical activity and sedentary variable respectively (more marked on the physical activity variable).
- The education level also plays a role in the cluster characterization. Higher educated individuals have lower odds ratios for the unhealthy diet, mixed lifestyle and substance usage clusters. This indicates that higher educated individuals tend to belong to the healthier cluster. Lower educated individuals also have a lower odds ratio for substance usage cluster, which indicates that this cluster is characterized by middle educated individuals.
- The income variable is inversely related to the substance usage and inactive lifestyle clusters as odds ratios are higher for individuals having more difficulties to end months. For the unhealthy diet cluster, individuals with some and great difficulties to end months are more likely to be part of it. On the other hand, individuals who can easily or very easily end months have a lower probability of belonging to it.
- The household type is a significant determinant as well. Individuals living alone are more likely to be part of the substance usage, mixed lifestyle and inactive clusters, while individuals living with parents, friends etc. are more likely to be part of the substance usage cluster.
- The unhealthy diet cluster is influenced by the Dutch variable, with individuals not speaking Dutch at home are less likely to be part of it. This can be explained by the significant association between the Dutch speaking at home variable and the education level variable. Indeed, there is a higher proportion of higher educated individuals who do not speak Dutch at home (45.48%) than do speak it (40.64%).

In summary, lifestyle clusters are influenced by the above socio-demographic factors. It is interesting to note that being a woman, being older, having no difficulties to meet the end of the month, having a higher education level and not living alone are factors that decrease the probability of engaging in an unhealthy lifestyle. Table 4.3.5 displays the distribution of socio-demographic variables within each cluster. These results are in coherence with the outcomes of the multinomial logistic regression. To have a visual illustration of this, refer to Figure 6.0.7 until Figure 6.0.13 in the Appendix section.

Demographics	Healthy lifestyle	Unhealthy diet	Substance usage	Mixed lifestyle	Inactive lifestyle
Sex					
Male	958 (42.65%)	271 (62.01%)	78 (69.64%)	394 (62.64%)	285 (48.55%)
Female	1284 (57.17%)	165 (37.76%)	33 (29.46%)	234 (37.20%)	302 (51.45%)
Other	4 (0.18%)	1 (0.23%)	1 (0.89%)	1 (0.16%)	0 (0%)
Age					
18–24	156 (6.77%)	62 (14.18%)	43 (38.39%)	66 (10.49%)	49 (8.35%)
25–34	265 (11.80%)	67 (15.33%)	30 (26.79%)	82 (13.04%)	101 (17.21%)
35–44	304 (13.54%)	67 (15.33%)	22 (19.64%)	85 (13.51%)	102 (17.38%)
45–54	357 (15.89%)	81 (18.54%)	13 (11.61%)	102 (16.22%)	102 (17.38%)
55–64	448 (19.95%)	75 (17.16%)	4 (3.57%)	148 (23.53%)	119 (20.27%)
65–74	424 (18.88%)	43 (9.84%)	0 (0%)	100 (15.90%)	52 (8.86%)
75+	292 (13.00%)	42 (9.61%)	0 (0%)	46 (7.31%)	62 (10.56%)
Education					
lower level	547 (24.35%)	135 (30.89%)	5 (4.46%)	176 (27.98%)	147 (25.04%)
middle level	654 (29.12%)	198(45.31%)	69 (61.61%)	251 (39.90%)	187 (31.86%)
higher level	1045 (46.53%)	104 (23.80%)	38 (34.82%)	202 (32.11%)	253 (43.10%)
Income					
great difficulties	37 (1.65%)	17 (3.89%)	4 (3.57%)	15 (2.38%)	29 (4.94%)
some difficulties	144 (6.41%)	35 (8.01%)	8 (7.14%)	57 (9.06%)	49 (8.35%)
fairly easily	385 (17.14%)	115 (26.32%)	18 (16.07%)	145 (23.05%)	118 (20.10%)
easily	788 (35.08%)	160 (36.61%)	32 (28.57%)	206 (32.75%)	195 (33.22%)
very easily	679 (30.23%)	95 (21.74%)	38 (33.93%)	160 (25.44%)	146 (24.87%)
BMI					
underweight	49 (2.18%)	7 (1.60%)	6 (5.36%)	13 (2.07%)	14 (2.39%)
normal	1093 (48.66%)	180 (41.19%)	65 (58.04%)	289 (45.95%)	227 (38.67%)
overweight	742 (33.04%)	163 (37.30%)	36 (32.14%)	234 (37.20%)	210 (35.78%)
obesity	362 (16.12%)	87 (19.91%)	5 (4.46%)	93 (14.79%)	136 (23.17%)
Household type					
alone	287 (12.78%)	50 (11.44%)	18 (16.07%)	95 (15.10%)	109 (18.57%)
alone + child	69 (3.07%)	16 (3.66%)	3 (2.68%)	35 (5.56%)	24 (4.09%)
couple	967 (43.05%)	143 (32.72%)	15 (13.39%)	228 (36.25%)	171 (29.13%)
couple + child	697 (31.03%)	146 (33.41%)	21 (18.75%)	176 (27.98%)	208 (35.43%)
with parent(s)	226 (10.06%)	82 (18.76%)	55 (49.11%)	95 (15.10%)	75 (12.78%)
Dutch					
yes	2121 (94.43%)	420 (96.11%)	103 (91.96%)	584 (92.85%)	539 (91.82%)
no	125 (5.57%)	17 (3.89%)	9 (8.04%)	45 (7.15%)	48 (8.18%)

Table 4.3.2: Distribution of demographics characteristics within each cluster

4.4 Objective 3 - Link with intentions

The idea behind this third analysis is to check how intentions are associated within each cluster. In other words, do people have the intention to change multiple unhealthy behaviors in a particular cluster. As described in Section 3.3, the Fisher's Exact tests is used to assess the relationship between the intention variables. Recall that the presence of the "Not applicable" category indicates that individuals who do not engage in unhealthy behavior with regard to alcohol, smoking and cannabis do not need to change, since they already show healthy behaviors (see Section 2.2.3 to recall to who these questions are applicable). As far as cannabis use is concerned, there are differences between the number of people with this unhealthy behavior and the number of people for whom the question is applicable. Unhealthy cannabis behavior is defined as cannabis use within the past 12 months, whereas the intention question was only asked to people who use cannabis at least once a month. The intention question therefore does not apply to people who use cannabis less than once a month. Note that the association analysis takes into account the "Not applicable" category.

Cluster 1: Healthy lifestyle
Table 4.4.1 displays frequencies of the future intentions to change unhealthy behaviors. It can be seen that a majority of individuals belonging this cluster wants to improve their eating habits (68.34%) and to be more physically active (71.37%). Willingness to make improvements in sedentary behaviour concerns 43.54% of the individuals. With regard to intentions to change substance use, not all of the individuals in the cluster are involved, as the majority do not engage in any of these behaviors. Alcohol consumption involves 1850 individuals, smoking involves 31 individuals and cannabis use involves 4 individuals. There is a difference between the intention to stop drinking and the intention to cut down, with 6.65% and 23.68% among the concerned individuals having this intention respectively. At last, 54.83% of the concerned individuals are willing to stop smoking and none among the 4 wants to stop cannabis. In addition to these information, Table 6.0.15 in the Appendix section provides an understanding of the interactions between future intentions.

N= 2246	Yes	No	Not applicable
Cannabis	0 (0.00%)	4	2242
Smoking	17 (54.83%)	14	2215
Alcohol - Reduction	438 (23.68%)	1412	396
Alcohol - Stop	123 (6.65%)	1727	396
Phy. Act.	1603 (71.37%)	643	/
Sitting	978 (43.54%)	1268	/
Diet	1535 (68.34%)	711	/

Table 4.4.1: Future intentions to change behaviors within cluster 1 - Frequency table

As explained above, the interest of this analysis is to check association between intention variables within a cluster. Table 4.4.2 exposes the p-values of the Fisher's Exact test corrected for multiplicity (refer to Section 3.3.2).

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	1	1	0.768	1	1	1
Smoking		1	1	0.042*	0.034*	0.179
Alc. Red.			<0.001*	<0.001*	<0.001*	<0.001*
Alc. Stop				<0.001*	<0.001*	<0.001*
Phy.Act.					<0.001*	<0.001*
Sitting						<0.001*

Table 4.4.2: Future intentions within cluster 1: P-value of the Fisher's Exact test

In cluster 1, the healthy lifestyle cluster, there are 12 out of the 21 pairs of future intention variables that are associated at a 0.05 level. Intention to stop cannabis does not have any significant relationship with the other variables. This can be explained by the fact that a too small number of individuals are concerned by cannabis use to detect any significant associations, resulting in a low statistical power. Future intention to stop smoking is only related to physical activity and sitting intentions. Finally, the rest of the intention variables (alcohol reduction, stop alcohol, physical activity, sedentary and diet) are associated with one another.

Cluster 2: Unhealthy diet

Table 4.4.3 enables to get an overview from the intention variables by displaying frequencies of the future intentions to change unhealthy behaviors within the unhealthy diet cluster. This cluster focuses on unhealthy diet with 57.9% of the individuals want to improve their diet. However, it can be seen that more individuals are willing to be more physically active (68.6%). Improving sedentary behavior is still significant, representing 38% of the individuals. Individuals in this cluster are less involved in substance use: 299 engage in alcohol consumption, 72 engage in smoking and none engage in cannabis use. Among the concerned individuals, 22.07% want to reduce alcohol consumption, 9.70% want to stop alcohol consumption and 40.28% want to stop smoking. In addition to these information, Table 6.0.1Â§ in the Appendix section provides an understanding of the interactions between future intentions.

N= 437	Yes	No	Not applicable
Cannabis	0 (/)	0	437
Smoking	29 (40.28%)	43	365
Alcohol - Reduction	66 (22.07%)	270	101
Alcohol - Stop	29 (9.70%)	270	101
Phy. Act.	300 (68.65%)	137	/
Sitting	166 (37.99%)	271	/
Diet	253 (57.89%)	184	/

Table 4.4.3: Future intentions to change behaviors within cluster 2 - Frequency table

Table 4.4.4 exposes the p-values of the Fisher's Exact test corrected for multiplicity for the unhealthy diet cluster in order to detected potential associations.

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	1	1	1	1	1	1
Smoking		0.899	0.987	0.293	0.012*	<0.001*
Alc. Red.			<0.001*	0.072	<0.001*	0.026*
Alc. Stop				0.987	0.032*	0.869
Phy.Act.					<0.001*	<0.001*
Sitting						<0.001*

Table 4.4.4: Future intentions within cluster 2: P-value of the Fisher's Exact test

In this cluster, 9 out of the 21 intention variables pairs are associated with each other at a 0.05 level. Improving sedentary and diet behaviors are related to all the other intention variables (expect cannabis and cannabis/alcohol stop respectively). Not surprisingly, alcohol reduction and stopping are related. Note that cannabis use does not concern any individuals, so it is not possible to detect potential association with other variables.

Cluster 3: Substance usage

Table 4.4.5 gives an overview from the intention variables by displaying frequencies of the future intentions to change unhealthy behaviors within cluster 3. A significant number of individuals in this cluster are also willing to change their physical activity (78.6%), sitting (35.7%) and diet (66.1%) habits. Cannabis use concerns 50 individuals, 30.00% of them intending to stop. Smoking concerns 51 individuals, 50.98% of them intending to stop. Alcohol consumption concerns 106 individuals, 4.72% of them wanting to stop, while 25.47% want to reduce their consumption. In addition to these information, Table 6.0.17 in the Appendix section provides an understanding of the interactions between future intentions.

N= 112	Yes	No	Not applicable
Cannabis	15 (30.00%)	35	62
Smoking	26 (50.98%)	25	61
Alcohol - Reduction	27 (25.47%)	79	6
Alcohol - Stop	5 (4.72%)	101	6
Phy. Act.	88 (78.63%)	24	/
Sitting	40 (35.72%)	72	/
Diet	74 (66.17%)	38	/

Table 4.4.5: Future intentions to change behaviors within cluster 3 - Frequency table

Table 4.4.6 exposes the p-values of the Fisher's Exact test corrected for multiplicity for the substance usage cluster in order to detected potential associations.

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	0.167	0.035*	0.039*	0.851	0.147	0.256
Smoking		0.930	0.930	0.187	0.187	0.025*
Alc. Red.			<0.001*	0.0014*	0.130	0.014*
Alc. Stop				0.230	0.471	0.471
Phy.Act.					<0.001*	<0.001*
Sitting						0.001*

Table 4.4.6: Future intentions within cluster 3: P-value of the Fisher's Exact test

In this cluster, there are 9 intention variable pairs out of the 21 that are associated at a 0.05 level. Cannabis and alcohol intentions are associated. Additionally, intention to reduce alcohol is related to intention to stop alcohol, to be more physically active and to have better a diet behavior. Once again, intention to change physical activity, sitting and diet habits are associated with each other.

Cluster 4: Mixed lifestyle

Table 4.4.7 gives an overview from the intention variables by displaying frequencies of the future intentions to change unhealthy behaviors within cluster 4. It can be seen that 71.2% of individuals have the intention to be more physically active and 70.9% have the intention to improve their diet habits. More than a fourth (45.5%) wants to improve its sedentary behavior. Concerning alcohol consumption, 25.3% of the concerned individuals (534) want to reduce it and 6.4% want to stop it. Smoking concerns 129 individuals, with 39.53% of them want to quit. Finally, cannabis use concerns 7 individuals and none of them has the intention to stop it. In addition to these information, Table 6.0.18 in the Appendix section provides an understanding of the interactions between future intentions.

N= 629	Yes	No	Not applicable
Cannabis	0 (0.00%)	7	622
Smoking	51 (39.53%)	78	500
Alcohol - Reduction	159 (25.30%)	375	95
Alcohol - Stop	40 (6.40%)	494	95
Phy. Act.	448 (71.26%)	181	/
Sitting	286 (45.52%)	343	/
Diet	446 (70.97%)	183	/

Table 4.4.7: Future intentions to change behaviors within cluster 4 - Frequency table

Table 4.4.8 exposes the p-values of the Fisher's Exact test corrected for multiplicity for the mixed lifestyle cluster in order to detected potential associations.

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	0.780	0.733	0.780	0.733	0.780	0.396
Smoking		0.022*	0.087	0.027*	0.010*	0.010*
Alc. Red.			<0.001*	<0.001*	<0.001*	0.003*
Alc. Stop				<0.001*	<0.001*	0.024*
Phy.Act.					<0.001*	<0.001*
Sitting						<0.001*

Table 4.4.8: Future intentions within cluster 4: P-value of the Fisher's Exact test

In this cluster, 14 out of the 21 intention variable pairs are related. Diet intention, sedentary intention, physical activity intention and alcohol reduction intention are all related to each of the other intention variables (except the cannabis intention). Additionally, intention to quit smoking and intention to stop alcohol are significantly associated with the other intention variables (except with the cannabis intention). Note that once again, cannabis use concerns a too small number of individuals to assess significant association.

Cluster 5: Inactive lifestyle

Table 4.4.9 gives an overview from the intention variables by displaying frequencies of the future intentions to change unhealthy behaviors within cluster 5. A significant part of the individuals in this cluster is willing to be physically more active (75.3%), be less sedentary (49.4%) and want to improve their diet habits (68.1%). This cluster contains 469 individuals engaging in alcohol consumption and among those, 4.26% have the intention to stop it while 22.39% want to reduce it. Smoking concerns 70 individuals and 40.00% of them want to quit it. Finally, cannabis use concerns only one person who does not have the intention to stop it. In addition to these information, Table 6.0.19 in the Appendix section provides an understanding of the interactions between future intentions.

N= 587	Yes	No	Not applicable
Cannabis	0 (0.00%)	1	586
Smoking	28 (40.00%)	42	517
Alcohol - Reduction	105 (22.39%)	364	118
Alcohol - Stop	20 (4.26%)	449	118
Phy. Act.	442 (75.39%)	145	/
Sitting	290 (49.45%)	297	/
Diet	400 (68.13%)	187	/

Table 4.4.9: Future intentions to change behaviors within cluster 5 - Frequency table

Table 4.4.10 exposes the p-values of the Fisher's Exact test corrected for multiplicity for the inactive lifestyle cluster in order to detected potential associations.

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	1	1	1	1	1	1
Smoking		0.201	0.018*	0.040*	0.744	0.113
Alc. Red.			<0.001*	0.156	0.243	0.031*
Alc. Stop				0.128	0.243	1
Phy.Act.					<0.001*	<0.001*
Sitting						<0.001*

Table 4.4.10: Future intentions within cluster 5: P-value of the Fisher's Exact test

In this cluster, there are 7 out of the 21 intention variable pairs that are associated at a 0.05 level. Intention to quit smoking is related to intention to stop alcohol and to be more physically active. Intention to reduce alcohol is associated to intention to have better diet habits and to stop alcohol. As in the other clusters, intention to change physical activity, sedentary and diet habits are associated with each other. Once again, alcohol reduction is related to quitting alcohol and smoking.

Conclusion

In examining the behavior change intentions in each cluster, it appears that there are similarities in the percentages of individuals wanting to change their behavior across the different clusters:

- Regardless of the cluster, more than 65.00% individuals want to improve their diet habits.
- The percentages of individuals with the willingness to be physically more active ranges from 68.65% to 78.63% across clusters. Note that the higher percentage concerns the inactive lifestyle cluster.
- The percentages of individuals with the willingness to improve sedentary behavior ranges from 35.72% to 49.45% across clusters. Note that the highest percentage concerns the inactive lifestyle cluster.
- The percentages of concerned individuals with the willingness to stop smoking ranges from 39.53% to 54.83% across clusters.
- Individuals engaged in cannabis use concerns mainly the substance usage cluster, with 30.00% of them having the intention to stop it. In the other clusters, none of the individuals involved in cannabis use have the intention to change.
- Across all clusters, there is a significant difference between the intention to reduce alcohol consumption and the intention to stop it. In each cluster, there is a higher percentage of the concerned individuals expressing the intention to reduce alcohol consumption (ranging from 22.07% to 25.47% across all clusters) compared to the intention to completely stop it (ranging from 4.26% to 9.70% across all clusters).

Although the percentage of individuals having the intention to change a specific behavior may be similar across clusters, the association between intentions shows both similarities and dissimilarities across clusters. Physical activity, sedentary behavior and diet habits intentions show systematic association across clusters, indicating an interconnection between these behavior change intentions. Additionally, alcohol reduction and stop alcohol intentions are consistently related. However, as it has been discussed, some dissimilarities between clusters can be noticed, which highlights specific combinations of intentional relationships within each cluster. It is interesting to point out that cannabis intention is associated to other intention variables only within the substance usage cluster due to the too small number of cannabis users in the other clusters. The mixed lifestyle cluster includes specific associations between smoking intention and four other intentions (alcohol reduction, physical activity, sedentary behavior and diet habits). In addition, the combination of smoking and eating habits is also found in the unhealthy diet cluster. These examples illustrate specific relevant combination of intentions within clusters.

Chapter 5

Study limitations

This study presents a number of limitations which should be mentioned and taken into consideration in future research. First, findings are highly influenced by the definitions of unhealthy behaviors and the cut-off used. Although thresholds are chosen based on national guidelines and on decisions in other studies, using these thresholds discards a lot of information. Indeed, it does not take into account contextual factors, individual's differences,... Moreover, dichitomizing continuous variables has an impact on the analysis. Often, dichotomization is not needed for statistical purpose but simplifies the interpretation. It results in a loss of information and a significant variability within each group. For example, individuals staying 7.5 hours per day seated and those staying 8.5 hours are located on either side of the cut-off (8 hours) and are classified in the two opposite groups, while they are rather similar. In addition, having only categorical variables has an impact on the statistical methods. Indeed, clustering methods using continuous data are more widely developed than for mixed or categorical data. Secondly, data are self-reported which are less reliable and can be a source of bias. It can be due to socially desirable answers: overestimation of the time being active or underestimation of alcohol consumption (in the case of excessive consumption). Thirdly, missing values can definitively influence the results of an analysis. In this work, we try to examine their patterns, the assumptions and some available methods to handle missing values. However, missing values in clustering techniques could be the topic of an entire research which is out of the scope of this study. Moreover, missing values present in this work seem to be linked with the difficulty to answer some questions (difficult to estimate the number of minutes someone remains sitting per day), future research should consider to build surveys differently in order to limit the number of unanswered questions. Finally, future researches should take into account the selection bias. Even if the study design has been carefully constructed, resulting in a representative sample in terms of sex, age and education level, the participation must be investigated. Indeed, if the health behaviors of the non-responders differ systematically from the ones of responders, or if the reserve individual with similar characteristics has different lifestyle habits from the initially selected individual, this can lead to bias.

Chapter 6

Conclusion

This study in collaboration with Sciensano enables to get a better understanding of the clustering of (un)healhty behaviors among the adult Flemish population and their links with intentions to change behaviors. It highlights a number of valuable findings to design targeted intervention policies. These findings were obtained by applying several statistical methods on the data: clustering methods, a multinomial logistic regression and an association analysis.

In the clustering analysis, we have compared 3 different methods and their results. The popular K-means clustering method has been adapted to handle categorical/mixed data using Gower distance. It results in 4 different clusters characterized by specific (un)healthy behaviors. Then, the fuzzy K-means has also been adapted to handle categorical/mixed data by using the Gower distance. It results in 4 more balanced clusters and membership values assigned to each observation. In this case, clusters were only characterized by 1 or 2 highly prevalent unhealthy behaviors. Finally, the LCA, which relies on another clustering approach, has been applied to the datasets. It results in 5 different clusters in which 1 was highly characterized by unhealthy diet habits. The LCA has been chosen for the rest of the analyses based on the model assumptions, the results and the objective of the study. From this clustering analysis, 5 clusters have been found, each one highlighting different behavioral patterns: the healthy lifestyle, the unhealthy diet, the substance usage, the mixed lifestyle and the sedentary lifestyle clusters. The majority of the population tends to engage in healthy behaviors and a minority of the population tends to engage in substance usage.

Once the clusters have been determined, a multinomial logistic regression has been applied by taking the healthy cluster as reference. This analysis highlights that men are more likely to engage in unhealthy behaviors (unhealthy diet, substance usage, mixed lifestyle and inactive lifestyle clusters). Individuals in younger age categories are associated with the unhealthy diet and substance usage clusters. Higher educated individuals tend to belong to the healthy lifestyle cluster while middle educated individuals tend to belong to the substance usage cluster. Additionally, individuals with (great) difficulties to end months are more likely to engage in unhealthy behaviors (unhealthy diet, substance usage and inactive lifestyle clusters). Individuals who are overweight or obese tend to engage in unhealthy diet behaviors and in an inactive lifestyle. Individuals living alone or living alone + child(ren) are more likely to be associated with the substance usage, mixed lifestyle and inactive lifestyle clusters. Moreover, individuals living with parents, friends etc. tend to engaging in substance use behaviors. At

last, individuals who speak Dutch at home tend to belong to the unhealthy diet cluster.

The third analysis has enabled to assess the association per cluster of the behavioral change intentions using Fisher Exact's tests corrected for multiplicity. From a descriptive analysis, we have found that a large majority of individuals are willing to change their diet and physical activity habits, regardless of the cluster. Moreover, a significant proportion of individuals wants to change their sedentary and smoking behaviors. The intention to reduce alcohol consumption concerns around a quarter of individuals, while the intention to stop drinking represents between 5% to 10%. At last, 30% of the individuals in the substance usage cluster are willing to stop cannabis. From the association analysis, we have found that intentions to change physical activity, sedentary and diet behaviours are associated regardless of the cluster. In the mixed lifestyle cluster, intention to quit smoking is related to the intention to change reduce alcohol, be more physical active, be less sedentary and have better diets habits. Moreover, the association between intentions to quit smoking and to change diet habits was found in the unhealthy diet cluster. At last, cannabis intention to change is related to other intention behaviors only within the substance usage cluster.

Finally, we have faced a significant number of missing item values. In order to treat them in the most appropriate way within the scope of this study, a sensitivity analysis has been conducted to compare 3 methods to handle them: the complete case analysis, the mice algorithm and the random forest for imputation algorithm. It results that the outcomes of the clustering analyses were not much influenced by the method used to handle missing data. The random forest algorithm was finally chosen as it requires fewer assumptions.

As discussed in the previous section, it is the first study of this kind in the Flemish population and future researches should take into account the listed limitations. There are numerous possibilities for further research. One would be to replicate this study in different regions of Belgium, in order to study potential regional disparities. In addition, the integration of behavioral factors related to mental health could provide valuable information. The potential for expanding this research is immense and multifaceted.

Appendix

Participtation rate

	Invited	Participant (%)	Paper questionnaire (%)
Total	19995	4011 (20.1%)	1485 (37.0%)
Sex			
male	10464	1990 (19.0%)	633 (31.8%)
female	9531	2021 (21.2%)	852 (42.2%)
Age			
18-24	2212	443 (20.0%)	75 (16.9%)
25-34	3565	549 (15.4%)	108 (19.7%)
35-44	3368	570 (16.9%)	133 (23.3%)
45-54	3279	695 (21.2%)	222 (31.9%)
55-64	2826	776 (27.5%)	330 (42.5%)
65-74	2207	597 (27.0%)	337 (56.5%)
75+	2538	381 (15.0%)	280 (73.5%)
Education level			
lower level	8137	1264 (15.5%)	665 (52.6%)
middle level	6807	1384 (20.3%)	479 (34.6%)
higher level	5051	1363 (27.0%)	341 (25.0%)
Health care region			
Aalst	1158	202 (17.4%)	71 (35.2%)
Antwerpen	2859	561 (19.6%)	71 (35.2%)
Brugge	983	199 (20.2%)	86 (43.2%)
Brusselse rand	1937	360 (18.6%)	117 (32.5%)
Genk	835	155 (18.6%)	62 (40.0%)
Gent	2774	548 (19.8%)	202 (36.9%)
Hasselt	1918	381 (19.9%)	149 (39.1%)
Kortrijk	1006	196 (19.5%)	75 (38.3%)
Leuven	1440	325 (22.6%)	119 (36.6%)
Mechelen	1230	270 (22.0%)	82 (30.4%)
Oostende	606	121 (20.0%)	55 (45.5%)
Roeselare	1112	220 (19.8%)	93 (42.3%)
Sint-Niklaas	766	159 (20.8%)	71 (44.7%)
Turnhout	1371	314 (22.9%)	120 (38.2%)

Table 6.0.1: Number of people invited to complete the survey and effective response rates by socio-demographic variables taken from the National Register and type of questionnaire

Questionnaire - Intention questions

Here is an example on how the questionnaire was constructed for the intention questions.

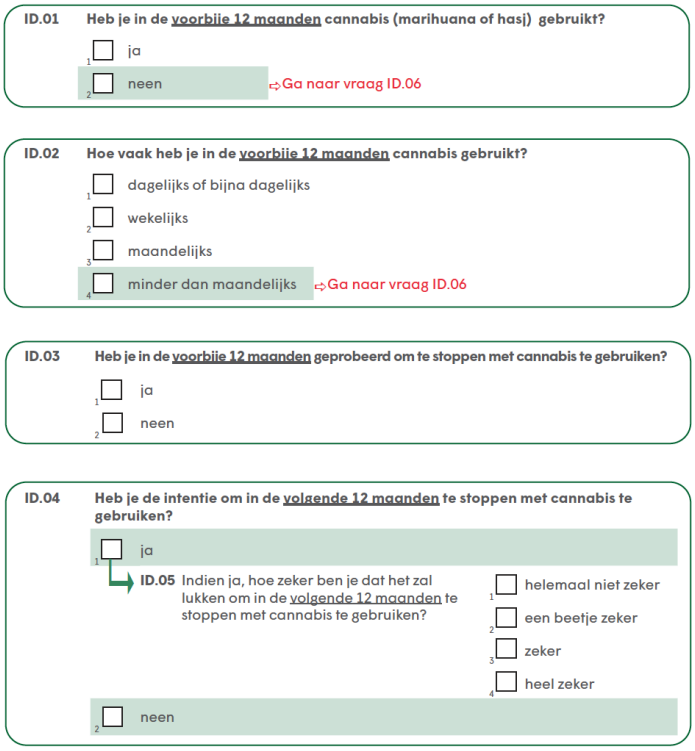

Figure 6.0.1

Tests of association

The choice of the test depends on the sample size, on the distribution of the data and on if data are paired. Some tests are not appropriated for small sample size as their ability to detect a significant association depends on the sample size. Moreover, some tests make assumptions about the distribution of the data (the Chi-square test make the assumption that frequencies for each category are large enough). The below table shows the guideline to use when choosing the appropriate statistical test.

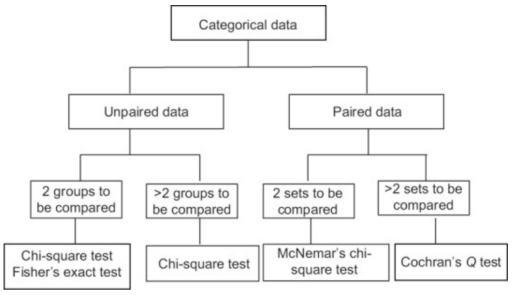

Figure 6.0.2: Statistical tests to analyse association between categorical variables [53]

Missing values

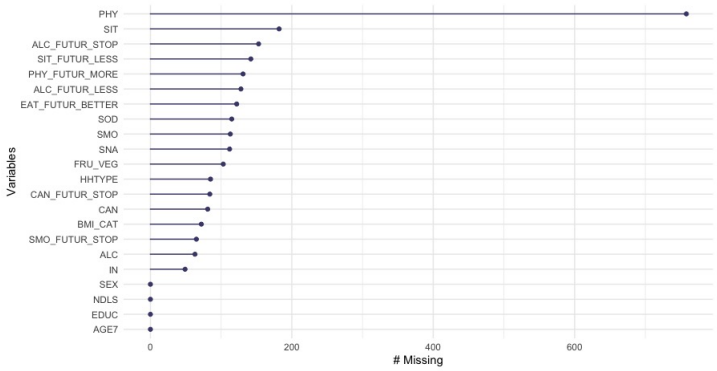

Figure 6.0.3: Missing values distribution: Number of missing values per variables.

Descriptive statistics

Sample distribution of unhealthy behaviors

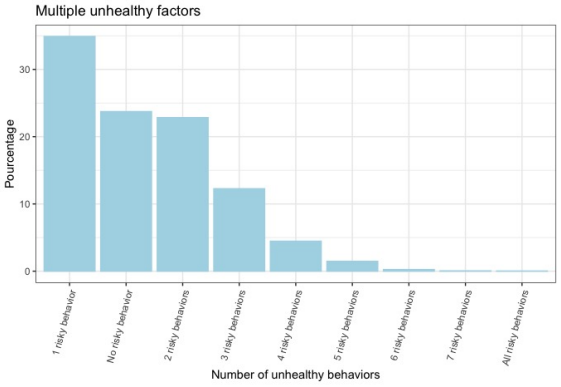

Figure 6.0.4: Forest sample distribution in terms of unhealthy behaviors combinations

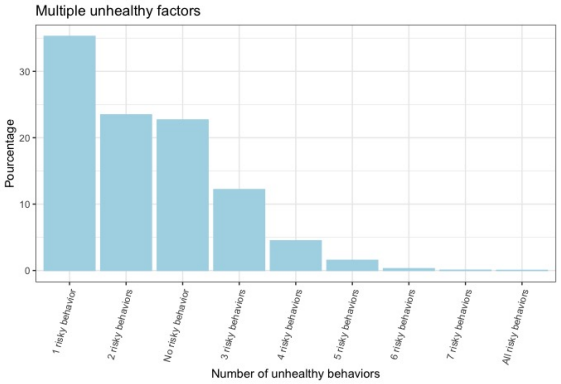

Figure 6.0.5: Mice sample distribution in terms of unhealthy behaviors combinations

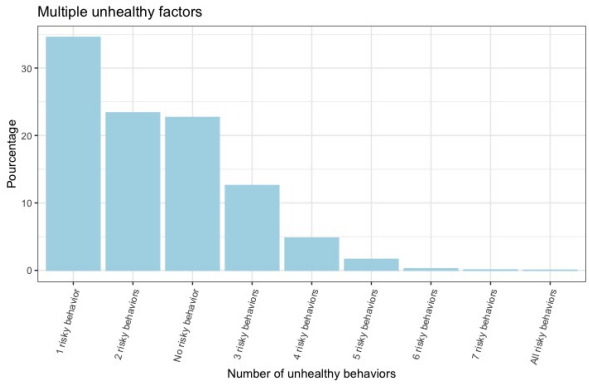

Figure 6.0.6: Complete sample distribution in terms of unhealthy behaviors combinations

Association between health related behaviors and socio-demographic factors

Variables	H. type	Educ.	Income	BMI	Dutch	Sex	Age
Cannabis	<0.001*	<0.001*	<0.001*	<0.001*	1	<0.001*	<0.001*
Smoking	<0.001*	<0.001*	<0.001*	0.666	0.666	<0.001*	<0.001*
Alcohol	<0.001*	<0.001*	<0.001*	<0.001*	0.605	<0.001*	<0.001*
Phy.Act.	<0.001*	0.457	<0.001*	<0.001*	0.045*	<0.001*	<0.001*
Sitting	<0.001*	<0.001*	<0.001*	0.034*	0.289	<0.001*	<0.001*
Fru.Veg.	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
Snack	0.766	<0.001*	<0.001*	<0.001*	<0.001*	0.054	0.640
Soda	<0.001*	<0.00*1	0.001*	0.001*	<0.001*	<0.001*	<0.001*

Table 6.0.2: Association between variables of the forest dataset: p-value of the Chi Squared test

Variables	H. type	Educ.	Income	BMI	Dutch	Sex	Age
Cannabis	<0.001*	<0.001*	<0.001*	<0.001*	1	<0.001*	<0.001*
Smoking	<0.001*	<0.001*	<0.001*	0.712	0.712	<0.001*	<0.001*
Alcohol	<0.001*	<0.001*	<0.001*	<0.001*	0.903	<0.001*	<0.001*
Phy.Act.	<0.001*	0.089	<0.001*	<0.001*	0.024*	<0.001*	<0.001*
Sitting	<0.001*	<0.001*	<0.001*	0.012*	0.317	<0.001*	<0.001*
Fru.Veg.	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
Snack	0.783	<0.001*	<0.001*	<0.001*	<0.001*	0.124	0.900
Soda	<0.001*	<0.00*1	0.001*	0.001*	<0.001*	<0.001*	<0.001*

Table 6.0.3: Association between variables of the mice dataset: p-value of the Chi Squared test

Variables	H. type	Educ.	Income	BMI	Dutch	Sex	Age
Cannabis	<0.001*	<0.001*	<0.001*	0.016*	0.975	<0.001*	<0.001*
Smoking	<0.001*	<0.001*	<0.001*	0.923	0.861	<0.001*	<0.001*
Alcohol	0.019*	<0.001*	<0.001*	0.019*	1	<0.001*	<0.001*
Phy.Act.	<0.001*	0.158	<0.001*	<0.001*	0.035*	<0.001*	<0.001*
Sitting	<0.001*	<0.001*	<0.001*	0.011*	0.556	<0.001*	<0.001*
Fru.Veg.	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
Snack	0.865	<0.001*	<0.001*	<0.001*	<0.001*	0.057	0.698
Soda	<0.001*	<0.00*1	0.001*	0.001*	<0.001*	<0.001*	<0.001*

Table 6.0.4: Association between variables of the forest dataset: p-value of the Chi Squared test

Association between health related behaviors

Variables	Cannabis	Smoking	Alcohol	Phy. act.	Sitting	Fru. veg.	Soda
Cannabis							
Smoking	<0.001*						
Alcohol	<0.001*	<0.001*					
Phy.Act.	<0.001*	<0.001*	<0.001*				
Sitting	<0.001*	<0.001*	<0.001*	<0.001*			
Fru.Veg.	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*		
Soda	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	
Snack	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*

Table 6.0.5: Association between health variables (for the original, the forest, the mice and complete datasets): p-value of the Chi Squared test

Clustering analysis - Probabilities of unhealthy behaviors

K-means

Variable	Cluster 1 (n=932)	Cluster 2 (n=508)	Cluster 3 (n=760)	Cluster 4 (n=714)
Cannabis	4.29%	6.10%	4.61%	5.60%
Smoking	11.6%	20.5%	12.4%	10.8%
Alcohol	10.8%	15.6%	10.00%	8.54%
Phy.Act.	0.00%	23.40%	69.30%	22.70%
Sitting	0.00%	14.8%	69.3%	16.00%
Soda	8.91%	15.20%	12.10%	23.20%
Snack	0.00%	0.00%	10.40%	100%
Fru/veg	0.00%	100%	14.70%	18.60%

Table 6.0.6: K-means under complete dataset - Percentages of individuals engaging in unhealthy behaviors within each cluster

Variable	Cluster 1 (n=1562)	Cluster 2 (n=1047)	Cluster 3 (n=781)	Cluster 4 (n=621)
Cannabis	3.78%	6.02%	2.82%	5.15%
Smoking	11.5%	19.2%	14.00%	9.18%
Alcohol	10.30%	12.80%	8.07%	8.86%
Phy.Act.	3.33%	34.6%	83.20%	3.54%
Sitting	17.40%	23.80%	19.30%	17.10%
Soda	9.22%	20.40%	16.00%	18.20%
Snack	0.83%	21.60%	28.90%	98.60%
Fru/veg	0.70%	98.50%	0.90%	0.32%

Table 6.0.7: K-means under mice dataset - Percentages of individuals engaging in unhealthy behaviors within each cluster

Variable	Cluster 1 (n=1562)	Cluster 2 (n=1047)	Cluster 3 (n=781)	Cluster 4 (n=621)
Cannabis	3.59%	5.49%	4.11%	4.82%
Smoking	11.80%	19.50%	13.40%	10.10%
Alcohol	9.79%	12.80%	10.10%	8.49%
Phy.Act.	0.00%	24.70%	72.60%	19.60%
Sitting	0.00%	11.70%	66.80%	12.50%
Soda	19.57%	19.60%	12.70%	20.90%
Snack	0.00%	20.3%	10.60%	100%
Fru/veg	0.00%	100%	14.40%	0.00%

Table 6.0.8: K-means under forest dataset - Percentages of individuals engaging in unhealthy behaviors within each cluster

Fuzzy K-means

Variable	Cluster 1 (n=923)	Cluster 2 (n=650)	Cluster 3 (n=813)	Cluster 4 (n=528)
Cannabis	3.68%	7.08%	5.54%	3.98%
Smoking	10.8%	20.5%	14.0%	6.82%
Alcohol	9.97%	15.5%	10.6%	7.20%
Phy.Act.	0.00%	25.1%	70.5%	13.6%
Sitting	0.00%	14.8%	65.2%	17.0%
Snack	0.00%	20.5%	16.4%	100%
Soda	8.56%	20.5%	16.2%	13.8%
Fru/veg	0.00%	98.6%	13.8%	0.00%

Table 6.0.9: Fuzzy K-means under complete dataset - Percentages of individuals engaging in unhealthy behaviors within each cluster

Variable	Cluster 1 (n=1479)	Cluster 2 (n=877)	Cluster 3 (n=939)	Cluster 4 (n=716)
Cannabis	2.57%	6.04%	6.18%	3.77%
Smoking	9.87%	20.9%	18.2%	7.26%
Alcohol	8.25%	14.4%	11.9%	6.98%
Phy.Act.	0.00%	20.2%	91.6%	14.0%
Sitting	13.7%	13.9%	47.5%	14.8%
Snack	0.00%	16.9%	23.2%	100%
Soda	7.80%	18.1%	23.1%	15.5%
Fru/veg	0.00%	98.7%	19.8%	0.00%

Table 6.0.10: Fuzzy K-means under mice dataset - Percentages of individuals engaging in unhealthy behaviors within each cluster

Variable	Cluster 1 (n=1526)	Cluster 2 (n=861)	Cluster 3 (n=899)	Cluster 4 (n=725)
Cannabis	2.82%	5.92%	5.90%	4.00%
Smoking	9.76%	20.4%	18.2%	7.86%
Alcohol	7.99%	13.9%	12.7%	7.86%
Phy.Act.	0.00%	19.3%	91.5%	13.4%
Sitting	12.9%	12.4%	50.6%	13.4%
Snack	0.00%	14.6%	25.1%	100%
Soda	8.13%	18.5%	22.2%	15.6%
Fru/veg	0.00%	99.0%	22.1%	0.00%

Table 6.0.11: Fuzzy K-means under forest dataset - Percentages of individuals engaging in unhealthy behaviors within each cluster

LCA

Variable	Cluster 1 (8.03%)	Cluster 2 (13.73%)	Cluster 3 (54.17%)	Cluster 4 (19.35%)	Cluster 5 (4.72%)
Cannabis	0.96%	2.22%	2.182%	3.53%	58.62%
Smoking	16.4%	16.52%	3.75%	17.97%	85.83%
Alcohol	2.29%	9.93%	7.33%	18.7%	36.94%
Phy.Act.	65.92%	0.00%	81.63%	1.00%	72.31%
Sitting	15.16%	48.60%	22.56%	16.20%	28.12%
Snack	62.02%	20.33%	33.09%	0.00%	32.16%
Soda	100%	12.37%	2.56%	9.06%	31.38%
Fru/veg	35.71%	48.16%	11.25%	43.18%	40.51%

Table 6.0.12: LCA under complete dataset - Conditional probabilities of an individual within clusters engaging in specific unhealthy behaviors

Variable	Cluster 1 (11.40%)	Cluster 2 (13.72%)	Cluster 3 (34.88%)	Cluster 4 (35.01%)	Cluster 5 (5.00%)
Cannabis	1.83%	4.42%	1.21%	1.90%	49.77%
Smoking	25.72%	7.50%	2.92%	12.95%	85.04%
Alcohol	6.26%	14.04%	6.40%	10.29%	35.01%
Phy.Act.	52.14%	43.93%	20.37%	22.28%	29.48%
Sitting	26.42%	99.97%	11.18%	0.00%	24.33%
Snack	51.99%	16.04%	49.63%	0.00%	30.77%
Soda	62.42%	1.55%	12.98%	4.85%	29.14%
Fru/veg	48.09%	28.75%	10.12%	31.37%	43.40%

Table 6.0.13: LCA under mice dataset - Conditional probabilities of an individual within clusters engaging in specific unhealthy behaviors

Variable	Cluster 1 (8.23%)	Cluster 2 (12.60%)	Cluster 3 (54.73%)	Cluster 4 (20.71%)	Cluster 5 (3.73%)
Cannabis	0.25%	1.61%	1.84%	2.72%	69.62%
Smoking	19.93%	16.58%	4.62%	21.13 %	80.03%
Alcohol	3.73%	9.44%	7.08%	17.52%	34.88%
Phy.Act.	31.07%	100%	17.02%	7.83%	26.31%
Sitting	9.98%	53.8%	18.58%	12.13%	28.34%
Snack	59.51%	23.13%	32.12%	0.00%	39.24%
Soda	100%	15.69%	2.93%	9.21%	30.62%
Fru/veg	32.08%	45.32%	12.47%	45.91%	40.83%

Table 6.0.14: LCA under forest dataset - Conditional probabilities of an individual within clusters engaging in specific unhealthy behaviors

Interactions between future intentions

These tables enable to get the percentages of individuals within each cluster who want to change one specific pair of 2 unhealthy variables.

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Smoking		0.17%	0.00%	0.62%	0.26%	0.58%
Alc. Red.			4.85%	16.03%	11.75%	15.63%
Alc. Stop				4.67%	3.65%	4.40%
Phy.Act.					39.18%	57.12%
Sitting						37.53%

Table 6.0.15: Future intentions within the healthy lifestyle cluster: Cross-Tabulation

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Smoking		0.68%	0.68%	4.80%	2.51%	3.89%
Alc. Red.			5.49%	12.12%	7.55%	10.75%
Alc. Stop				4.80%	2.28%	4.11%
Phy.Act.					34.32%	45.99%
Sitting						57.89%

Table 6.0.16: Future intentions within the unhealthy diet cluster: Cross-Tabulation

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	5.35%	5.35%	1.78%	11.60%	1.78%	9.82%
Smoking		6.25%	1.78%	21.42%	10.71%	16.96%
Alc. Red.			3.57%	23.21%	12.50%	21.42%
Alc. Stop				4.46%	2.67%	4.46%
Phy.Act.					34.82%	58.92%
Sitting						30.36%

Table 6.0.17: Future intentions within the substance usage cluster: Cross-Tabulation

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Smoking		3.17%	0.79%	6.20%	4.29%	6.35%
Alc. Red.			5.08%	20.50%	15.26%	20.19%
Alc. Stop				6.04%	4.92%	5.56%
Phy.Act.					41.17%	58.98%
Sitting						40.06%

Table 6.0.18: Future intentions within the mixed lifestyle cluster: Cross-Tabulation

Variables	Smoking	Alc. Red.	Alc. Stop	Phy. Act.	Sitting	Diet
Cannabis	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Smoking		0.85%	0.34%	3.06%	2.21%	3.57%
Alc. Red.			3.06%	14.31%	10.22%	13.96%
Alc. Stop				2.38%	2.04%	2.38%
Phy.Act.					44.97%	41.05%
Sitting						41.05%

Table 6.0.19: Future intentions within the inactive lifestyle cluster: Cross-Tabulation

Proportion of non-responses by age category

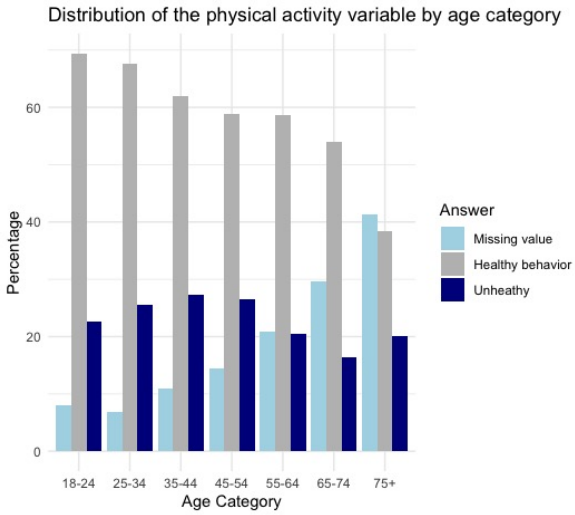

Figure 6.0.7: Distribution of responses to the physical activity variable by age category

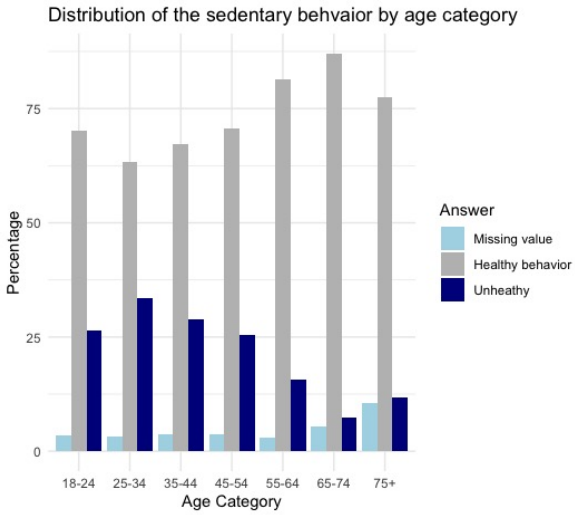

Figure 6.0.8: Distribution of responses to the physical activity variable by age category

Cluster profiles - Distribution of health related behaviors in the sample

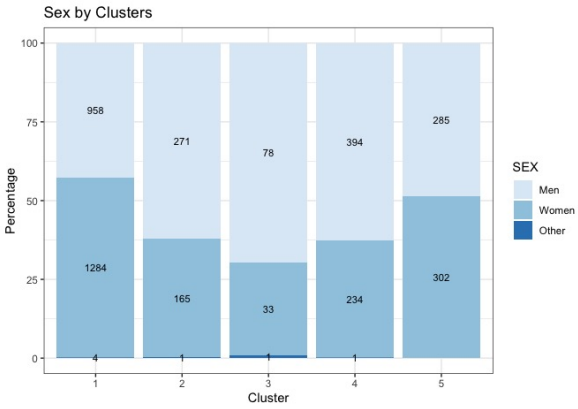

Figure 6.0.9: Distribution of the sex variable within each cluster

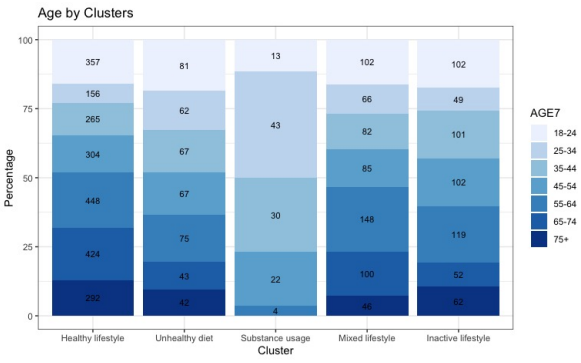

Figure 6.0.10: Distribution of the age variable within each cluster

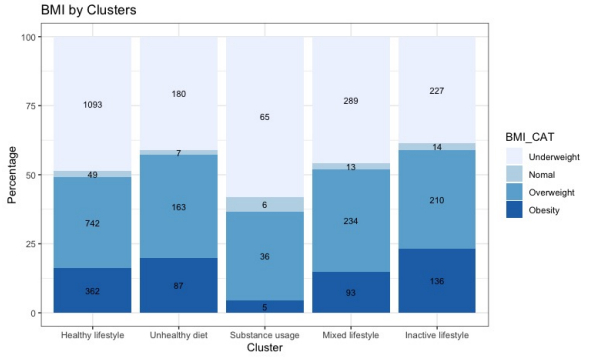

Figure 6.0.11: Distribution of the BMI variable within each cluster

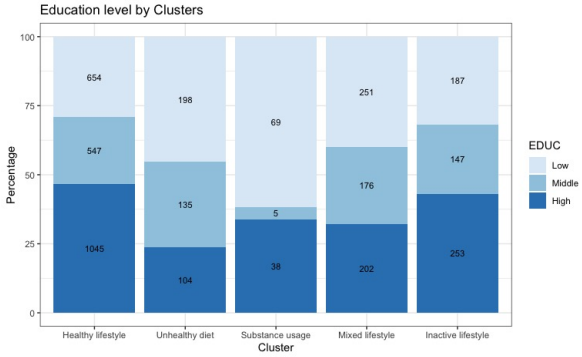

Figure 6.0.12: Distribution of the education level variable within each cluster

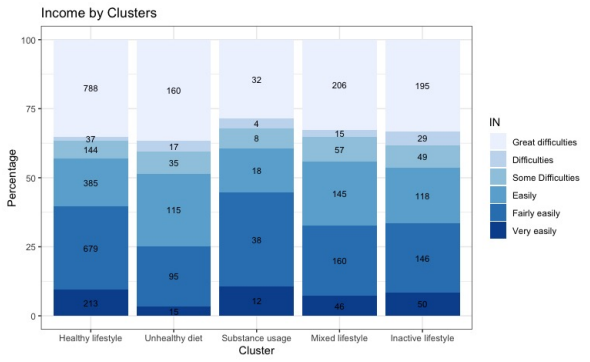

Figure 6.0.13: Distribution of the income variable within each cluster

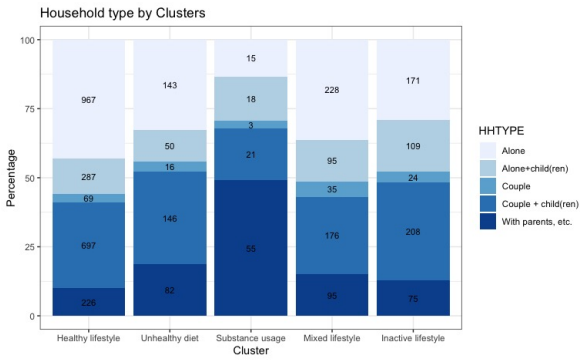

Figure 6.0.14: Distribution of the household type variable within each cluster

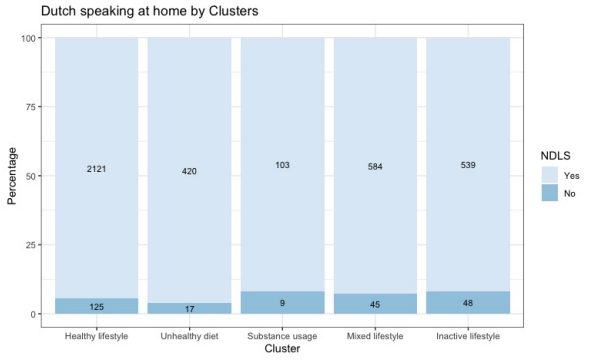

Figure 6.0.15: Distribution of the dutch speaking at home variable within each cluster

Bibliography

[1] World Health Organization. (2018). Noncommunicable diseases country profiles 2018. World Health Organization. https://apps.who.int/iris/handle/10665/274512

[2] Ford, E. S., Bergmann, M. M., Boeing, H., Li, C., Capewell, S. (2012). Healthy lifestyle behaviors and all-cause mortality among adults in the United States. Preventive Medicine, 55(1), 23-27. https://doi.org/10.1016/j.ypmed.2012.04.016

[3] Noble, N., Paul, C., Sanson-Fisher, R., Turon, H., Turner, N., Conigrave, K. M. (2016). Ready, set, go : a cross-sectional survey to understand priorities and preferences for multiple health behaviour change in a highly disadvantaged group. BMC Health Services Research, 16(1). https://doi.org/10.1186/s12913-016-1701-2

[4] Noble, N., Paul, C., Turon, H., Oldmeadow, C. (2015). Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity (SNAP) health risk factors. Preventive Medicine, 81, 16 - 41. https://doi.org/10.1016/j.ypmed.2015.07.003

[5] Hobbs, M., Duncan, M. J., Collins, P., McKenna, J. A., Schoeppe, S., Rebar, A. L., Alley, S., Short, C., Vandelanotte, C. (2019). Clusters of health behaviours in Queensland adults are associated with different socio-demographic characteristics. Journal of Public Health, 41(2), 268 - 277. https://doi.org/10.1093/pubmed/fdy043

[6] Britt, E., Hudson, S. R., Blampied, N. M. (2004). Motivational interviewing in health settings : a review. Patient Education and Counseling, 53(2), 147155. https://doi.org/10.1016/s07383991(03)001411

[7] Braekman, E., Fiers, S. (2023). Preventiebarometer: Beweging en sedentair gedrag. Sciensano.be. Retrieved from https://www.sciensano.be/en/biblio/preventiebarometer-beweging-en-sedentair-gedrag 24

[8] Deliu, M., Sperrin, M., Belgrave, D., Custovic, A. (2016). Identification of Asthma Subtypes Using Clustering Methodologies. Pulmonary therapy, 2(1), 19â41. https://doi.org/10.1007/s41030-016-0017-z

[9] Geneva: World Health Organization; 2020 [cited 2023 Feb 3]. Report No.: Licence: CC BY-NC-SA 3.0 IGO. Available from: https://www.who.int/publications-detail-redirect/9789240015128

[10] Matthys, Frieda Zeeuws, D.. (2018). Risico's van alcoholgebruik (HGR NR 9438)

[11] Conseil Superieur de la Sante. Recommandations nutritionnelles pour la Belgique 2016. Bruxelles: CSS; 2016. Avis n° 9285. 31

[12] Shentu, Y., Xie, M. (2010). A note on dichotomization of continuous response variable in the presence of contamination and model misspecification. Statistics in Medicine, 29(21), 2200 – 2214. https://doi.org/10.1002/sim.3966

[13] R Core Team. (2023). cluster: Functions for Cluster Analysis. R package version 2.1.4. Retrieved from https://www.rdocumentation.org/packages/cluster/versions/2.1.4/topics/pam

[14] R Core Team. (2023). cluster: Functions for Cluster Analysis. R package version 2.1.4. Retrieved from https://www.rdocumentation.org/packages/cluster/versions/2.1.4/topics/fanny

[15] Linzer, D. A., Lewis, J. A. (2011). poLCA : AnRPackage for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, 42(10). https://doi.org/10.18637/jss.v042.i10

[16] Venables, W. N., Ripley, B. D. (2023). nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models. R package version 7.3 – 16. Retrieved from https://cran.microsoft.com/web/packages/nnet/nnet.pdf

[17] R Core Team. (2023). stats: Functions for Statistical Calculations. R package version 3.6.2. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fisher.test

[18] Artzner P. Delbaen F. Eber J. Heath D. 1999. Coherent measures of risk. Mathematical finance. 9, 203-228.

[19] Jie, C., Jiyue, Z., Junhui, W., Yusheng, W., Huiping, S., Kaiyan, L. (2020). Review on the Research of K-means Clustering Algorithm in Big Data. https://doi.org/10.1109/icece51594.2020.9353036

[20] Patil, S., Nuli, U. A. (2019). A Review of Clustering and Clustering Quality Measurement. International Journal of Computer Engineering in Research Trends. https://doi.org/10.22362/ijcert/2018/v5/i12/v5i1205

[21] Ali, B. B., Massmoudi, Y. (2013). Kmeans clustering based on gower similarity coefficient : A comparative study. Dans International Conference on Modeling, Simulation, and Applied Optimization. https://doi.org/10.1109/icmsao.2013.6552669

[22] Jamotton, Charlotte ; Hainaut, Donatien ; Hames, Thomas. Insurance analytics with clustering techniques. LIDAM Discussion Paper ISBA ; 2023/02 (2023) 27 pages 33

[23] Yuan, C., Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. J, 2(2), 226â235. https://doi.org/10.3390/j2020016

[24] Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), 857. https://doi.org/10.2307/2528823

[25] Dehariya, V. K., Shrivastava, S. K., Jain, R. K. (2010). Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms. Dans International Conference on Computational Intelligence and Communication Networks. https://doi.org/10.1109/cicn.2010.80

[26] Chan, K. M., Cheung, Y. (1992). Clustering of clusters. Pattern Recognition, 25(2), 211 – 217. https://doi.org/10.1016/0031-3203(92)90102-o

[27] Bezdek, J. C., Hall, L. D., Clarke, L. P. (1993). Review of MR image segmentation techniques using pattern recognition. Medical Physics, 20(4), 1033â1048. https://doi.org/10.1118/1.597000

[28] Atzendorf J, Apfelbacher C, Gomes de Matos E, et al Patterns of multiple lifestyle risk factors and their link to mental health in the German adult population: a cross-sectional study BMJ Open 2018;8:e022184. doi: 10.1136/bmjopen-2018-022184

[29] Aflaki, K., Vigod, S. N., Ray, J. G. (2022). Part I : A friendly introduction to latent class analysis. Journal of Clinical Epidemiology, 147, 168–170. https://doi.org/10.1016/j.jclinepi.2022.05.008

[30] Sinha, P., Calfee, C. S., Delucchi, K. L. (2021b). Practitioners Guide to Latent Class Analysis : Methodological Considerations and Common Pitfalls. Critical Care Medicine, 49(1), e63-e79. https://doi.org/10.1097/ccm.0000000000004710

[31] Harrouti, T. E., Azhari, M., Deqqaq, H., Abouabdellah, A., Aidi, S. E., Chaoui, H. (2022). Using Clustering Methods to Detect the Revealed Preferences of Moroccans towards the Electric Vehicles: Latent Class Analysis (LCA) and K-Modes Algorithm (K-MA). Mathematics and Statistics, 10(5), 971-980. https://doi.org/10.13189/ms.2022.100508

[32] El-Habil, A. M. (2012). An Application on Multinomial Logistic Regression Model. Pakistan Journal of Statistics and Operation Research, 8(2), 271. https://doi.org/10.18187/pjsor.v8i2.234

[33] Hashimoto, E. M., Ortega, E. M. M., Cordeiro, G. M., Suzuki, A. K., Kattan, M. W. (2020). The multinomial logistic regression model for predicting the discharge status after liver transplantation : estimation and diagnostics analysis. Journal of Applied Statistics, 47(12), 2159-2177. https://doi.org/10.1080/02664763.2019.1706725

[34] Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep. net/stat/mlelr. pdf, 83

[35] Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Crit Care. 2005 Feb;9(1):112-8. doi: 10.1186/cc3045. Epub 2005 Jan 13. PMID: 15693993; PMCID: PMC1065119.

[36] Bewick, V., Cheek, L. Ball, J. Statistics review 8: Qualitative data - tests of association. Crit Care 8, 46 (2003). https://doi.org/10.1186/cc2428

[37] Kwong, K. S., Holland, B., Cheung, S. G. (2002). A modified Benjamini-Hochberg multiple comparisons procedure for controlling the false discovery rate. Journal of Statistical Planning and Inference, 104(2), 351-362. https://doi.org/10.1016/s0378-3758(01)00252-x

[38] Efron, B. (1994). Missing Data, Imputation, and the Bootstrap. Journal of the American Statistical Association, 89(426), 463-475. https://doi.org/10.1080/01621459.1994.10476768

[39] Golino, H., Gomes, C. M. (2016). Random forest as an imputation method for education and psychology research : its impact on item fit and difficulty of the Rasch model. International Journal of Research Method in Education, 39(4), 401-421. https://doi.org/10.1080/1743727x.2016.1168798

[40] Mack C, Su Z, Westreich D. Managing Missing Patient Data in Patient Registries. White Paper, addendum to Registries for Evaluating Patient Outcomes: A Userâs Guide, Third Edition. (Prepared by LM Policy Research, LLC, under Contract No. 290-2014- 00004-C.) AHRQ Publication No. 17(18)-EHC015-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2018. www.effectivehealthcare.ahrq.gov. DOI: https://doi.org/10.23970/AHRQREGISTRIESMISSDATA

[41] Zimmermann, P., Mazouch, P., Tesarkova, K. H. (2014). Missing Categorical Data Imputation and Individual Observation Level Imputation. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62(6), 15271534. https://doi.org/10.11118/actaun201462061527

[42] Cro, S., Morris, T., Kenward, M. G., Carpenter, J. R. (2020). Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation : A practical guide. Statistics in Medicine, 39(21)

[43] Jamshidian, M., Mata, M. (2007). Advances in Analysis of Mean and Covariance Structure when Data are Incomplete. In Elsevier eBooks (pp. 21â44). https://doi.org/10.1016/b978-044452044-9/50005-7, M., Mata, M. (2007). Advances in Analysis of Mean and Covariance Structure when Data are Incomplete. Dans Elsevier eBooks (p. 21-â44). Elsevier BV. https://doi.org/10.1016/b978–044452044–9/50005–7

[44] Azur, M., Stuart, E. A., Frangakis, C., Leaf, P. J. (2011). Multiple imputation by chained equations : what is it and how does it work ? International Journal of Methods in Psychiatric Research, 20(1), 40â49. https://doi.org/10.1002/mpr.329

[45] Stekhoven, D. J., Buhlmann, P. (2012). MissForestnonparametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112118. https://doi.org/10.1093/bioinformatics/btr597

[46] Jakobsen, J. C., Gluud, C., Wetterslev, J., Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. BMC Medical Research Methodology, 17(1). https://doi.org/10.1186/s12874-017-0442-1

[47] Halkidi, M., Batistakis, Y. Vazirgiannis, M. On Clustering Validation Techniques. Journal of Intelligent Information Systems 17, 107-145 (2001). https://doi.org/10.1023/A:1012801612483

[48] Kaufman, L., Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley Sons.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc